Supporting Information

Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication

Devin R. Scannell, A. Carolin Frank, Gavin C. Conant, Kevin P. Byrne, Megan Woolfit,
Kenneth H. Wolfe

This Supporting Information (SI) contains:

SI Methods

Genome survey sequencing of *Kluyveromyces polysporus* and *Kluyveromyces phaffii*

The type strains of *Kluyveromyces polysporus* (DSMZ 70294) and *Kluyveromyces phaffii* (MUCL 31247) were obtained from the culture collections of the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen) and MUCL (Mycothèque de l'Université catholique de Louvain). DNA cloning and sequencing was done by GATC-Biotech (Konstanz, Germany). Genomic DNA was sheared by nebulization and random fragments of 1-2 kb were cloned into plasmids. Both ends of the inserts in 384 plasmids from each species were sequenced. Genes were identified by BLASTX and the gene order in fragments containing >1 gene was compared to other hemiascomycetes. In both *K. polysporus* and *K. phaffii* we found examples of neighboring genes that were close, but not immediate neighbors, in non-WGD species. This suggested that *K. polysporus* and *K. phaffii* are post-WGD species.

Scaffold Assembly

Sequence coverage in the Phrap assembly is 7.8x. We manually ordered and oriented 90% of the contigs into 41 supercontigs (SI Figure 6, which is published as supporting information on the PNAS web site), using a combination of physical scaffolds constructed by the program Bambus (1) based on fosmid read-pair information, and gene order information from comparisons to other yeast genomes. Within the supercontigs, adjacent contigs with overlapping or consecutive genes at their ends (as inferred by comparison with the non-WGD species *A. gossypii*, *K. waltii* and *K. lactis*) were physically joined by a stretch of 100 N's into longer contigs, reducing the total number of contigs from 546 to 424. The set of 290 contigs that are larger than 2 kb was retained for subsequent annotation and analysis. The total size of these contigs is 14,703,743 bp, and their N_{50} value is 125,449 bp (that is, half of the bases are in contigs of this size or larger). N_{50} for the supercontigs is 421,604 bp.

Annotation

We wrote a suite of Perl modules to automate identification of conserved features in the genome of *K. polysporus*. The modules provide data-structures to represent genomes at various levels of resolution from exons to scaffolds and wrappers to run external applications. We performed a three-step annotation. First, tRNAscan-SE (2) was used to identify tRNA genes and HMMER v1.8.4 (3) was used to identify putative telomeres and introns. Next, open reading frames (ORFs) above a context-dependent minimum length were identified and all possible gene structures were constructed by merging ORFs across introns, possible sequencing errors and scaffold gaps. Finally, a single gene structure was selected at each locus and all gene structures were evaluated with respect to conservation of sequence in other sequenced yeast genomes, synteny, learned codon-usage patterns and other heuristics. In total, 5927 possible protein-coding genes were identified and 5652 were retained as likely real genes. Perl modules are available on request from D.R.S. (email: dscannell@lbl.gov). Genes were initially named using the scheme *Kpol_{contig_number}.{gene_number}* where the gene numbers were consecutive within the contig. Subsequent manual curation resulted in the elimination of some numbered genes, and the discovery of some extra genes that were given names with lettered suffixes. Sequences have been deposited in GenBank with accession number AAZN00000000 and the data can be browsed in the Yeast Gene Order Browser (YGOB) platform at http://wolfe.gen.tcd.ie/ygob.

Gene Ontology annotation mapping and statistical tests

We mapped Gene Ontology terms to 3252 ancestral loci that satisfy YGOB's quality criteria (4). Among these, in *S. cerevisiae* 2819 ancestral loci have been returned to single-copy (singletons) and 433 ancestral loci have retained both gene copies (ohnologs), while in *K. polysporus* there are 2802 singletons and 450 ohnolog pairs.

In the analysis shown in SI Table 2, which is published as supporting information on the PNAS web site, for each GO term we counted the number of singletons in *S. cerevisiae* annotated with the term and the number of ohnolog loci at which both gene copies had been annotated with the term. For ohnolog loci at which a GO term had been assigned to only one of an ohnolog pair, the ohnolog count was incremented by one half. We identified GO terms that are either under- or over-represented among ohnolog loci relative to singleton loci using a two-sided Fisher's exact test and report all terms for which the P-value is less than or equal to 0.05, after applying the Benjamini and Hochberg correction for multiple-testing. We transferred all GO annotations mapped to *S. cerevisiae* genes present at an ancestral locus (either a singleton or an ohnolog pair) to the *K. polysporus* genes at that locus and identified GO terms that are either under- or over-represented among ohnolog loci relative to singleton loci as described above.

In *SI Appendix,* section 4 we describe two methods to calculate the expected number of shared duplicate pairs between *S. cerevisiae* and *K. polysporus* and the significance of the observed deviation from these values. In Figure 3 we calculated the expected number of shared duplicate pairs for individual GO categories using Method 2 (which accounts for the presence of a shared evolutionary branch) with the additional assumption that the proportion of loci preserved in duplicate on the shared evolutionary branch is the same as the genome average (1.93% / 7.35% = 0.26) and does not vary among GO categories.

Phylogenetic analysis

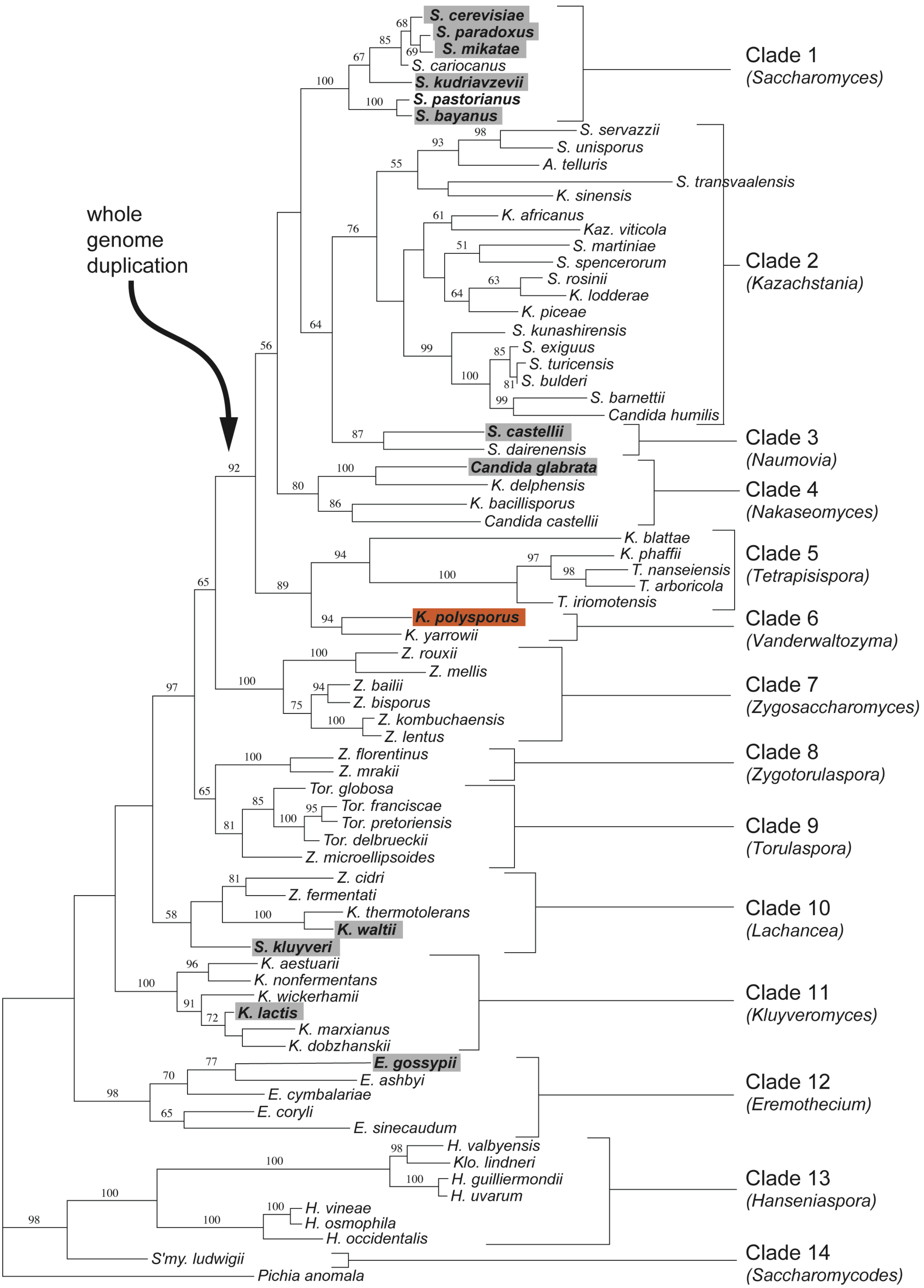
We used YGOB to select loci that have been retained in duplicate since the WGD by *S. cerevisiae, S. bayanus, C. glabrata, S. castellii* and *K. polysporus* and for which single-copy orthologs were also available in four additional yeast species (*K. lactis, K. waltii, A. gossypii* and *C. albicans*). Ignoring the *K. polysporus* genes, we first used YGOB to determine which of the two gene copies in *S. bayanus, C. glabrata* and *S. castellii* are orthologous to each of the two gene copies in *S. cerevisiae*. We were able to partition these duplicates into two clades (DC1, DC2), each consisting of four syntenic orthologs, for 92 loci.

Because of the high level of reciprocal gene loss between *K. polysporus* and *S. cerevisiae* we used phylogenetic methods rather than YGOB (which relies on conservation of synteny) to determine which of the two gene copies in *K. polysporus* is orthologous to each of the two gene copies in *S. cerevisiae.* For each locus we used ClustalW (5) and Gblocks (6) to generate an alignment from all 14 sequences and used Shimodaira-Hasegawa tests (7) (implemented in Tree-Puzzle (8)) to determine whether one of the two possible topologies was preferred: either *K. polysporus* copy 1 clusters with DC1 and *K. polysporus* copy 2 clusters with DC2 or *vice versa*. Loci at which there was significant (α = 0.05 level) support for one topology over the other were retained.

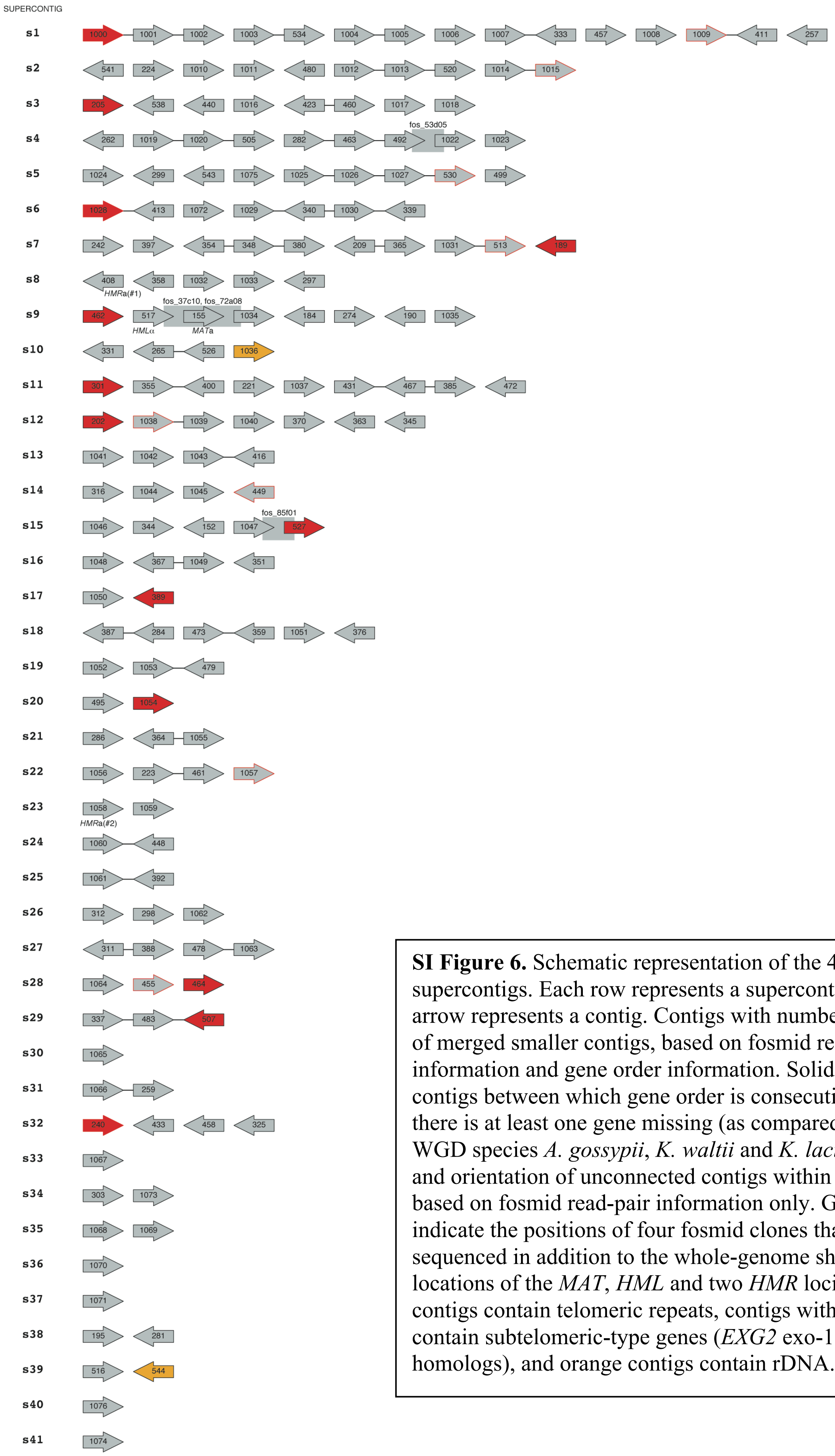
We also sought to exclude loci that may have undergone gene conversion (9). We used Phyml (10) to draw unconstrained trees for each locus with all five pairs of duplicates and the corresponding single ortholog in *K. lactis*. Any loci for which either DC1 or DC2 (including the appropriate *K. polysporus* ortholog) were not reconstructed were discarded. Eleven loci were retained for further analysis (*S. cerevisiae* gene names: *YBP2*/*YBP1*, *SWH1*/*OSH2*, *HST1*/*SIR2*, *FAR10*/*VPS64*, *SBE2*/*SBE22*, *GEA1*/*GEA2*, *SDT1*/*PHM8*, *SIR3*/*ORC1*, *FSH2*/*FSH3*, *CDC50*/*YNR048W* and *TRF4*/*TRF5*), and super-alignments of these loci were used for phylogenetic analysis.

At any given locus all the gene copies in DC1 (or DC2) are orthologous to one another and are paralogous to the gene copies in DC2 (or DC1). There is however no relationship between the gene copies in DC1 at one locus and the gene copies in DC1 at other loci. It is therefore possible to concatenate gene copies from DC1 at one locus with gene copies from DC2 at other loci (provided all gene copies in DC1 are treated consistently) when constructing a super-alignment. We used this fact to exclude the possibility that generating a single super-alignment might result in concatenation of the faster-evolving clades (DC1 and DC2 can evolve at very different rates) at several loci. Instead, we generated 100 super-alignments (4045 amino acid sites each) in which the DC1/DC2 designation was randomly reversed with probability 0.5 for each locus. Finally, for each of the 100 super-alignments a single bootstrap-replicate was generated using 'seqboot' in the Phylip package and these – rather than the original super-alignments – were retained for phylogenetic reconstruction.

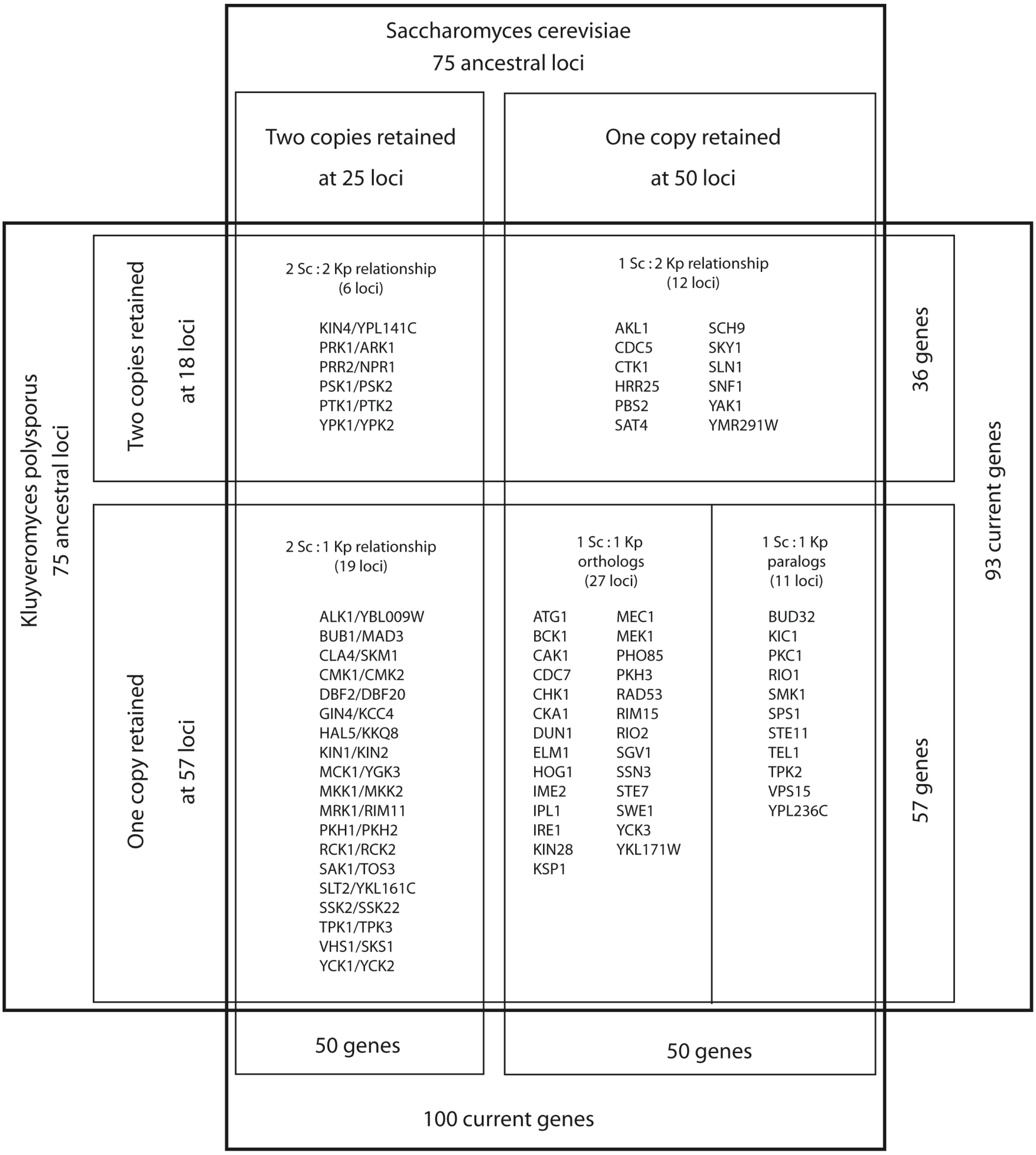
Because the phylogenetic relationships between the yeasts used in this study are known (11, 12) we optimized branch-lengths but not the topology (modified to include *K. polysporus*) for each of 100 bootstrap-replicates using a WAG + I + G(8) + F model. Finally, branch-lengths were averaged between duplicate clades and across all 100 bootstrap-replicates to obtain the tree in Figure 2C. We did not correct the tree in Figure 2C for the effect of accelerated protein sequence evolution after WGD because we found that the method used in (11) yielded a small negative length for the branch between the WGD and the *K. polysporus* divergence (D.R.S. and K.H.W., in preparation).



SI Figure 5. Phylogenetic tree of the 14 clades of hemiascomycetes, redrawn from Kurtzman and Robnett (12, 13). Species with sequenced genomes are highlighted and the inferred position of the WGD is indicated.



SI Figure 6. Schematic representation of the 41 *K. polysporus* supercontigs. Each row represents a supercontig, and each arrow represents a contig. Contigs with numbers >1000 consist of merged smaller contigs, based on fosmid read-pair information and gene order information. Solid lines connect contigs between which gene order is consecutive, but where there is at least one gene missing (as compared to the non-WGD species *A. gossypii*, *K. waltii* and *K. lactis*). The order and orientation of unconnected contigs within a supercontig is based on fosmid read-pair information only. Gray rectangles indicate the positions of four fosmid clones that we completely sequenced in addition to the whole-genome shotgun phase. The locations of the *MAT*, *HML* and two *HMR* loci are shown. Red contigs contain telomeric repeats, contigs with red outline contain subtelomeric-type genes (*EXG2* exo-1,3-beta-glucanase homologs), and orange contigs contain rDNA.



SI Figure 7. Differential resolution of protein kinase gene pairs in *K. polysporus* and *S. cerevisiae*. Genes are identified by their *S. cerevisiae* names. The set of genes is based on (14). Protein kinases that are not listed could not be scored on both tracks in both species, due to sequence gaps or lack of synteny.

SI Table 2. All Gene Ontology (GO) terms that are significantly under- or over-represented among loci retained in duplicate since the WGD in *K. polysporus* relative to single-copy genes.

Gene Ontology Term	Ohnologs		Singletons		Corrected P-value
	Count	Percentage	Count	Percentage	
Death	17	3.78%	16	0.57%	3.87E-07
cell death	16.5	3.67%	16	0.57%	1.42E-06
regulation of biological process	90	20.00%	295.5	10.55%	3.21E-06
Cytosol	49.5	11.00%	129.5	4.62%	5.41E-06
Aging	14	3.11%	14	0.50%	6.37E-06
regulation of physiological process	87.5	19.44%	290.5	10.37%	7.68E-06
regulation of cellular physiological process	84	18.67%	282	10.06%	1.15E-05
regulation of cellular process	84	18.67%	282	10.06%	1.15E-05
Cytosolic ribosome (sensu Eukaryota)	26.5	5.89%	51	1.82%	1.66E-05
Golgi-associated vesicle	16.5	3.67%	21.5	0.77%	2.24E-05
cell aging	13.5	3.00%	14	0.50%	2.28E-05
COPII vesicle coat	6	1.33%	1	0.04%	4.51E-05
ER to Golgi transport vesicle membrane	6	1.33%	1	0.04%	4.51E-05
Vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
cytoplasmic vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
cytoplasmic membrane-bound vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
membrane-bound vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
RNA processing	11.5	2.56%	216.5	7.73%	7.14E-05
G1/S transition of mitotic cell cycle	12	2.67%	13.5	0.48%	7.91E-05
interphase	19	4.22%	34.5	1.23%	7.98E-05
interphase of mitotic cell cycle	19	4.22%	34.5	1.23%	7.98E-05
Cytosolic small ribosomal subunit (sensu Eukaryota)	13	2.89%	18	0.64%	0.000135
eukaryotic 48S initiation complex	13	2.89%	18	0.64%	0.000135
replicative cell aging	10.5	2.33%	10	0.36%	0.000136
eukaryotic 43S preinitiation complex	15	3.33%	24	0.86%	0.000138
positive regulation of cellular process	15	3.33%	24	0.86%	0.000138
positive regulation of cellular physiological process	15	3.33%	24	0.86%	0.000138
positive regulation of physiological process	15	3.33%	24	0.86%	0.000138
positive regulation of transcription	14	3.11%	21	0.75%	0.000139
carbohydrate metabolism	31.5	7.00%	81	2.89%	0.000162
ER to Golgi transport vesicle	9.5	2.11%	7.5	0.27%	0.000169
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	14	3.11%	22	0.79%	0.000199
positive regulation of transcription, DNA-dependent	13	2.89%	19	0.68%	0.000199
organellar ribosome	0	0.00%	61	2.18%	0.000212
mitochondrial ribosome	0	0.00%	61	2.18%	0.000212
positive regulation of biological process	16.5	3.67%	29	1.03%	0.000245
RNA metabolism	21.5	4.78%	296	10.56%	0.000254
regulation of progression through cell cycle	25.5	5.67%	61	2.18%	0.000261
regulation of cell cycle	25.5	5.67%	61	2.18%	0.000261
cell wall organization and biogenesis	25	5.56%	60.5	2.16%	0.000261
external encapsulating structure organization and biogenesis	25	5.56%	60.5	2.16%	0.000261
cellular carbohydrate metabolism	29.5	6.56%	74.5	2.66%	0.000276
positive regulation of cellular metabolism	14	3.11%	23	0.82%	0.000279
positive regulation of metabolism	14	3.11%	23	0.82%	0.000279

protein amino acid O-linked glycosylation	5.5	1.22%	1	0.04%	0.00028
coated vesicle	17.5	3.89%	32.5	1.16%	0.000295
regulation of metabolism	59	13.11%	202.5	7.23%	0.00034
Golgi apparatus	29	6.44%	77.5	2.77%	0.00037
response to oxidative stress	12.5	2.78%	19	0.68%	0.000586
regulation of cellular metabolism	54.5	12.11%	189	6.75%	0.00063
oxygen and reactive oxygen species metabolism	12.5	2.78%	20	0.71%	0.000819
transport vesicle membrane	6	1.33%	4	0.14%	0.000934
Golgi-associated vesicle membrane	6	1.33%	4	0.14%	0.000934
glucose metabolism	13	2.89%	23.5	0.84%	0.00103
mitotic cell cycle	36.5	8.11%	111.5	3.98%	0.001044
monosaccharide metabolism	17	3.78%	36.5	1.30%	0.001062
mRNA processing	3	0.67%	91	3.25%	0.001145
hexose metabolism	16	3.56%	33.5	1.20%	0.001406
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	47.5	10.56%	164.5	5.87%	0.001616
response to chemical stimulus	31	6.89%	96	3.43%	0.001744
small nuclear ribonucleoprotein complex	0	0.00%	46	1.64%	0.001885
DNA binding	28.5	6.33%	85	3.03%	0.002318
transcription factor activity	10	2.22%	17	0.61%	0.002444
transport vesicle	9.5	2.11%	13.5	0.48%	0.002578
regulation of transcription	42.5	9.44%	147	5.25%	0.002614
mitochondrial envelope	11.5	2.56%	173	6.17%	0.002733
plasma membrane	25	5.56%	74.5	2.66%	0.003303
bud neck	18	4.00%	47	1.68%	0.003468
response to abiotic stimulus	38	8.44%	133	4.75%	0.003509
cell cycle	55	12.22%	211	7.53%	0.003532
phosphotransferase activity, alcohol group as acceptor	25	5.56%	76	2.71%	0.003606
organelle lumen	40.5	9.00%	413.5	14.76%	0.003963
membrane-enclosed lumen	40.5	9.00%	413.5	14.76%	0.003964
mitochondrion	58.5	13.00%	554.5	19.79%	0.004202
kinase activity	28.5	6.33%	89	3.18%	0.004268
bud	22	4.89%	64	2.28%	0.004297
polysome	4	0.89%	2	0.07%	0.004511
positive regulation of gene expression, epigenetic	4	0.89%	1.5	0.05%	0.004512
loss of chromatin silencing	4	0.89%	1.5	0.05%	0.004513
regulation of translational fidelity	4	0.89%	2	0.07%	0.004515
progressive alteration of chromatin during cell aging	4	0.89%	1.5	0.05%	0.004516
translation elongation factor activity	4	0.89%	2	0.07%	0.004517
Rho GTPase activator activity	4	0.89%	2	0.07%	0.004518
development	51	11.33%	194.5	6.94%	0.004553
Golgi membrane	11	2.44%	23	0.82%	0.005099
specific RNA polymerase II transcription factor activity	8	1.78%	12.5	0.45%	0.005396
vesicle coat	8	1.78%	13	0.46%	0.005397
alcohol metabolism	23.5	5.22%	67.5	2.41%	0.005429
bud tip	10.5	2.33%	20	0.71%	0.00588
enzyme regulator activity	27	6.00%	89	3.18%	0.006631
ribosome biogenesis	8	1.78%	127.5	4.55%	0.006799
macromolecule biosynthesis	65	14.44%	268.5	9.58%	0.006933

antioxidant activity	5	1.11%	5	0.18%	0.007282
phosphatase regulator activity	5	1.11%	5	0.18%	0.007284
protein phosphatase regulator activity	5	1.11%	5	0.18%	0.007286
GTPase activator activity	8	1.78%	13.5	0.48%	0.007469
cytoplasmic vesicle membrane	8	1.78%	14	0.50%	0.007471
vesicle membrane	8	1.78%	14	0.50%	0.007473
coated vesicle membrane	8	1.78%	14	0.50%	0.007475
spliceosome complex	1	0.22%	52	1.86%	0.007565
positive regulation of transcription from RNA polymerase II promoter	9	2.00%	17.5	0.62%	0.008829
regulation of mitosis	9	2.00%	17.5	0.62%	0.008831
cell wall glycoprotein biosynthesis	4	0.89%	3	0.11%	0.009425
cell wall mannoprotein biosynthesis	4	0.89%	3	0.11%	0.009427
mannoprotein biosynthesis	4	0.89%	3	0.11%	0.00943
mannoprotein metabolism	4	0.89%	3	0.11%	0.009432
age-dependent general metabolic decline	4	0.89%	3	0.11%	0.009434
mitochondrial membrane	10.5	2.33%	151.5	5.41%	0.009475
signal transduction	24.5	5.44%	79	2.82%	0.009812
regulation of glycolysis	3	0.67%	1	0.04%	0.009867
rDNA binding	3	0.67%	1	0.04%	0.00987
RNA splicing, via transesterification reactions	3	0.67%	71.5	2.55%	0.010173
major (U2-dependent) spliceosome	0	0.00%	34	1.21%	0.010942
reproductive physiological process	27	6.00%	93	3.32%	0.011191
reproductive cellular physiological process	27	6.00%	93	3.32%	0.011194
monosaccharide catabolism	7.5	1.67%	12	0.43%	0.011288
sphingolipid metabolism	7.5	1.67%	12	0.43%	0.011291
vacuolar transport	1	0.22%	49	1.75%	0.011339
translational elongation	5	1.11%	6	0.21%	0.011905
mRNA catabolism, deadenylylation-dependent decay	5	1.11%	6	0.21%	0.011908
nuclear lumen	27.5	6.11%	288	10.28%	0.012326
ribosome	33	7.33%	120	4.28%	0.012517
cell wall	11.5	2.56%	24.5	0.87%	0.012622
external encapsulating structure	11.5	2.56%	24.5	0.87%	0.012625
cell wall (sensu Fungi)	11.5	2.56%	24.5	0.87%	0.012629
nucleoplasm	13	2.89%	164	5.85%	0.012925
cell communication	26	5.78%	88	3.14%	0.013328
membrane coat	8	1.78%	16	0.57%	0.013386
coated membrane	8	1.78%	16	0.57%	0.013389
rRNA processing	6	1.33%	101.5	3.62%	0.014194
nuclear mRNA splicing, via spliceosome	3	0.67%	68	2.43%	0.014494
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3	0.67%	69	2.46%	0.014541
regulation of transcription, DNA-dependent	37.5	8.33%	141.5	5.05%	0.015548
carbohydrate kinase activity	4.5	1.00%	4	0.14%	0.016912
regulation of cyclin dependent protein kinase activity	4	0.89%	4	0.14%	0.016917
glucose catabolism	6.5	1.44%	10	0.36%	0.017169
hexose catabolism	6.5	1.44%	10	0.36%	0.017174
carbohydrate catabolism	8.5	1.89%	17	0.61%	0.01736
cellular carbohydrate catabolism	8.5	1.89%	17	0.61%	0.017364
actin cortical patch	8	1.78%	17	0.61%	0.017369
organellar large ribosomal subunit	0	0.00%	32	1.14%	0.01747

mitochondrial large ribosomal subunit	0	0.00%	32	1.14%	0.017475
rRNA metabolism	7	1.56%	105.5	3.77%	0.017625
hydrolase activity, hydrolyzing O-glycosyl compounds	5	1.11%	7	0.25%	0.018206
regulation of mRNA stability	5	1.11%	7	0.25%	0.01821
glycolysis	5.5	1.22%	7	0.25%	0.018215
regulation of RNA stability	5	1.11%	7	0.25%	0.01822
cytosolic large ribosomal subunit (sensu Eukaryota)	11	2.44%	28	1.00%	0.018383
transferase activity, transferring hexosyl groups	13.5	3.00%	35	1.25%	0.01902
regulation of endocytosis	2	0.44%	0	0.00%	0.02
protein phosphatase inhibitor activity	2	0.44%	0	0.00%	0.020005
positive regulation of glycolysis	2	0.44%	0	0.00%	0.02001
ligase activity, forming carbon-carbon bonds	2	0.44%	0	0.00%	0.020016
proton-transporting ATP synthase, catalytic core (sensu Eukaryota)	2	0.44%	0	0.00%	0.020021
proton-transporting ATP synthase, catalytic core	2	0.44%	0	0.00%	0.020026
protein desumoylation	2	0.44%	0	0.00%	0.020031
eukaryotic translation elongation factor 1 complex	2	0.44%	0	0.00%	0.020036
re-entry into mitotic cell cycle	2.5	0.56%	0	0.00%	0.020042
glutathione peroxidase activity	2	0.44%	0	0.00%	0.020047
ubiquitin-like-protein-specific protease activity	2	0.44%	0	0.00%	0.020052
re-entry into mitotic cell cycle after pheromone arrest	2.5	0.56%	0	0.00%	0.020057
SUMO-specific protease activity	2	0.44%	0	0.00%	0.020062
phosphatase inhibitor activity	2	0.44%	0	0.00%	0.020068
1,3-beta-glucan synthase complex	2	0.44%	0	0.00%	0.020073
protein biosynthesis	57.5	12.78%	242.5	8.65%	0.020323
alcohol catabolism	7.5	1.67%	13.5	0.48%	0.020497
site of polarized growth	21	4.67%	68.5	2.44%	0.020629
glycoprotein biosynthesis	13.5	3.00%	36	1.28%	0.020783
reproduction	33.5	7.44%	127	4.53%	0.020855
response to stimulus	62.5	13.89%	268.5	9.58%	0.021549
programmed cell death	3	0.67%	2	0.07%	0.02246
loss of chromatin silencing during replicative cell aging	3	0.67%	1.5	0.05%	0.022466
apoptosis	3	0.67%	2	0.07%	0.022472
carbohydrate transporter activity	3	0.67%	2	0.07%	0.022477
progressive alteration of chromatin during replicative cell aging	3	0.67%	1.5	0.05%	0.022483
response to reactive oxygen species	3	0.67%	2	0.07%	0.022489
glycoprotein metabolism	13.5	3.00%	37	1.32%	0.022938
small GTPase regulator activity	10	2.22%	24.5	0.87%	0.024216
actin filament organization	10.5	2.33%	24.5	0.87%	0.024223
intracellular signaling cascade	17	3.78%	54.5	1.95%	0.026041
regulation of RNA metabolism	5	1.11%	8	0.29%	0.026491
tRNA modification	0	0.00%	28	1.00%	0.026764
spindle checkpoint	4	0.89%	5	0.18%	0.027464
chronological cell aging	4.5	1.00%	5	0.18%	0.027471
nuclear nucleosome	4	0.89%	5	0.18%	0.027478
mitotic spindle checkpoint	4	0.89%	5	0.18%	0.027486
nucleosome	4	0.89%	5	0.18%	0.027493
mitotic checkpoint	4	0.89%	5	0.18%	0.0275
RNA splicing	4.5	1.00%	82	2.93%	0.027527

GTPase regulator activity	12	2.67%	32.5	1.16%	0.028193
DNA-directed RNA polymerase II, holoenzyme	2	0.44%	54	1.93%	0.029519
condensed chromosome	2	0.44%	53.5	1.91%	0.029527
protein kinase activity	18.5	4.11%	59	2.11%	0.030446
endocytosis	12	2.67%	33.5	1.20%	0.030515
response to stress	47	10.44%	199.5	7.12%	0.030805
budding cell bud growth	6	1.33%	12	0.43%	0.031601
non-developmental growth	6	1.33%	12	0.43%	0.03161
cysteine-type peptidase activity	6	1.33%	12	0.43%	0.031618
signal transducer activity	10.5	2.33%	26.5	0.95%	0.031816
growth	18	4.00%	59.5	2.12%	0.031846
biopolymer glycosylation	12	2.67%	35	1.25%	0.033459
protein amino acid glycosylation	12	2.67%	35	1.25%	0.033467
enzyme activator activity	12	2.67%	35	1.25%	0.033476
endomembrane system	38	8.44%	156	5.57%	0.035082
cellular lipid metabolism	29	6.44%	112	4.00%	0.036473
small GTPase mediated signal transduction	9	2.00%	23	0.82%	0.036688
regulation of protein kinase activity	5	1.11%	9	0.32%	0.036818
regulation of kinase activity	5	1.11%	9	0.32%	0.036828
COPI-coated vesicle	5	1.11%	9	0.32%	0.036838
cyclin-dependent protein kinase regulator activity	5	1.11%	9	0.32%	0.036847
regulation of transferase activity	5	1.11%	9	0.32%	0.036857
RNA modification	1	0.22%	38	1.36%	0.037176
sporulation	16	3.56%	53.5	1.91%	0.039203
age-dependent response to oxidative stress	3	0.67%	3	0.11%	0.040754
age-dependent general metabolic decline during chronological cell aging	3	0.67%	3	0.11%	0.040765
age-dependent response to oxidative stress during chronological cell aging	3	0.67%	3	0.11%	0.040776
regulation of translation	6.5	1.44%	12.5	0.45%	0.041094
regulation of protein biosynthesis	6.5	1.44%	12.5	0.45%	0.041105
ER-associated protein catabolism	6	1.33%	12.5	0.45%	0.041115
tRNA metabolism	4	0.89%	71	2.53%	0.041167
biosynthesis	93	20.67%	443.5	15.83%	0.041357
condensed nuclear chromosome	2	0.44%	49.5	1.77%	0.041625
biopolymer methylation	0	0.00%	25	0.89%	0.042012
mitochondrial small ribosomal subunit	0	0.00%	26	0.93%	0.04285
organellar small ribosomal subunit	0	0.00%	26	0.93%	0.042862
outer membrane	3	0.67%	60.5	2.16%	0.042875
organelle outer membrane	3	0.67%	60.5	2.16%	0.042886
mitochondrial outer membrane	3	0.67%	60.5	2.16%	0.042897
lipid metabolism	30	6.67%	120	4.28%	0.043283
main pathways of carbohydrate metabolism	11.5	2.56%	31.5	1.12%	0.044875
cellular polysaccharide metabolism	8.5	1.89%	18.5	0.66%	0.046224
translation factor activity, nucleic acid binding	8	1.78%	19	0.68%	0.046236
polysaccharide metabolism	8.5	1.89%	18.5	0.66%	0.046248
actin cytoskeleton organization and biogenesis	15.5	3.44%	50	1.78%	0.047829
nucleic acid binding	54.5	12.11%	242.5	8.65%	0.048341

SI Table 3. All Gene Ontology (GO) terms that are significantly under- or over-represented among loci retained in duplicate since the WGD in *S. cerevisiae* relative to single-copy genes.

Gene Ontology Term	Ohnologs		Singletons		Corrected P-value
	Count	Percentage	Count	Percentage	
cytosolic ribosome (sensu Eukaryota)	42.5	9.82%	35	1.24%	4.83E-17
cytosol	65	15.01%	114	4.04%	6.31E-14
cytosolic large ribosomal subunit (sensu Eukaryota)	23	5.31%	16	0.57%	4.78E-11
eukaryotic 48S initiation complex	19	4.39%	12	0.43%	8.77E-10
cytosolic small ribosomal subunit (sensu Eukaryota)	19	4.39%	12	0.43%	8.77E-10
structural constituent of ribosome	42	9.70%	81	2.87%	1.33E-08
ribosome	47	10.85%	106	3.76%	6.46E-08
eukaryotic 43S preinitiation complex	19	4.39%	20	0.71%	1.41E-07
RNA processing	7	1.62%	221	7.84%	4.53E-07
organelle lumen	27	6.24%	427	15.15%	2.06E-06
membrane-enclosed lumen	27	6.24%	427	15.15%	2.06E-06
RNA metabolism	15.5	3.58%	302	10.71%	4.89E-06
ribosome biogenesis	2.5	0.58%	133	4.72%	1.84E-05
macromolecule biosynthesis	74.5	17.21%	259	9.19%	3.16E-05
phosphotransferase activity, alcohol group as acceptor	29	6.70%	72	2.55%	5.69E-05
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	0	0.00%	72	2.55%	5.91E-05
mRNA processing	1	0.23%	93	3.30%	5.91E-05
biosynthesis	108.5	25.06%	428	15.18%	5.98E-05
cellular carbohydrate metabolism	30	6.93%	74	2.63%	6.15E-05
protein kinase activity	24.5	5.66%	53	1.88%	6.23E-05
large ribosomal subunit	23	5.31%	48	1.70%	6.25E-05
carbohydrate metabolism	31.5	7.27%	81	2.87%	7.33E-05
structural molecule activity	50	11.55%	161	5.71%	8.28E-05
cellular biosynthesis	98.5	22.75%	384	13.62%	8.49E-05
nuclear lumen	18.5	4.27%	297	10.54%	8.57E-05
kinase activity	32.5	7.51%	85	3.02%	9.35E-05
nuclear mRNA splicing, via spliceosome	0	0.00%	71	2.52%	9.75E-05
small ribosomal subunit	19	4.39%	38	1.35%	0.000109
cell wall organization and biogenesis	25.5	5.89%	60	2.13%	0.000151
external encapsulating structure organization and biogenesis	25.5	5.89%	60	2.13%	0.000151
plasma membrane	27.5	6.35%	72	2.55%	0.000286
nucleoplasm	8	1.85%	169	6.00%	0.000295
mitochondrial ribosome	0	0.00%	61	2.16%	0.000347
organellar ribosome	0	0.00%	61	2.16%	0.000347
rRNA processing	2.5	0.58%	105	3.72%	0.000384
protein biosynthesis	63	14.55%	237	8.41%	0.000471
cell wall	13	3.00%	23	0.82%	0.000528
external encapsulating structure	13	3.00%	23	0.82%	0.000528
cell wall (sensu Fungi)	13	3.00%	23	0.82%	0.000528
biopolymer biosynthesis	7.5	1.73%	6	0.21%	0.000641
polysaccharide biosynthesis	7.5	1.73%	6	0.21%	0.000642
RNA splicing, via transesterification reactions	0.5	0.12%	74	2.63%	0.000704
mRNA metabolism	5	1.15%	124	4.40%	0.000715

protein amino acid phosphorylation	18	4.16%	41	1.45%	0.000719
spliceosome complex	0	0.00%	53	1.88%	0.000811
protein serine/threonine kinase activity	14.5	3.35%	28	0.99%	0.000828
rRNA metabolism	3.5	0.81%	109	3.87%	0.000929
biopolymer metabolism	91.5	21.13%	883	31.32%	0.001091
cyclin-dependent protein kinase regulator activity	7	1.62%	7	0.25%	0.001136
phosphorylation	23.5	5.43%	63	2.23%	0.00116
signal transduction	26.5	6.12%	77	2.73%	0.001247
energy reserve metabolism	8.5	1.96%	10	0.35%	0.001319
glycogen biosynthesis	4.5	1.04%	1	0.04%	0.001449
glucan biosynthesis	5.5	1.27%	3	0.11%	0.001714
regulation of cyclin dependent protein kinase activity	5	1.15%	3	0.11%	0.001715
cellular polysaccharide metabolism	10	2.31%	17	0.60%	0.001824
polysaccharide metabolism	10	2.31%	17	0.60%	0.001824
35S primary transcript processing	0	0.00%	49	1.74%	0.002013
regulation of cell redox homeostasis	3.5	0.81%	0	0.00%	0.002424
cell redox homeostasis	3.5	0.81%	0	0.00%	0.002425
glucan metabolism	8	1.85%	12	0.43%	0.002963
small nuclear ribonucleoprotein complex	0	0.00%	46	1.63%	0.003123
transferase activity, transferring phosphorus-containing groups	38.5	8.89%	137	4.86%	0.003191
cell communication	27	6.24%	87	3.09%	0.003253
RNA splicing	2.5	0.58%	84	2.98%	0.00352
nucleus	116.5	26.91%	1047	37.14%	0.003801
alcohol metabolism	22	5.08%	69	2.45%	0.005108
regulation of cellular process	69	15.94%	297	10.54%	0.005213
regulation of cellular physiological process	69	15.94%	297	10.54%	0.005215
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	84	19.40%	772	27.39%	0.005518
carbohydrate biosynthesis	11.5	2.66%	25	0.89%	0.006134
response to abiotic stimulus	36	8.31%	135	4.79%	0.006136
reproductive cellular physiological process	27	6.24%	93	3.30%	0.006547
reproductive physiological process	27	6.24%	93	3.30%	0.006549
endocytosis	13.5	3.12%	32	1.14%	0.00657
regulation of transferase activity	6	1.39%	8	0.28%	0.006723
glycogen metabolism	6.5	1.50%	8	0.28%	0.006725
regulation of protein kinase activity	6	1.39%	8	0.28%	0.006727
regulation of kinase activity	6	1.39%	8	0.28%	0.006728
transcription factor activity	9	2.08%	18	0.64%	0.006842
RNA splicing factor activity, transesterification mechanism	0	0.00%	38	1.35%	0.007156
small GTPase mediated signal transduction	10	2.31%	22	0.78%	0.007418
regulation of physiological process	70	16.17%	308	10.93%	0.007517
phosphorus metabolism	26.5	6.12%	88	3.12%	0.007759
phosphate metabolism	26.5	6.12%	88	3.12%	0.007761
cytoplasm organization and biogenesis	9.5	2.19%	149	5.29%	0.007765
ribosome biogenesis and assembly	9.5	2.19%	149	5.29%	0.007767
condensed chromosome	0.5	0.12%	55	1.95%	0.007979
DNA-directed RNA polymerase II, holoenzyme	1	0.23%	55	1.95%	0.007981
regulation of biological process	71.5	16.51%	314	11.14%	0.00803
pyrimidine base metabolism	4	0.92%	3	0.11%	0.008174

UDP-glucosyltransferase activity	4	0.92%	3	0.11%	0.008177
chromosome	12	2.77%	167	5.92%	0.008494
enzyme regulator activity	26	6.00%	90	3.19%	0.008505
oxidoreductase activity, acting on the CH-CH group of donors, quinone or related compound as acceptor	3	0.69%	1	0.04%	0.0088
succinate dehydrogenase (ubiquinone) activity	3	0.69%	1	0.04%	0.008802
thiol-disulfide exchange intermediate activity	3	0.69%	1	0.04%	0.008804
intracellular membrane-bound organelle	233	53.81%	1903	67.51%	0.008919
membrane-bound organelle	233	53.81%	1903	67.51%	0.008921
ribonucleoprotein complex	51	11.78%	215	7.63%	0.00975
protein complex	92.5	21.36%	827	29.34%	0.009848
vacuolar transport	1	0.23%	49	1.74%	0.011299
condensed nuclear chromosome	0.5	0.12%	51	1.81%	0.011526
endomembrane system	14	3.23%	180	6.39%	0.011606
reproduction	33.5	7.74%	127	4.51%	0.013068
G1/S transition of mitotic cell cycle	8.5	1.96%	17	0.60%	0.013835
organelle organization and biogenesis	67	15.47%	614	21.78%	0.013866
mitochondrial lumen	6	1.39%	104	3.69%	0.014064
mitochondrial matrix	6	1.39%	104	3.69%	0.014068
intracellular signaling cascade	17.5	4.04%	54	1.92%	0.014292
DNA recombination	1.5	0.35%	63	2.23%	0.014337
bud tip	9.5	2.19%	21	0.74%	0.014462
lipid metabolism	31	7.16%	119	4.22%	0.014874
ribonucleotide biosynthesis	7	1.62%	14	0.50%	0.016603
response to chemical stimulus	27	6.24%	100	3.55%	0.017144
organellar large ribosomal subunit	0	0.00%	32	1.14%	0.017259
mitochondrial large ribosomal subunit	0	0.00%	32	1.14%	0.017263
major (U2-dependent) spliceosome	0	0.00%	34	1.21%	0.018314
ATP-dependent helicase activity	0	0.00%	34	1.21%	0.018318
septin ring assembly	2	0.46%	0	0.00%	0.01838
thioredoxin peroxidase activity	2	0.46%	0	0.00%	0.018385
glycogen synthase kinase 3 activity	2	0.46%	0	0.00%	0.018389
tRNA-pseudouridine synthase activity	2	0.46%	0	0.00%	0.018394
regulation of glycogen catabolism	2	0.46%	0	0.00%	0.018399
septin ring organization	2	0.46%	0	0.00%	0.018403
ligase activity, forming carbon-carbon bonds	2	0.46%	0	0.00%	0.018408
regulation of glycogen biosynthesis	2.5	0.58%	0	0.00%	0.018413
small GTPase regulator activity	10.5	2.42%	24	0.85%	0.019155
helicase activity	1.5	0.35%	58	2.06%	0.019815
transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	3	0.69%	2	0.07%	0.019981
pyrimidine base biosynthesis	3	0.69%	2	0.07%	0.019986
disulfide oxidoreductase activity	3	0.69%	2	0.07%	0.019991
organelle membrane	32.5	7.51%	333	11.81%	0.021585
protein kinase regulator activity	8	1.85%	19	0.67%	0.02237
glucosyltransferase activity	4	0.92%	5	0.18%	0.023811
proteolysis	7	1.62%	106	3.76%	0.023904
covalent chromatin modification	1	0.23%	43	1.53%	0.024994
chromosome, pericentric region	0.5	0.12%	43	1.53%	0.025001
histone modification	1	0.23%	43	1.53%	0.025007
regulation of progression through cell cycle	19.5	4.50%	67	2.38%	0.02567

regulation of cell cycle	19.5	4.50%	67	2.38%	0.025676
interphase of mitotic cell cycle	13.5	3.12%	40	1.42%	0.025728
interphase	13.5	3.12%	40	1.42%	0.025735
monosaccharide metabolism	13.5	3.12%	40	1.42%	0.025741
regulation of enzyme activity	6.5	1.50%	12	0.43%	0.026287
signal transducer activity	10	2.31%	27	0.96%	0.026421
phosphoric monoester hydrolase activity	12	2.77%	34	1.21%	0.027044
protein amino acid acetylation	0	0.00%	30	1.06%	0.027405
nuclear envelope-endoplasmic reticulum network	4.5	1.04%	85	3.02%	0.02766
proteolysis during cellular protein catabolism	5	1.15%	86	3.05%	0.027671
ribonucleotide metabolism	7	1.62%	16	0.57%	0.027761
meiotic recombination	0	0.00%	31	1.10%	0.02868
transcription factor complex	3.5	0.81%	77	2.73%	0.029567
protein serine/threonine phosphatase activity	5	1.15%	9	0.32%	0.031296
phosphoric ester hydrolase activity	12	2.77%	37	1.31%	0.034093
hexose metabolism	12.5	2.89%	37	1.31%	0.034102
development	45.5	10.51%	200	7.09%	0.034667
DNA metabolism	28	6.47%	283	10.04%	0.034691
nucleolus	11	2.54%	139	4.93%	0.035721
Ras protein signal transduction	4.5	1.04%	6	0.21%	0.035763
antioxidant activity	4	0.92%	6	0.21%	0.035772
oxidoreductase activity, acting on the CH-CH group of donors	4	0.92%	6	0.21%	0.035782
actin cap	4	0.92%	6	0.21%	0.035791
regulation of translational fidelity	3	0.69%	3	0.11%	0.036319
response to salt stress	3	0.69%	3	0.11%	0.036329
translation elongation factor activity	3	0.69%	3	0.11%	0.036338
mitochondrial transport	3	0.69%	3	0.11%	0.036348
kinetochore	0.5	0.12%	40	1.42%	0.037222
ubiquitin-dependent protein catabolism	5	1.15%	84	2.98%	0.038345
modification-dependent protein catabolism	5	1.15%	84	2.98%	0.038355
cytoplasm	311	71.82%	1713	60.77%	0.0398
cortical cytoskeleton	9.5	2.19%	25	0.89%	0.040261
cortical actin cytoskeleton	9.5	2.19%	25	0.89%	0.040272
nuclear chromosome	12	2.77%	144	5.11%	0.040654
methyltransferase activity	1.5	0.35%	52	1.84%	0.04162
mitochondrial small ribosomal subunit	0	0.00%	26	0.92%	0.042086
organellar small ribosomal subunit	0	0.00%	26	0.92%	0.042097
ubiquitin ligase complex	0	0.00%	26	0.92%	0.042108
growth	17.5	4.04%	60	2.13%	0.042478
transferase activity, transferring one-carbon groups	1.5	0.35%	53	1.88%	0.042772
purine ribonucleotide biosynthesis	6	1.39%	14	0.50%	0.043797
specific RNA polymerase II transcription factor activity	6.5	1.50%	14	0.50%	0.043808
generation of precursor metabolites and energy	25.5	5.89%	99	3.51%	0.045675
energy derivation by oxidation of organic compounds	22.5	5.20%	86	3.05%	0.046304
cellular protein catabolism	5.5	1.27%	89	3.16%	0.046985

SI Appendix

Section 1. Notes on the gene content of *K. polysporus*.

Mating type loci: The life cycle of *K. polysporus* has been described in detail (15, 16). It is homothallic, and we identified a homolog (*Kpol_1054.32*) of the *HO* endonuclease gene, which catalyzes mating-type switching in *S. cerevisiae*. *K. polysporus* has been reported to grow primarily as a haploid (zygotes do not bud but instead sporulate soon after formation) (16), but our sequenced isolate was either diploid or contained a mixture of *MAT***a** and *MAT*α haploid cells. We identified eight clones in our fosmid library with ~40 kb inserts spanning the *MAT* locus (in supercontig s9; SI Figure 6), of which five contained a *MAT***a** allele and three contained a *MAT*α allele, as determined by sequencing the fosmids with a primer flanking the *MAT* locus. We completely sequenced the inserts in one *MAT*α fosmid (fos_37c10) and one *MAT***a** fosmid (fos_72a08) and found that they had no sequence differences other than the α-specific and **a**-specific "Y" regions of the *MAT* locus. Unusually, the *K. polysporus* genome sequence includes three silent copies of mating-type information: two *HMR***a**-like loci (in supercontigs s8 and s23) and one *HML*α-like locus (in supercontig s9, 100 kb from the *MAT* locus). Like *Candida glabrata* (17), the genome of *K. polysporus* does not contain a homolog of the *S. cerevisiae* silencing gene *SIR1*, although *SIR2*, *SIR3* and *SIR4* homologs are present. (The *K. polysporus* ohnolog pair *Kpol_1032.18* and *Kpol_479.28* corresponds to the *S. cerevisiae* ohnolog pair *SIR2* and *HST1*; the pair *Kpol_1001.11* and *Kpol_520.35* corresponds to the pair *SIR3* and *ORC1*; *Kpol_269.1* is an ortholog of *SIR4*.)

Genes for pheromones and their receptors: *K. polysporus* has two copies (ohnologs) of the α-pheromone gene. One copy (*Kpol_1002.67*) codes for five identical repeats of the peptide WHWLELDNGQPIY, and the other (*Kpol_1033.32*) codes for four identical repeats of the peptide WHWLRLRYGEPIY. The 9/13 amino acid match between these two putative pheromone peptides is surprisingly low. Interestingly, *K. polysporus* retains two ohnolog copies of the *STE2* α-pheromone receptor (*Kpol_1011.19* and *Kpol_1058.22*), so it is possible that there are two separately interacting pheromone/receptor pairs in this species. The only **a**-pheromone genes in *K. polysporus* (*Kpol_1039.70*, *Kpol_1039.70a*, and *Kpol_1039.70b*) are in a triple tandem repeat at a locus that is in a paralogous relationship (reciprocal gene loss after WGD) with *S. cerevisiae MFA2*. *K. polysporus* retains a single ortholog of the *STE3* **a**-factor receptor gene (*Kpol_2001.38*).

Subtelomeric regions: The subtelomeric regions of the *K. polysporus* genome contain multiple genes (at least 19 copies) for exo-1,3-beta-glucanase, an enzyme that degrades the cell wall polymer beta-glucan. In *S. cerevisiae* there are only three exo-1,3-beta-glucanase genes (*SPR1*, *EXG1* and *EXG2*), and they function in cell wall assembly and spore wall morphogenesis (18, 19). The amplification of this family in *K. polysporus* is possibly related to its multi-spored phenotype.

Protein Complexes: Protein complexes and genes coding for their components tend to be lost and gained relatively rarely during evolution. However, we noticed that the genes coding for all three subunits (*SSY1*, *SSY5* and *PTR3*) of the SPS extracellular amino acid sensor system (20), and several subunits of dynein and dynactin (discussed in main text) are absent from the genome of *K. polysporus*, as are genes for enzymes of the DAL pathway (*DAL1*, *DAL2*, *DAL3*, *DAL4*, *DAL7* and *DCG1*; these are not known to form a complex) (21). In addition, six (*SFB3*, *SEC13*, *SEC16*, *SEC23*, *SEC31* and *SEC24*/*SFB2*) of the seven genes coding for subunits of the COPII vesicle complex are retained as ohnolog pairs in *K. polysporus*. Only *SEC24*/*SFB2* is present in duplicate in *S. cerevisiae* and *SAR1* is duplicated in neither species. COPII proteins coat and direct the formation of vesicles that transport proteins from the ER to the golgi and may also have a role in 'cargo' protein selection (22). Genes coding for COPII subunits are evolutionarily well conserved and most have single orthologs in mammals (22). Three interacting subunits of the F_1 portion of the mitochondrial F_1F_0-ATPase (*ATP1*, *ATP2* and *ATP5*) have also been retained as ohnolog pairs in *K. polysporus* but not in other post-WGD yeasts.

Species-specific genes: The *K. polysporus* genome contains some multicopy gene families that have no homologs in other yeasts. A similar situation exists in *S. castellii* (23). Representative members of *K. polysporus*-specific families are *Kpol_489.2* and *Kpol_1035.52*. Other *K. polysporus* gene families, such as those represented by *Kpol_387.6* and *Kpol_487.8*, lack homologs in *S. cerevisiae* but are also multigene families in other yeasts such as *S. castellii* or *C. glabrata*. None of these genes have functionally characterized homologs in any other organism. We also noticed that *K. polysporus* has a gene (*Kpol_520.25*) coding for a protein in the Argonaute family. Argonaute proteins bind small RNAs and usually function in gene silencing. Although present in most eukaryotes, including the filamentous euascomycetes and *Schizosaccharomyces pombe*, there are no Argonaute homologs in *S. cerevisiae*. The *K. polysporus* Argonaute gene has a WGD-derived paralog in *S. castellii* (*Scas_719.65)* but not in any of the other species (post-WGD or pre-WGD) in YGOB. There is also an Argonaute homolog in *C. albicans* (24).

Transposable elements: We identified at least 39 LTR (long terminal repeat) retrotransposons, similar to the Ty elements of *S. cerevisiae.* The exact number of retroelements is uncertain because many of them cause gaps between contigs. We named the elements Tkp1, Tkp3, Tkp4 and Tkp5, following the nomenclature of ref. (25), of which the most common type of solo LTR is Tkp5. Although most retroelements are inserted near tRNA or rRNA genes or in telomeric regions, there are two cases where a Tkp5 element interrupts an otherwise intact protein coding gene (*Kpol_1036.28* and *Kpol_2000.48*), suggesting that the insertions are recent and that Tkp5 is an active element.

Section 2. Measuring the effect of the ortholog-paralog bias in YGOB's tracking algorithm.

YGOB uses an algorithm based on shared gene content in a local (41 locus) sliding window to assign orthology of the sister genomic regions (tracks) among different post-WGD species (4), but the high levels of independent gene loss that have occurred between *K. polysporus* and the other post-WGD yeasts make this assignment difficult in most parts of the genome. In the region shown in Figure 1, for example, there are two places where YGOB's algorithm 'changes its mind' about how orthology and paralogy are assigned between *K. polysporus* and *S. cerevisiae* chromosomes. We refer to the process of identifying orthologous chromosomal regions between species as 'tracking'.

In the whole-genome comparison of the 3252 ancestral loci that could be reliably scored as present or absent in both *K. polysporus* and *S. cerevisiae*, YGOB scored 44.7% of loci as single-copy orthologs and 34.6% as single-copy paralogs (reciprocal gene losses) (Table 1). Because YGOB's algorithm works on the principle that orthologous regions should have higher similarity of gene content than paralogous regions, and because it operates on a local window, it has a built-in bias that will cause it to overestimate the number of orthologs in situations where the true numbers of orthologs and paralogs are similar.

We measured the effect of this bias by using the YGOB engine to create and score 100 *K. polysporus* pseudo-genomes in which any possible signal of shared ancestry with *S. cerevisiae* was obliterated. While scoring the real *K. polysporus* genome against the ancestral gene order ('Real genome' columns in SI Table 4) we created 100 pseudo-genomes where at every locus with a syntenic *K. polysporus* presence on one track and a syntenic *K. polysporus* absence on the other track, we swapped the syntenic gene from its chromosome into the syntenic gap in the chromosome on the other track with a probability of 0.5. This procedure means that the pseudo-genomes must, on average, contain equal numbers of orthologs and paralogs of the *S. cerevisiae* single-copy genes. We then used the YGOB engine to score these 100 pseudo-genomes, calculating a mean and standard deviation for each locus class (SI Table 4). As would be expected due to the randomizations' breaking of chromosomes into smaller syntenic fragments, the number of scoreable loci in the pseudo-genomes is less than in the real genome. Nevertheless the average proportions of single-copy orthologs (43.42% ± s.d. 2.23%) and paralogs (33.80% ± s.d. 2.64%) reported in the pseudo-genomes are the same as in the real data, instead of being equal to each other.

Thus, the reported excess of orthologs over paralogs in Table 1 may be due to YGOB's bias towards reporting orthologs. These results fail to reject the null hypothesis of no shared gene losses on the phylogenetic branch between the WGD and the common ancestor of *K. polysporus* and *S. cerevisiae*, such as would occur if they had undergone completely independent WGD events. However, modeling gene losses using a likelihood approach does reveal a signal of shared ancestry (*SI Appendix*, section 5).

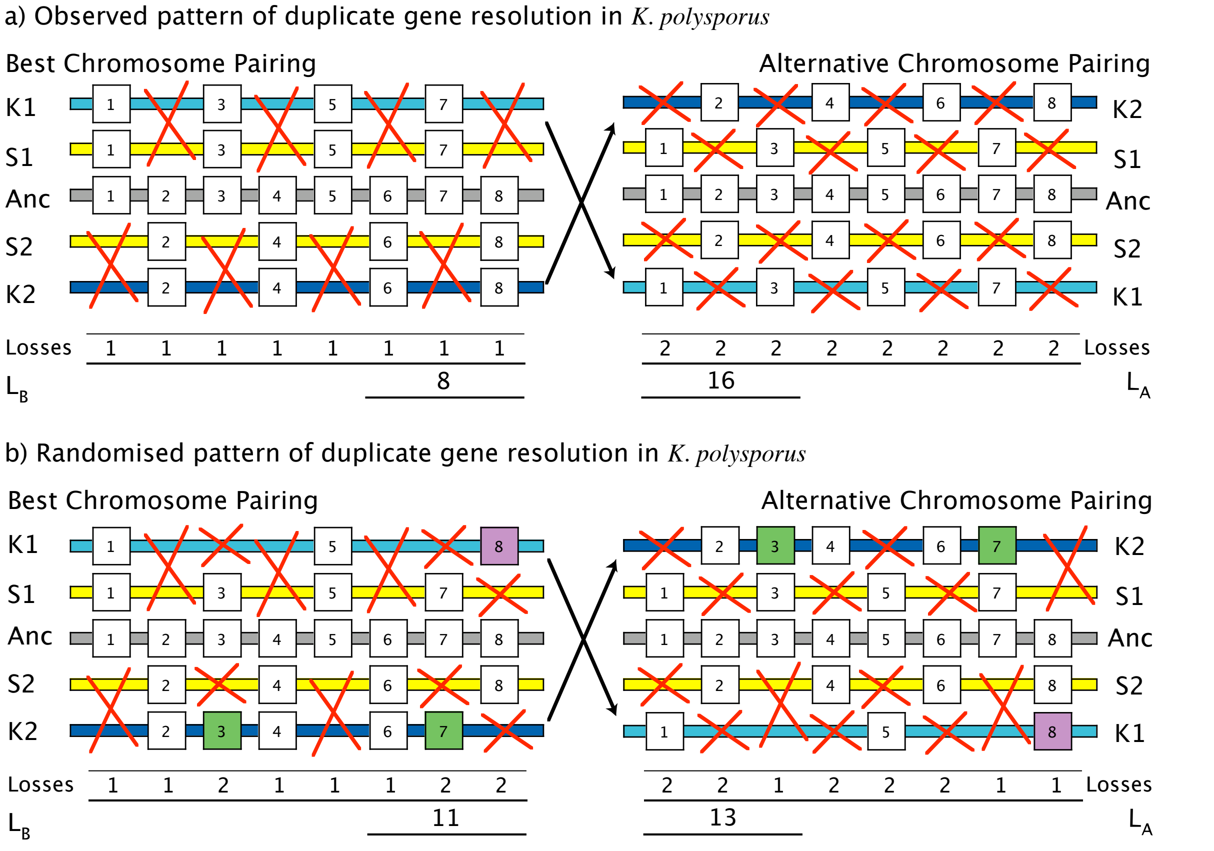
SI Table 4. Percentages of loci in different retention classes between *S. cerevisiae* and the real *K. polysporus* genome, and in 100 pseudo-genomes where the tracking of *K. polysporus* single-copy genes was randomized.

Locus class (*K. pol.*:*S. cer.*)	Real genome		Pseudo-genomes			
			Number of loci		Percent	
	Number of loci	Percent	Mean	S.D.	Mean	S.D.
2:2	212	6.52%	209.17	1.81	7.60%	0.87%
2:1	238	7.32%	234.90	1.65	8.53%	0.70%
1:2	221	6.80%	183.27	5.00	6.66%	2.73%
1:1 orthologs	1455	44.74%	1195.72	26.65	43.42%	2.23%
1:1 paralogs	1126	34.62%	930.61	24.61	33.80%	2.64%
Total	3252		2753.67			
Proportion of paralogs among 1:1 loci		44%			44%	1%

Section 3. Relationship between the estimated fraction of paralogous single-copy genes, and the confidence of YGOB's orthologous track assignment between *K. polysporus* and *S. cerevisiae.*

Our estimate that 44.7% of single-copy loci in *K. polysporus* and *S. cerevisiae* are paralogs (Table 1) is based on scoring all 3252 ancestral loci that can be compared between the two species, using the YGOB engine (4). The accuracy of this estimate depends on the accuracy with which YGOB identifies, in any genomic region, the correct overall orthology and paralogy relationships among the two *K. polysporus* genomic tracks (K1 and K2 in SI Figure 8, below) and the two *S. cerevisiae* genomic tracks (S1 and S2). We refer to this identification process as 'tracking'. If the tracking of a particular genomic region is incorrect, individual single-copy loci within that region will be mis-called (orthologs will be misidentified as paralogs, and *vice versa*).

We were concerned that our estimate of the proportion of paralogs in the genome might be inflated by the inclusion of mis-tracked genomic regions in the analysis. However, using a heuristic measure of the confidence of tracking, we show below that there are few regions of the genome where the percentage of single-copy loci that are paralogs is less than 20%, and that the fraction of paralogs is at least 30% in the half of the genome that is most confidently tracked.


a) Observed pattern of duplicate gene resolution in K. polysporus
Best Chromosome Pairing
Alternative Chromosome Pairing
K1
S1
Anc
S2
K2
Losses 1 1 1 1 1 1 1 1
L_B 8
2 2 2 2 2 2 2 2 Losses
16 L_A
b) Randomised pattern of duplicate gene resolution in K. polysporus
Best Chromosome Pairing
Alternative Chromosome Pairing
Losses 1 1 2 1 1 1 2 2
L_B 11
2 2 1 2 2 2 1 1 Losses
13 L_A

SI Figure 8. Method for estimating confidence of orthologous track assignment. See text for details. 'Anc' represents the ancestral gene order before WGD.

We used YGOB to find pairs of homologous chromosomal segments in the genomes of both *S. cerevisiae* and *K. polysporus*, that have remained unrearranged since the WGD and where no sequence gaps exist in the *K. polysporus* assembly. We retrieved 98 such 'blocks' (a pair of contiguous homologous chromosomal segments from *S. cerevisiae* and the corresponding pair of regions from *K. polysporus*), ranging in length from 10 to 73 genes, and containing a total of 1765 ancestral loci.

For each block we considered the two possible orthologous chromosomal pairings between the *S. cerevisiae* and *K. polysporus* segments (*i.e*., S1 orthologous to K1 and S2 orthologous to K2, or S1 orthologous to K2 and S2 orthologous to K1). We counted the number of gene losses, *L*, required to account for the observed pattern of gene loss in each case. We assumed that all gene losses were of single genes (26) and that where a gene is missing from an orthologous locus (in the context of the pairing being considered) in both species, it was lost in the common ancestor. We refer to the chromosomal pairing that requires the fewest gene losses (L_B in SI Figure 8a) as the 'best' pairing and the other possible pairing as the 'alternative' pairing (which requires L_A losses). $D = L_A - L_B$ gives the number of loci that support the best pairing over the alternative pairing and has a value between 0 and the length of the block.

If there are many more single-copy orthologs (which can be explained by single gene losses in the common ancestor of *S. cerevisiae* and *K. polysporus*) in the best chromosomal pairing than in the alternative pairing, *D* is large and parsimony favors the best pairing as the true orthologous pairing (in the example in SI Figure 8a, $D = 8$). By contrast, if the numbers of single-copy orthologs in the best and alternative pairings are approximately equal, *D* will be close to zero and neither chromosomal pairing is well supported. We assigned significance to *D* by comparing the observed value of *D* for the best pairing (D_{Real}) to a null distribution obtained by calculating *D* for randomized blocks (D_{Rand}). Randomizations preserved the number of genes retained in each genome but randomized the pattern of duplicate gene resolution by reassigning genes from *K. polysporus* segment K1 to the paralogous locus on *K. polysporus* segment K2 with a probability of 0.5 (compare loci 3, 7, and 8 between panels **a** and **b** in SI Figure 8). The percentage of randomized datasets for which D_{Rand} is less than D_{Real} is a measure of our confidence that the best pairing reflects a correct assignment of orthologous tracks.

We found that orthologous chromosomes can be inferred with reasonable confidence in some regions of the genome, but that in others (even where relatively large contiguous regions exist in both *S. cerevisiae* and *K. polysporus*) the pattern of gene loss is not significantly different from that predicted by independent WGD events (*i.e.,* no shared history). For instance, although block 91 is 57 genes long, the best chromosome pairing requires only 3 fewer losses to explain than the alternative, which is better than only 25% of randomized datasets. By contrast, for block 43 (15 genes long) the best pairing involves 9 fewer losses than the alternative, which is better than 99% of randomizations.

We stratified blocks according to intervals of our confidence statistic (SI Table 5) and calculated the percentage of single-copy orthologs and single-copy paralogs in each stratum. The estimated proportion of orthologs decreases as the tracking confidence decreases. This is as expected, because a block with a high content of orthologs should be easy to track. No matter what the average proportion of orthologs is across the whole genome, we would expect there to be some regional variation (purely by chance) resulting in some blocks with confident tracking and high ortholog content, and other blocks with lower tracking confidence and lower ortholog content.

SI Table 5 indicates that, even in the most confidently-tracked blocks in the genome (containing 12.7% of the studied loci), 17.4% of single-copy loci are paralogs between *K. polysporus* and *S. cerevisiae*. Among the best-tracked 55.8% of loci (the top four strata), the estimated fraction of paralogs is 31.7%. Similar to YGOB's estimate for the whole genome (Table 1), we estimate that among all 98 blocks considered here the proportion of single-copy loci that are paralogs is 38.9%.

SI Table 5. Estimated proportions of orthologous and paralogous loci between *K. polysporus* and *S. cerevisiae*, in 98 genomic blocks stratified according to confidence of track assignment.

Tracking confidence percentile	Number of blocks	Total number of loci	Number of valid single-copy loci	Number of single-copy orthologs	Number of single-copy paralogs	Proportion of orthologs (%)	Cumulative proportion of orthologs (%)*	Cumulative proportion of paralogs (%)*	Cumulative proportion of loci (%)*
81-100	12	225	109	90	19	82.6	82.6	17.4	12.7
61-80	17	274	134	96	38	71.6	76.5	23.5	28.3
41-60	10	170	82	55	27	67.1	74.2	25.8	37.9
21-40	17	315	155	87	68	56.1	68.3	31.7	55.8
1-20	10	228	107	58	49	54.2	65.8	34.2	68.7
0	32	553	298	155	143	52.0	61.1	38.9	100.0

* Cumulative proportions calculated across the confidence percentile intervals 81-100%, 61-100%, 41-100%, 21-100%, 1-100% and 0-100%.

Section 4. Calculating the expected number of shared ohnolog pairs between *S. cerevisiae* and *K. polysporus*

The high level of paralogy (~44.7%) among genes that are single-copy in both *S. cerevisiae* and *K. polysporus* indicates that the fates of most duplicated loci were not determined at the time of divergence of these two species. Indeed, our model indicates that 79% of loci were still duplicated and in the **U** ('undecided') state at this time (Fig. 2; *SI Appendix,* section 5). Since 47% of loci that are currently duplicated in *K. polysporus* are also present in duplicate in *S. cerevisiae* (212 of 450, among the 3252 loci studied in Table 1), this suggests substantial convergent preservation of duplicates. We estimated the number of duplicate genes that were preserved convergently in two different ways.

Method 1: Assuming negligible shared ancestry

Because *S. cerevisiae* and *K. polysporus* diverged very soon after the WGD we estimated the number of loci that would be preserved in duplicate under the assumption of negligible shared ancestry (*i.e*., the length of the shared evolutionary branch after WGD is effectively zero) and in the absence of selection. Although this is a very naïve calculation it serves as an estimate of the number of duplicate pairs that will be shared due to chance alone. In the genomes of *S. cerevisiae* and *K. polysporus* 13% and 14% of loci respectively are present in duplicate and the expected number of shared duplicate loci is therefore 0.13 * 0.14 * 3252 = 60 loci. Since the observed number of shared duplicates is 212 (approximately 3.5 times the expected), this represents an excess of 152 loci.

Method 2: Accounting for the shared evolutionary branch

Using the model described in *SI Appendix,* section 5 it is possible to estimate the number of loci that were preserved in duplicate in the common ancestor of *S. cerevisiae* and *K. polysporus.* Note that the model estimates were calculated on a reduced dataset of 2299 loci, which contains exactly 169 loci (7.35%) in each of three configurations: duplicated in *S. cerevisiae* only; duplicated in *K. polysporus* only; and duplicated in both species. The model estimates that 1.93% of loci (44.4 loci) were fixed in duplicate prior to the divergence of *S. cerevisiae* and *K. polysporus,* and 5.42% of loci (7.35% - 1.93% = 5.42%; 124.6 loci) must therefore have been preserved in duplicate convergently.

Using the same approach as in Method 1 (above) it is now possible to calculate how many loci were preserved in duplicate convergently in excess of that expected by chance. At the time of divergence between *S. cerevisiae* and *K. polysporus* 1808 loci (79% of the original total) were still duplicated and in the **U** ('undecided') state and 16.24% ((169+124.6)/1808 = 0.1624) of these were preserved in duplicate in each lineage after this time. We therefore expect 0.1624 * 0.1624 * 1808 = 47.7 loci to be preserved in duplicate in both lineages by chance alone. The total expected number of shared duplicates is therefore 92.1 loci (44.4 on the shared branch and 47.7 due to sampling) and the ratio of the observed to the expected is 169/92.1 = 1.84-fold. This represents an excess of 76.9 loci and suggests that a significant number of loci have been independently preserved in duplicate in *S. cerevisiae* and *K. polysporus*.

We tested whether the observed excess of shared ohnolog pairs was statistically significant using a hypergeometric probability. Considering only the 124.6 duplicate pairs inferred to have been preserved in duplicate convergently on the *S. cerevisiae* and *K. polysporus* lineages, we calculated the probability of observing this number or greater by chance given that 293.6 (= 124.6 + 169) duplicate pairs were preserved independently on each lineage and that 1808 duplicate pairs in total were available for preservation. The probability of observing this by chance is effectively zero ($P = 2.4 \times 10^{-33}$).

Section 5. Modeling the resolution of genome duplication.

We developed a mathematical model of the loss or fixation of duplicated genes after WGD. This model is significantly more powerful and flexible than the approach we took in ref. (11). Our model assumes that the observed genomic sequences are related to each other by an (unknown) bifurcating phylogenetic topology. It attempts to explain the observed frequencies of duplicates and of the shared or divergent losses of duplicates among the five genomes (*K. polysporus, S. castellii, C. glabrata, S. cerevisiae* and *S. bayanus*). Thus, we create an 'alignment' of five species. Each site in this alignment represents an ancestral locus was duplicated in the WGD. For each species, we used YGOB to determine if that locus is still duplicated (state D_O) or had lost the first copy of the duplicate pair (S_1) or the second copy (S_2). We excluded from our analysis sites where both duplicates appear to have been lost. We use YGOB to assign consistent definitions of S_1 and S_2 across the five species (4, 11).

Our model (DL-SUBF) is in the spirit of likelihood models of character state evolution proposed by Lewis (27). We assume that a pair of loci formed by WGD can be in one of 6 possible states, and that transitions between states are possible (with rates specified by the parameters α,β and γ) as summarized in SI Figure 9A.

Initially all genes are assumed to be duplicated (i.e. $P(U|t_0)$=1.0). The instantaneous transition probabilities given in SI Figure 9A were used to construct a system of linear differential equations, which were symbolically solved using *Mathematica* 5.2. The probability of observing each state for each ancestral locus after a given time *t* is thus given by:

$$P(U \to U \mid t) = e^{-(2+2\beta+\gamma)\alpha t}$$

$$P(U \to S_1 \mid t) = \frac{(1+2\beta)\cdot(1+\beta+\gamma) - (1+\beta)\cdot(1+\gamma)\cdot e^{-(2+2\beta+\gamma)\alpha t} - \beta(2+2\beta+\gamma)\cdot e^{-(1+\gamma)\alpha t}}{(1+2\beta)\cdot(1+\gamma)\cdot(2+2\beta+\gamma)}$$

$$P(U \to F \mid t) = \frac{\gamma\cdot\left((1+2\beta)\cdot(1+2\beta+\gamma) - (1+\gamma)\cdot e^{-(2+2\beta+\gamma)\alpha t} - 2\beta\cdot(2+2\beta+\gamma)\cdot e^{-(1+\gamma)\alpha t}\right)}{(1+2\beta)\cdot(1+\gamma)\cdot(2+2\beta+\gamma)}$$

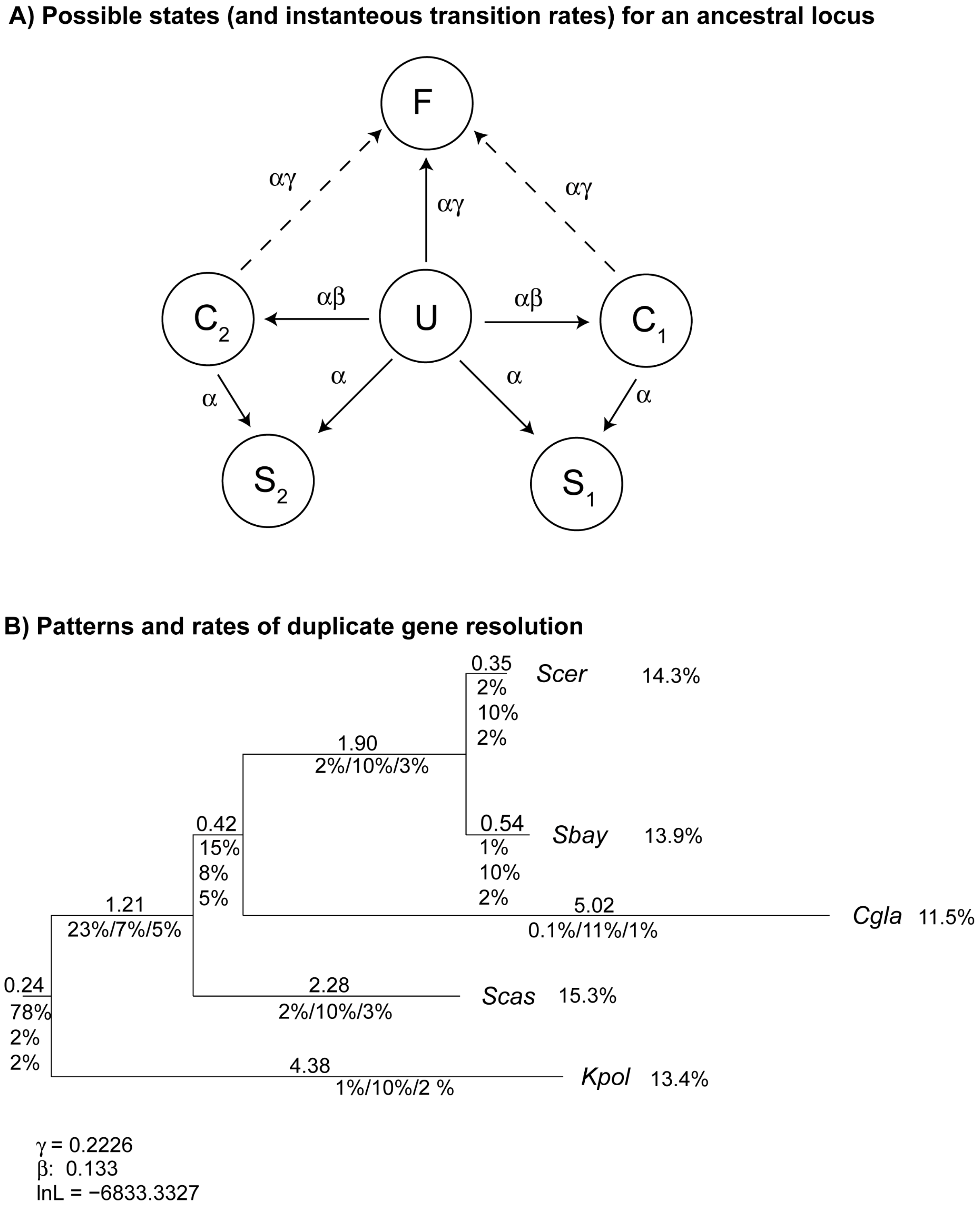
$$P(U \to C_1 \mid t) = \frac{\beta\cdot\left(e^{-(1+\gamma)\alpha t} - e^{-(2+2\beta+\gamma)\alpha t}\right)}{1+2\beta}$$

$$P(C_1 \to C_1 \mid t) = e^{-(1+\gamma)\alpha t}$$

$$P(C_1 \to S_1 \mid t) = \frac{1-e^{-(1+\gamma)\alpha t}}{1+\gamma}$$

$$P(C_1 \to F \mid t) = \frac{\gamma\cdot\left(1-e^{-(1+\gamma)\alpha t}\right)}{1+\gamma}$$

Here *U* is a state where both duplicates are present and redundant (meaning that the loss of one or the other is selectively equivalent). When one copy of a duplicate is lost, the locus transitions to state S_1 or S_2. Note that these two states are completely symmetrical and hence that equations for state S_2 are not shown above. Duplicates can also be fixed: once in state *F* neither copy of a duplicate pair can be lost.



SI Figure 9. Modeling the resolution of WGD. **(A)** The 6 model states and the rates of the possible transitions between them (see equations above). (**B)** Maximum likelihood phylogeny for the 5 species under this model inferred from 2299 conservative sites identified by YGOB. Numbers above branches are branch lengths (see text). Numbers below the branches are the percentages of the original duplicate pairs that are in states *U*, *F*, and C_1+C_2, respectively.

Our previous analysis suggested that there is an excess of convergent losses of duplicated genes (cases where two species share a loss pattern than cannot be attributed to common ancestry) (11). We incorporated this feature into the model by creating states C_1 and C_2. Genes in these states are duplicated, but if a loss is to occur from this state it will always be to state S_1 or S_2, respectively. Such loci can alternatively become fixed. Thus, an initial partial loss of function mutation in the second copy of a gene predisposes that duplicate to be lost (entering state C_1). If further mutations accumulate, that copy is lost (transition to state S_1). If the first copy instead undergoes a partial loss of function, the two copies can be fixed by subfunctionalization, with each performing a subset of the ancestral functions (state *F*). Because these convergent duplicated states can be inherited, they allow us to explain the observation of convergent losses. Note that states *F*, C_1, C_2, and *U* are degenerate with respect to our data – we can only identify observed duplicate gene pairs D_O, so for each such pair we sum over the likelihood of the four possible duplicated states in the model. By partitioning states S_1 and S_2 into separate states for convergent and non-convergent losses, we can also infer what proportion of losses along any branch are convergent. A similar approach can be taken for the fixed duplicates to determine if they were directly fixed from state *U* or by first passing through states C_1 or C_2.

Given a bifurcating phylogenetic topology τ, values of β and γ and of the *2n-1* branch lengths (αt above, where *n* is the number of taxa in our analysis), we can calculate the likelihood of the data using our own implementation of the tree-transversal algorithm of Felsenstein (28). We then use standard numerical optimization (29) to find maximum likelihood estimates of the branch lengths and of β and γ. Note that because this model is not time-reversible, our inferences are performed on rooted topologies. In practice, we infer the phylogenetic relationship of the genomes in question with an exhaustive search across all possible topologies τ, retaining the topology with the highest likelihood. The results of applying this model to our data are shown in SI Figure 9B. Above each branch is given the branch length in terms of $x = (2 + 2\beta + \gamma)\alpha t$. Taking e^{-x} gives the probability of a duplicate gene remaining in state *U* along that branch. Below each branch are the percentages of the total set of genes duplicated at WGD that are still in the duplicated states *U*, *F*, and C_1+C_2, respectively. We simulate data under the inferred maximum likelihood tree to estimate the statistical error associated with the model parameters. Doing this constitutes an implicit hypothesis test of the topology shown in SI Figure 9B. We find that this topology is strongly supported (99% confidence intervals do not overlap zero on any branch).

Degenerate forms of the above model can also be constructed so as to disallow certain evolutionary possibilities. Thus, duplicate fixation can be forbidden by setting $\gamma = 0$ (DL-C); likewise convergence by setting $\beta = 0$ (DL-F). Subfunctionalization can be precluded by letting γ and β be nonzero but forbidding transitions from C_1 and C_2 to *F* (*i.e.*, removing the dashed lines in SI Figure 9A, DL-CF). Of course fixation and convergence can also be simultaneously disallowed by setting both γ and β to zero (DL). By simulating data under these more simple models, we can test the hypotheses that duplicate fixation, convergence, and subfunctionalization are statistically significant effects. In all four cases (alternative and null models DL-F and DL, DL-C and DL, DL-CF and DL-F, and DL-SUBF and DL-CF, respectively), we find the alternative models with these effects fit the data significantly better than the null models (**P** < 0.001).

The model DL-SUBF assumes that the instantaneous rate of duplicate loss and fixation from states C_1 and C_2 (C_x) is the same as that rate from state *U.* It is possible to relax this assumption, allowing more or less rapid rates of this processes after entering state C_x. Upon applying this more complex model (DL-SUBF-2) we found that while it offered a higher likelihood than the DL-SUBF model (2ΔlnL=135.8), it was not significantly better than a model where the *U-F* transition was forbidden (DL-SUBF-2 vs. DL-SUBF-C, 2ΔlnL =1.4). Effectively, the DL-SUBF-C model thus requires all fixations to pass through states C_x. Both model DL-SUBF-C and model DL-SUBF-2 have transition probabilities that are significantly more complicated than DL-SUBF. Moreover, the improvements seen using these two models are no longer significant if *C. glabrata* and *S. bayanus* are removed from the analysis (data not shown). For reasons of clarity we have therefore chosen to report our results in terms of the simpler model. We note that our general conclusions are not altered by using these more complex models.

One hypothesis of interest is whether the whole-genome duplication observed in *K. polysporus* is actually the same event as those seen in the other four species. Were they different events, the length of the root branch, which separates *K. polysporus* from the other four taxa, would have length 0. We can test if the inferred length of this branch in SI Figure 9B is significantly different from zero by simulating data under the hypothesis that this branch has length zero and using a likelihood ratio test to compare the null to the alternative hypothesis. When we do so, we find strong evidence that this branch has non-zero length and hence that all five species underwent the same duplication event (**P** < 0.001).

Our analysis uses YGOB (4) to infer orthology between the duplicated regions of these five genomes. There are occasions, however, when this inference can be problematic. In some cases, data may be missing from the genome sequence of one organism, making it impossible to determine whether a particular WGD locus is retained in duplicate in that species. There are also cases where single copy genes in a species cannot be confidently assigned as either orthologs or a paralogs of the corresponding WGD loci in the other species (for instance if that gene resides alone on its contig). We omit all such ambiguous sites in the estimates presented here. However, adding data where one or more species is ambiguous at certain sites produces essentially identical results (data not shown).

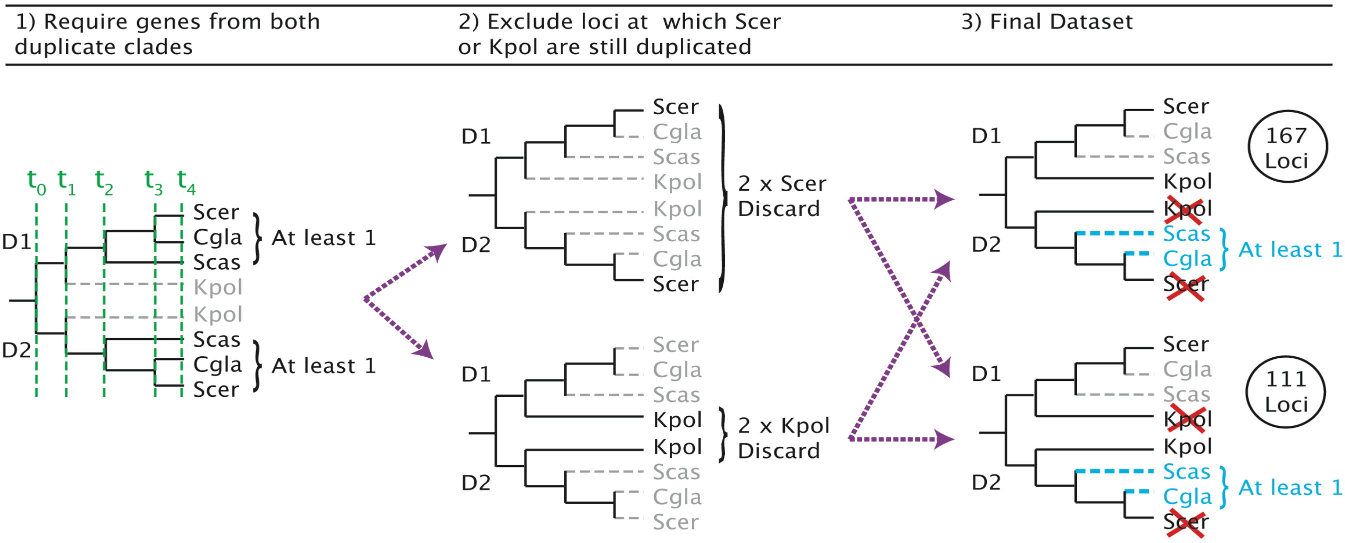
The problem of determining whether single copy genes in one species are true orthologs to their homologs in other species is especially pronounced in *K. polysporus* due to this species' early divergence from the other four species. Given this fact, it is possible that our scoring approach using YGOB could tend to over or under-estimate the proportion of shared gene losses at the root of the tree in SI Figure 9B above (further details are given in *SI Appendix*, sections 2 and 3). We can test whether this problem is misleading us by discarding the information as to which copy (S_1 or S_2) is present in *K. polysporus* and treating all single copy loci in this species as ambiguous with respect to the remaining four species (S_x). When we re-estimate the model parameters by maximum likelihood, the probability of each single copy site in *K. polysporus* is the sum of the probability for states S_1 and S_2 above. Doing so actually increases the inferred number of shared losses on the root branch of the tree in SI Figure 9B, suggesting that our original analysis is conservative in its estimate of the degree of shared ancestry between *S. cerevisiae* and *K. polysporus*. To test whether we would observe such a long root branch were the genome duplication not shared between the five species, we simulated data under the assumption of no shared ancestry between *K. polysporus* and the other four taxa. We then discarded information on which single copy genes were present for *K. polysporus* (creating the same ambiguities as above) and optimized the resulting datasets under the assumption of a zero length root branch and without this constraint. None of these simulated datasets showed an improvement in likelihood after constraint relaxation that was as large as seen in the real data (**P** < 0.001). This is strong evidence that our scoring approach has not misled us into inferring a single duplication event. It is also an encouraging signal that many of our other conclusions would be robust to incorrect tracking.

Section 6. Direct comparison of representative K_A values between convergently and divergently resolved loci.

To exclude the possibility that the result shown in Figure 4 could be caused by a general trend towards resolving slower-evolving loci at later time points, we tested whether loci undergoing convergent loss at later time points tended to be biased towards slower-evolving loci, in the same way as loci undergoing RGL are biased.

We assembled sets of loci at which either convergent gene loss (orthologs lost in two independent lineages; single-copy orthologs retained) or divergent gene loss (paralogs lost in two independent lineages; single-copy paralogs retained) have occurred between *S. cerevisiae* and *K. polysporus*. We excluded the possibility that loci in our convergent gene loss dataset were products of a single gene loss on a shared branch by requiring that the missing gene copy be still present in either *S. castellii* or *C. glabrata.* Although divergent gene loss at an ancestrally duplicated locus cannot be explained by a single gene loss on a shared branch, we imposed the same phylogenetic criterion when selecting convergently and divergently resolved loci so the two datasets could be compared directly.

In brief, we used YGOB to select loci at which one gene copy from each duplicate clade was retained in at least one of *S. cerevisiae*, *C. glabrata* or *S. castellii* (SI Figure 10 panel 1). All loci selected on this basis must have been retained in duplicate on the lineage leading to *S. cerevisiae* until at least the divergence of *S. castellii* (t_2 in panel 1). We then discarded any loci at which duplicates have been retained in either *S. cerevisiae* or *K. polysporus* (panel 2) and partitioned the remaining loci into those at which single-copy orthologs (167 loci) and single-copy paralogs (111 loci) were retained between *S. cerevisiae* and *K. polysporus* (panel 3).


1) Require genes from both duplicate clades
2) Exclude loci at which Scer or Kpol are still duplicated
3) Final Dataset
t_0 t_1 t_2 t_3 t_4
D1
D2
Scer
Cgla
Scas
Kpol
Kpol
Scas
Cgla
Scer
At least 1
At least 1
2 x Scer Discard
2 x Kpol Discard
167 Loci
111 Loci

SI Figure 10. Method of selection of sets of genes that have either been convergently or divergently resolved between *S. cerevisiae* and *K. polysporus*. Because all of these loci were retained in duplicate on the *S. cerevisiae* lineage until at least the divergence of *S. castellii*, they must all have involved at least two independent gene losses: one on the *K. polysporus* lineage in the interval between t_1 and t_4 and one on the *S. cerevisiae* lineage between t_2 and t_4.

For each locus in both datasets we calculated 'representative' K_A values between the orthologous genes in *K. lactis* and *A. gossypii*, $K_{A(\text{Klac-Agos})}$ (11), because this provides a measure of the intrinsic rate of evolution of the gene unaffected by any possible rate acceleration after gene duplication (30). We find that the median $K_{A(\text{Klac-Agos})}$ in single-copy orthologs is significantly greater than that amongst single-copy paralogs (0.3732 vs. 0.3315; $P = 0.006$ by one-sided Wilcoxon rank-sum test), indicating that RGL occurs preferentially at slow-evolving loci.

Although we used the same procedure to select loci for our single-copy ortholog and single-copy paralog datasets, it is possible that these datasets may be enriched for loci with different patterns of gene loss in *S. castellii* and *C. glabrata* and that it may therefore not be appropriate to compare them directly. To exclude this possibility we paired loci between our single-copy ortholog and single-copy paralog datasets whose patterns of gene loss were identical in all species except that the single-copy ortholog had retained the same (syntenic ortholog) gene copy in both *S. cerevisiae* and *K. polysporus* while the single copy paralog had retained alternative gene copies in these species. This produced 106 locus pairs whose only systematic difference is that one locus in each pair had lost orthologous gene copies independently in *S. cerevisiae* and *K. polysporus* and the second locus had independently lost paralogous gene copies. We performed this matching procedure 100 times and found that in 79 of 100 replicates, the $K_{\mathrm{A(Klac\text{-}Agos)}}$ values for single-copy paralogs were significantly lower ($P < 0.05$ by one-sided Wilcoxon rank-sum test) than those for single-copy orthologs.

These results are consistent with hypothesis that RGL is more likely to occur at loci where duplicates are functionally interchangeable (11) and that this condition is more likely to be met by slowly evolving loci.

Section 7. The proportion of partisan gene losses increases on successive branches after the WGD

As shown in Figure 2C the percentage of partisan losses (**C→S** transitions) as a fraction of all gene loss events (**U→S** and **C→S** transitions) inferred by our model of gene loss increases on successive branches after the WGD. It rises from 1% on the earliest branch after the WGD to 40% on the terminal *S. cerevisiae* branch. Because neutral losses (**U→S** transitions) arise from state **U** (which initially contains 100% of loci and must therefore decrease) while partisan losses arise from state **C** (which initially contains 0% of loci and must therefore decrease), we wanted to exclude the possibility that the increasing prevalence of partisan loss relative to neutral loss was a trivial consequence of the structure of our model. We therefore used a method that does not rely on the model to estimate the proportions of neutral and partisan gene losses at two different timepoints after the WGD and verified that the fraction of partisan gene losses is significantly higher at the later timepoint.

A simple method to estimate the proportion of neutral and partisan losses using gene loss data from post-WGD genome trios is described in ref. (11). Because any three post-WGD genomes can be resolved into a pair of ingroup genomes and a single outgroup genome, it is possible to identify loci that have been returned to single-copy independently in the outgroup genome and one of the in-group genomes by selecting loci that are still duplicated in the second ingroup genome (See Fig. 2, Classes 2C – 2F in ref. (11)). We can then compare the proportions of loci at which orthologous and paralogous gene copies (using synteny information to distinguish syntenic orthologs from non-syntenic paralogs) have been retained between the single-copy outgroup and ingroup genomes. Moreover, since any excess of orthologous over paralogous gene losses must be attribuSI Table to events on the shared evolutionary branch between the WGD and the divergence of the three species of interest, we can examine the effect of the time since duplicate gene divergence by selecting genome trios whose common ancestor existed at different timepoints after the WGD.

We used a genome trio composed of *(K. polysporus, (S. castellii, S. cerevisiae))* and one composed of *(S. castellii, (C. glabrata, S. cerevisiae))* to identify sets of genes that were resolved independently in two lineages after the divergence of *K. polysporus* (Kpol-Trio) or *S. castellii* (Scas-Trio) from the *S. cerevisiae* lineage respectively. Following exclusion of any loci that did not satisfy the synteny quality criteria required by the Yeast Gene Order Browser (4), we obtained 130 loci from the Kpol-Trio and 83 loci from the Scas-Trio for which independent resolution of gene duplicates in two lineages could be inferred with confidence. As can be seen from SI Table 6 (below), the proportion of orthologous and paralogous gene losses is close to equal for the Kpol-Trio (77 orthologous gene losses compared to 53 paralogous gene losses in the combined dataset) but very skewed for the Scas-Trio (65 orthologous gene losses compared to 18 paralogous gene losses in the combined dataset). These are significantly different in a chi-squared test of homogeneity (P = 0.006) indicating that the proportion of orthologous and paralogous gene losses depends on the time since the WGD. In addition, the direction of the change in the relative proportions of orthologous and paralogous gene losses (increase in the former relative to the latter at the later timepoint) is consistent with the idea that proportion of orthologous gene losses (and hence partisan losses; SI Table 6) increases with time since the duplication. These data indicate that the conclusion that the proportion of partisan gene losses is higher at later timepoints is not solely due to the structure of our likelihood model but is a property of the data.

SI Table 6. Estimated percentage of partisan gene losses at two different timepoints based on counts of orthologous and paralogous gene losses from two genome trios.

Outgroup	Ingroup		Gene Losses					
Single-copy	Single-copy	Double-copy	Orthologous losses	Paralogous losses	Total	Neutral losses*	Partisan losses*	% Partisan losses
Kpol	Scer	Scas	47	28	75	56	19	25.3%
Kpol	Scas	Scer	30	25	55	50	5	9.1%
	Combined		77	53	130	106	24	18.5%
Scas	Scer	Cgla	26	9	35	18	17	48.6%
Scas	Cgla	Scer	39	9	48	18	30	62.5%
	Combined		65	18	83	36	47	56.6%

* The number of neutral gene losses was estimated as twice the number of paralogous gene losses and the number of partisan gene losses was calculated as the number of orthologous gene losses minus the number of paralogous gene losses. See ref. (11) for justification. Note that because of the method by which these loci were selected (duplicates were required in at least one species) the proportions of orthologous and paralogous (or neutral and partisan) losses are not the same as those estimated by the model (Fig. 2C). The latter are based on a much larger and less biased dataset and should be more accurate.

SI References

1. Pop, M., Kosack, D. S. & Salzberg, S. L. (2004) *Genome Res* **14,** 149-59.
2. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res* **25,** 955-64.
3. Eddy, S. R., Mitchison, G. & Durbin, R. (1995) *J Comput Biol* **2,** 9-23.
4. Byrne, K. P. & Wolfe, K. H. (2005) *Genome Res* **15,** 1456-61.
5. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res* **22,** 4673-4680.
6. Castresana, J. (2000) *Mol Biol Evol* **17,** 540-52.
7. Shimodaira, H. & Hasegawa, M. (2001) *Bioinformatics* **17,** 1246-7.
8. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002) *Bioinformatics* **18,** 502-4.
9. Sugino, R. P. & Innan, H. (2005) *Genetics* **171,** 63-9.
10. Guindon, S. & Gascuel, O. (2003) *Syst Biol* **52,** 696-704.
11. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. (2006) *Nature* **440,** 341-5.
12. Kurtzman, C. P. & Robnett, C. J. (2003) *FEMS Yeast Res* **3,** 417-32.
13. Kurtzman, C. P. (2003) *FEMS Yeast Res* **4,** 233-45.
14. Hunter, T. & Plowman, G. D. (1997) *Trends Biochem Sci* **22,** 18-22.
15. van der Walt, J. P. (1956) *Antonie van Leeuwenhoek* **22,** 265-272.
16. Roberts, C. J. & van der Walt, J. P. (1959) *Compt Rend Lab Carlsberg* **31,** 129-148.
17. Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B. & Fairhead, C. (2005) *Mol Biol Evol* **22,** 856-873.
18. Muthukumar, G., Suhng, S. H., Magee, P. T., Jewell, R. D. & Primerano, D. A. (1993) *J Bacteriol* **175,** 386-94.
19. Esteban, P. F., Vazquez de Aldana, C. R. & del Rey, F. (1999) *Yeast* **15,** 91-109.
20. Forsberg, H. & Ljungdahl, P. O. (2001) *Mol Cell Biol* **21,** 814-26.
21. Wong, S. & Wolfe, K. H. (2005) *Nature Genetics* **37,** 777-82.
22. Kirchhausen, T. (2000) *Nat Rev Mol Cell Biol* **1,** 187-98.
23. Cliften, P. F., Fulton, R. S., Wilson, R. K. & Johnston, M. (2006) *Genetics* **172,** 863-72.
24. Nakayashiki, H., Kadotani, N. & Mayama, S. (2006) *J Mol Evol* **63,** 127-35.
25. Neuveglise, C., Feldmann, H., Bon, E., Gaillardin, C. & Casaregola, S. (2002) *Genome Res* **12,** 930-43.
26. Byrnes, J. K., Morris, G. P. & Li, W. H. (2006) *Mol Biol Evol* **23,** 1136-43.
27. Lewis, P. O. (2001) *Syst Biol* **50,** 913-25.
28. Felsenstein, J. (1981) *J Mol Evol* **17,** 368-376.
29. Press, W. H., Teukolsky, S. A., Vetterling, W. A. & Flannery, B. P. (1992) *Numerical Recipes in C* (Cambridge University Press, New York).
30. Davis, J. C. & Petrov, D. A. (2004) *PLoS Biol* **2,** E55.