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, and Kenneth H. Wolfe

PNAS published online May 9, 2007;
doi:10.1073/pnas.0608218104

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, and Kenneth H. Wolfe*

Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

Edited by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved March 15, 2007 (received for review September 18, 2006)

Among yeasts that underwent whole-genome duplication (WGD), *Kluyveromyces polysporus* represents the lineage most distant from *Saccharomyces cerevisiae*. By sequencing the *K. polysporus* genome and comparing it with the *S. cerevisiae* genome using a likelihood model of gene loss, we show that these species diverged very soon after the WGD, when their common ancestor contained >9,000 genes. The two genomes subsequently converged onto similar current sizes (5,600 protein-coding genes each) and independently retained sets of duplicated genes that are strikingly similar. Almost half of their surviving single-copy genes are not orthologs but paralogs formed by WGD, as would be expected if most gene pairs were resolved independently. In addition, by comparing the pattern of gene loss among *K. polysporus*, *S. cerevisiae*, and three other yeasts that diverged after the WGD, we show that the patterns of gene loss changed over time. Initially, both members of a duplicate pair were equally likely to be lost, but loss of the same gene copy in independent lineages was increasingly favored at later time points. This trend parallels an increasing restriction of reciprocal gene loss to more slowly evolving gene pairs over time and suggests that, as duplicate genes diverged, one gene copy became favored over the other. The apparent low initial sequence divergence of the gene pairs leads us to propose that the yeast WGD was probably an autopolyploidization.

genomics | polyploidy | reciprocal gene loss | *Vanderwaltozyma polyspora*

An ancestor of *Saccharomyces cerevisiae* underwent whole-genome duplication (WGD) after it had diverged from non-WGD yeast lineages such as *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Ashbya gossypii* (1–4). The WGD had a major impact on the evolution of *S. cerevisiae* and its relatives, most notably by facilitating their adaptation to anaerobic growth (5) and contributing to their rapid speciation (6). In *S. cerevisiae*, ≈20% of genes are members of duplicated pairs that were formed in the WGD (7). The other loci became single-copy again during the sorting-out process (genome reduction) that occurred after the WGD. Similar large-scale loss of copies of duplicated genes from paleopolyploid genomes has occurred during the evolution of plants such as grasses and crucifers (8–11).

Because the *S. cerevisiae* genome sequence is a single observation of the evolutionary result of the WGD that occurred in a yeast ancestor, it has not been clear whether the set of genes that survived the sorting-out process in *S. cerevisiae* was an inevitable outcome of the WGD, or whether stochastic processes played a major role. Two questions need to be answered. First, are the loci that remain duplicated in *S. cerevisiae* a special subset of the pre-WGD genome, that were somehow more amenable to retention in duplicate after WGD? Second, for loci that are now single-copy in *S. cerevisiae*, was retention of one particular gene copy preferred over the other? These questions are best addressed by studying the genomes of other yeast species that are descended from the same WGD event. Unfortunately, the post-WGD species whose genomes have been sequenced so far are so closely related to each other that the gene loss process was already nearly complete by the time they diverged (6). Ideally, we would like to compare genomes that diverged as soon as possible after the WGD, so that relatively little of the sorting-out process occurred on a shared evolutionary branch.

In this study, we show that *Kluyveromyces polysporus* is a member of the post-WGD lineage that is most divergent from *S. cerevisiae* and that the vast majority of genes were still duplicated when the lineages leading to these species diverged. We take advantage of the fact that most duplicate gene pairs were resolved twice (once on the *K. polysporus* lineage and once on the *S. cerevisiae* lineage) to study the extent to which the process of gene loss or retention in duplicate was nonrandom. We find that the two species show similar biases toward retaining duplicated loci with particular biological functions but that, for some functions, the actual genes retained in duplicate are often different. For loci that have become single-copy again, we find that the "choice" of which copy was discarded became increasingly nonrandom as time elapsed after the WGD.

Results and Discussion

***Kluyveromyces polysporus* Is a Member of the Post-WGD Clade That Is Most Divergent from *S. cerevisiae*.** The phylogeny of hemiascomycete yeasts was recently resolved into 14 clades by Kurtzman and Robnett (12) [supporting information (SI) Fig. 5]. The post-WGD species with sequenced (4, 13–16) or surveyed (17–19) genomes lie in clades 1–4, whereas clades 7–14 are outgroups lacking the duplication (20). Clades 5 and 6 are monophyletic and sister to clades 1–4, but it was not known whether they underwent the WGD or whether this event occurred after clades 1–4 split from clades 5 and 6. We sequenced a few hundred random genomic fragments (*SI Methods*) from *K. polysporus* (in clade 6) and *Kluyveromyces phaffii* (in clade 5). These data suggested that *K. polysporus* and *K. phaffii* both underwent genome duplication, and hence are representative of the WGD lineage most deeply diverged from *S. cerevisiae*. We chose the type strain of *K. polysporus*, originally isolated from soil in South Africa (21), for more extensive whole-genome shotgun sequencing.

Genome Sequence and Gene Content of *K. polysporus*. Our *K. polysporus* 7.8× coverage draft genome sequence consists of 290 contigs totaling 14.7 Mb, organized into 41 supercontigs

Author contributions: D.R.S., G.C.C., and K.H.W. designed research; D.R.S., A.C.F., G.C.C., K.P.B., and M.W. performed research; D.R.S., A.C.F., G.C.C., and K.P.B. analyzed data; and D.R.S. and K.H.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: WGD, whole-genome duplication; RGL, reciprocal gene loss; YGOB, Yeast Gene Order Browser; GO, Gene Ontology.

Data deposition: The sequence of *Kluyveromyces polysporus* reported in this paper has been deposited in the GenBank database (accession no. AAZN00000000).

*To whom correspondence should be addressed. E-mail: khwolfe@tcd.ie.

This article contains supporting information online at www.pnas.org/cgi/content/full/0608218104/DC1.

EVOLUTION

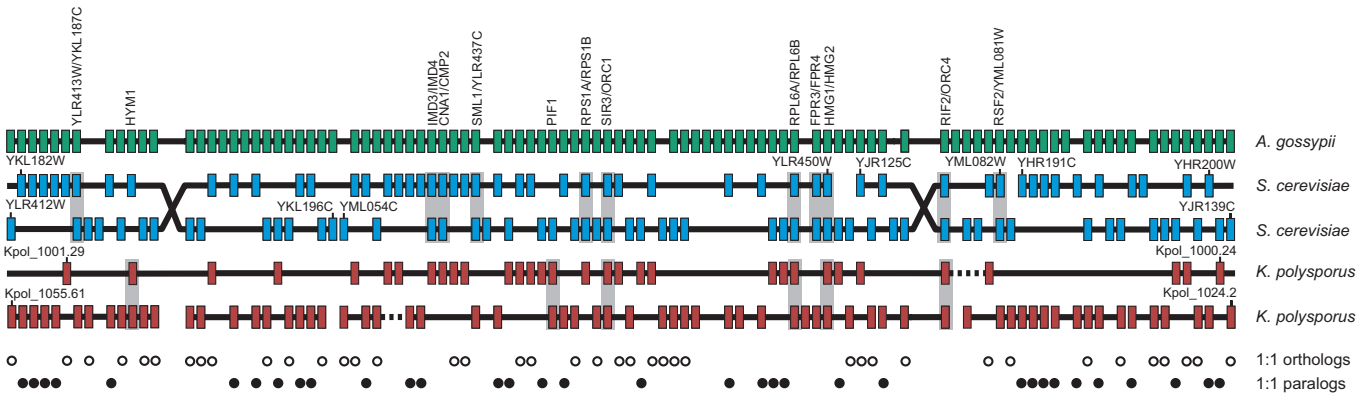


Fig. 1. Gene order relations in the genomic region around the *SIR3/ORC1* gene pair. There are two genomic tracks for each of the post-WGD species *K. polysporus* and *S. cerevisiae* and a single track for the non-WGD species *A. gossypii*. Colored rectangles represent genes, and genes in the same column are homologs. Retained duplicated genes in the post-WGD species are highlighted by gray shading and their *S. cerevisiae* names are shown at the top. Solid black lines connect genes that are immediate neighbors on a chromosome or contig. Dashed black lines in *K. polysporus* connect genes that are neighbors on the same supercontig, but between which there is a gap in the genome sequence. The tracks have been drawn to show how YGOB assigns orthology and paralogy between *K. polysporus* and *S. cerevisiae*: The upper tracks in the two species are considered orthologous, as are the two lower tracks. The two X symbols in *S. cerevisiae* show places where YGOB's orthology/paralogy assignments switch between chromosomes. Open and filled circles show how YGOB scored the 74 single-copy loci in this region as 40 orthologs and 34 paralogs, respectively.

(*SI Methods* and SI Fig. 6). We identified 5,652 protein-coding genes, 251 tRNAs and at least 39 LTR retrotransposons. The sequence has been submitted to GenBank and can be compared with other yeast genomes by using the Yeast Gene Order Browser (YGOB) (7). In general, the genome is similar in size and gene content to that of *S. cerevisiae*, but some notable differences exist (*SI Appendix*, section 1). For instance, several *S. cerevisiae* genes for components of dynein and dynactin (*DYN1*, *DYN3*, *PAC11*, *ARP1*, *JNM1*, and *NIP100*) have no homologs in *K. polysporus*. It is likely that these gene losses relate to a major phenotypic difference between *K. polysporus* and other yeasts: its asci typically contain 50–100 spores, which are formed by extra mitotic replications after meiosis (21, 22). In *S. cerevisiae*, dynein and dynactin serve to position the mitotic spindle across the bud neck (23), but the extra mitoses in *K. polysporus* occur in cells without buds.

The Genomes of *S. cerevisiae* and *K. polysporus* Are Superficially Similar but Very Different in Detail. The genome sequence data confirm that *K. polysporus* has undergone WGD. Like *S. cerevisiae*, its genome consists of pairs of sister chromosomal regions that contain some duplicated genes and show a double conserved synteny relationship with single genomic regions in non-WGD species such as *A. gossypii* (Fig. 1). Among the 3,252 ancestral loci that we could reliably compare between the *K. polysporus* and *S. cerevisiae* genomes using the YGOB engine (7), we identified 450 gene pairs formed by WGD (ohnologs) that have been retained in *K. polysporus* (Table 1). Thus, the overall fraction of ancestral loci retained in duplicate in *K. polysporus* is similar to that in *S. cerevisiae* (13.8% and 13.3%, respectively, for the data set in Table 1). However, beneath this superficial similarity, the details of gene loss are so different between the species that it is difficult to tell which of the two sister regions in *K. polysporus* is orthologous to which of the two sister regions in *S. cerevisiae* (Fig. 1). By contrast, orthologous sister regions are readily identifiable among the other post-WGD species *S. cerevisiae*, *S. castellii* and *C. glabrata* because they share many gene losses that differentiate them from their paralogous sisters (6).

Approximately Equal Numbers of Single-Copy Orthologs and Paralogs Between *K. polysporus* and *S. cerevisiae*. When two closely related genomes are compared, any gene in one species almost invariably has an ortholog in the other species. However, we estimate that only 56% of loci that are single-copy in both *K. polysporus* and *S. cerevisiae* are orthologs (genes that diverged in the speciation event) and the remaining 44% are paralogs (these genes became duplicated in the WGD, and after speciation the two species reciprocally lost different copies) (Table 1). The almost equal numbers of orthologs and paralogs around *SIR3/ORC1* (Fig. 1) are typical of the whole genome, as is the loss of approximately equal numbers of genes from both sister regions. Even the apparent small excess of putative orthologs over putative paralogs in Table 1 may be an artifact of the algorithm used by YGOB, which assumes that the genomic regions with the

Table 1. Patterns of differential gene retention between *K. polysporus* and *S. cerevisiae*

Copy no. relationship (*K. polysporus*:*S. cerevisiae*)	No. of ancestral loci	Fraction among all loci, %	Fraction among single-copy loci, %
2:2	212	6.5	–
2:1	238	7.3	–
1:2	221	6.8	–
1:1 (orthologous)	1,455	44.7	56.4
1:1 (paralogous)	1,126	34.6	43.6
Total	3,252	100.0	100.0

Only the 3,252 ancestral loci that could be scored reliably (6, 7) on both sister tracks in both species were counted here. The total numbers of ohnologs are at least 551 in *S. cerevisiae* (7) and at least 492 in *K. polysporus*, but interspecies rearrangements and gaps in the *K. polysporus* sequence cause some of these loci to be scorable in only one species.

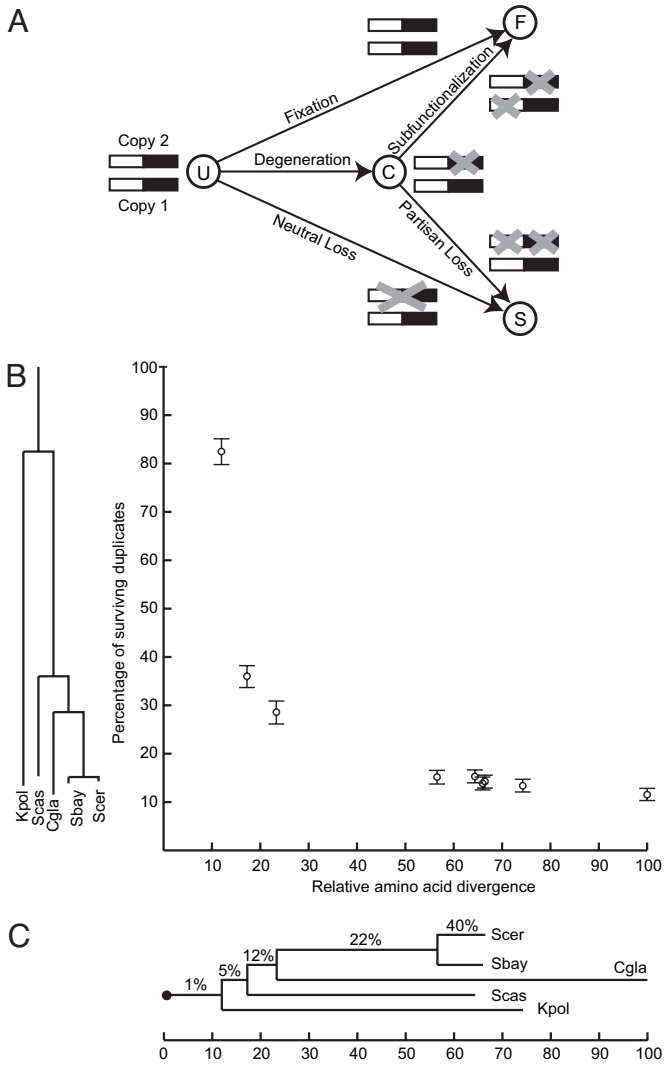


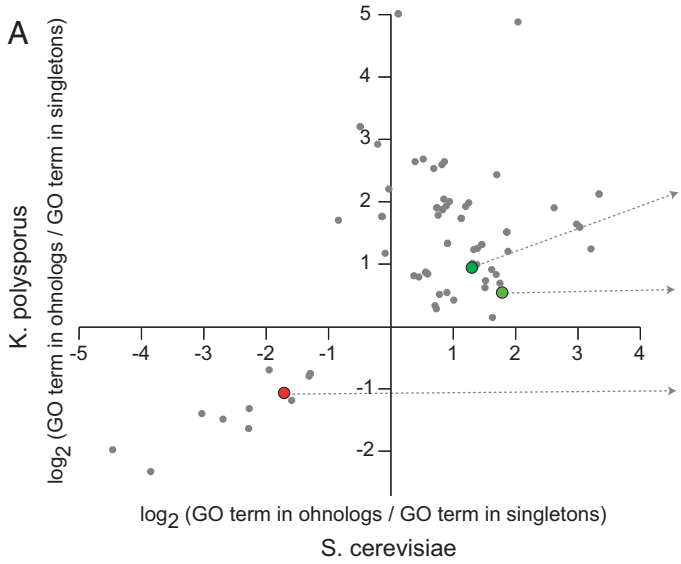
Fig. 2. Modeling gene pair evolution reveals a changing pattern of gene loss after WGD. (*A*) Our likelihood model of gene pair evolution, showing the four possible states of a pair (U, C, S, F; defined in *A Likelihood Model of Gene Loss After WGD That Incorporates Partisan Gene Loss*), and the permissible transitions between them (arrows). A hypothetical gene pair (copy 1 and copy 2) is shown, containing two domains (white and black boxes). Gray X symbols represent loss-of-function mutations that inactivate either a single domain or a whole gene and cause a pair to move from one state to another. (*B*) Likelihood estimates of the process of gene loss after WGD. Each point on the graph represents the estimated proportion of loci remaining duplicated at a node on the phylogenetic tree. *y* axis values come from the branch lengths of the tree on the left, which was obtained by optimizing the topology and parameters in our likelihood model of gene pair evolution (*SI Appendix*, section 5). *y* axis values are the total proportion of loci in states U + C + F, and their error bars were obtained by parametric bootstrapping. The *x* axis values correspond to amino acid divergence and are taken from the tree in *C*; we did not enforce a molecular clock to convert amino acid divergence into time units. (*C*) Tree reconstructed from protein sequences of 11 genes that are duplicated in all five species. Branch-lengths of duplicated branches have been averaged to obtain a species tree. The black dot indicates the time of divergence of duplicated gene pairs. On each branch on the lineage leading to *S. cerevisiae*, the estimated proportion of partisan gene losses (C → S transitions) is shown as a percentage of all loci returned to single-copy on that branch.

greatest shared gene content between species are orthologous (7). Indeed, the observed 56:44 ratio of orthologs to paralogs among single-copy genes is not significantly different from the 50:50 ratio that would be expected if the two species had gone through completely independent processes of gene loss after WGD (*SI Appendix*, section 2). Importantly, the conclusion that a high proportion of paralogs exists is robust to possible track-assignment errors in YGOB (*SI Appendix*, section 3). The extent of paralogy of single-copy genes observed between *K. polysporus* and *S. cerevisiae* greatly exceeds the levels previously documented in other pairs of species (6, 24). Our discovery that orthologs do not exist at many loci has negative implications for the prospect of using nuclear gene sequences to resolve the phylogenetic relationships among any group of paleopolyploid species that diverged soon after a WGD.

Similar Numbers and Types of Duplicate Gene Pairs Retained in *K. polysporus* and *S. cerevisiae*. The high proportion of paralogs seen between *K. polysporus* and *S. cerevisiae* indicates that these species must have diverged very soon after the WGD and undergone largely independent processes of gene loss. This result was perhaps expected given the phylogenetic position of *K. polysporus*, and is consistent with a Dobzhansky–Muller mechanism of speciation in post-WGD yeasts by reciprocal loss of duplicated genes (6, 25, 26). Using a likelihood model of the process of resolution of duplicated gene pairs (described below; Fig. 2*A*) we estimate that 82% of loci were still duplicated at the time that *S. cerevisiae* and *K. polysporus* diverged (Fig. 2*B*) and the common ancestor of these two species thus had at least 9,000 genes (assuming that the pre-WGD yeast had 5,000 genes; $5{,}000 \times 1.82 = 9{,}100$). Viewed from this perspective it is striking that, after speciation, the numbers of retained duplicates in the two species subsequently dropped independently to the same level (13–14% of the original gene set). Despite this independent history, 47% of the ohnolog pairs in *K. polysporus* have also been retained in duplicate in *S. cerevisiae* (212 of 450; Table 1). The number of shared ohnologs is 1.9-fold higher than expected by chance, even allowing for some shared ancestry, and must indicate convergent evolution of genome content ($P < 5 \times 10^{-33}$ by hypergeometric distribution; *SI Appendix*, section 4). More generally, we find that Gene Ontology (GO) terms that are significantly over- or underrepresented among the ohnologs of one yeast species, relative to its singletons, tend to be similarly biased in the other species (Fig. 3*A*). Both species show significant underrepresentation of genes involved in RNA metabolism, mRNA processing, and rRNA processing among duplicates relative to singletons, and significant overrepresentation of duplicated genes for cytosolic ribosomal proteins, protein kinases, and carbohydrate metabolism.

The Pattern of Duplicate Gene Preservation Varies Among Functional Categories. Surprisingly, however, the similarities of GO category biases among duplicates and singletons in the two species do not necessarily mean that the same loci have been retained in duplicate in both. We find that in GO categories that are underrepresented among ohnologs relative to singletons, such as "RNA metabolism" and "nucleoplasm," the degree to which ohnologs are shared by the two species is greater than in the genome at large (Fig. 3*B*). In these categories relatively few loci were retained in duplicate but both species tended to retain the same genes. Conversely, in GO categories that are overrepresented among ohnologs relative to singletons, such as "kinase activity," the level of ohnolog sharing between species is less than the genome average and no more than expected by chance (Fig. 3*B*; *SI Appendix*, section 4). Detailed analysis of a curated set of 75 ancestral protein kinase loci (a subset of the GO term kinase activity) shows that *S. cerevisiae* retains 25 duplicated pairs and *K. polysporus* retains 18 pairs, but only six of these pairs are the same; the others are in 2:1 or 1:2 relationships (SI Fig. 7). These data suggest that the GO categories that are overrepresented among ohnologs are overrepresented because certain types of gene (as opposed to particular genes) are favored for preservation in duplicate (10, 11, 27–29). Thus, in answer to the first

EVOLUTION



B

Gene Ontology Category	K. pol (%) 1	K. pol (%) 2	S. cer (%) 1	S. cer (%) 2	2 K. pol : 2 S. cer Obs	Exp	Ratio	Sig
kinase activity	3.2	6.3	3.0	7.5	10.5	10.8	1.0	
cell wall organization / biogenesis	2.2	5.6	2.1	5.9	14.0	10.4	1.3	*
carbohydrate metabolism	2.9	7.0	2.9	7.3	17.5	12.6	1.4	*
regulation of cell cycle	2.2	5.7	2.4	4.5	13.5	8.6	1.6	*
structural constituent of ribosome	3.5	5.3	2.9	9.7	21.0	12.3	1.7	***
regulation of cellular metabolism	6.8	12.1	7.1	9.8	29.5	16.0	1.8	***
Genome Average	86.2	13.8	86.7	13.3	212	112.3	1.9	***
RNA metabolism	10.6	4.8	10.7	3.6	8.5	3.23	2.6	**
nucleoplasm	5.9	2.9	6.0	1.9	6.0	2.1	2.9	**

Fig. 3. Duplicate gene retention in different GO categories in *K. polysporus* and *S. cerevisiae*. (*A*) Ratios of occurrence of particular GO terms among duplicates, relative to single-copy genes, in the two species. Each point represents a GO term; only terms that are significantly overrepresented or underrepresented in at least one of the two species ($\alpha < 0.001$ by Fisher's exact test) are shown. Colored data-points and dashed arrows show GO terms that also appear in *B*. Ratios are presented on a $\log_2$ scale, so 0 indicates a term that is equally frequent among ohnologs and singletons; 3 indicates 8-fold overrepresentation of a GO term among ohnologs relative to singletons, and -3 indicates 8-fold underrepresentation. Note that GO terms are not mutually exclusive so it is not appropriate to calculate a correlation. Details are given in SI Table 2 and SI Table 3. (*B*) Variation in the extent of overlap between species, within GO categories, of the genes retained in duplicate. The color scale indicates the ratio (Ratio) of the observed number of loci with a GO term retained in duplicate in both species (Obs) to the expected number (Exp). Observed values were obtained from YGOB. Expected values were calculated from the product of the duplicate preservation rates in each species after correcting for the shared evolutionary branch (*SI Appendix*, sections 4 and 5). Asterisks show Obs/Exp ratios significantly greater than one (hypergeometric probability: *, $P \leq 0.05$; **, $P \leq 10^{-3}$; ***, $P \leq 10^{-5}$). The other columns show the frequency of the GO term in each species among singletons and among ohnologs (columns labeled 1 and 2, respectively).

question we posed in the introduction, there is evidence that *K. polysporus* and *S. cerevisiae* independently converged toward similar categories of retained duplicate genes after WGD. The outcome of the WGD was therefore surprisingly predictable in terms of the functions of retained genes and the eventual overall level of gene retention, but generally unpredictable at the level of the fate of individual genes.

Convergent Loss of Gene Duplicates. To explore the second question (whether the two copies of a gene are equally prone to loss), we included several modes of duplicate gene loss in our likelihood model, and fitted its parameters to YGOB data for five post-WGD species (*SI Appendix*, section 5). In our previous study of *S. castellii*, *C. glabrata*, and *S. cerevisiae* (6), we found that, at loci where two of the species had each lost one member of an ohnolog pair through independent loss events, convergent losses of orthologous copies were seen about three times more frequently than reciprocal losses of paralogous copies, instead of the 50:50 ratio expected for independent events (classes 2C/2D and 2E/2F in ref. 6). This result suggested that there were selective differences between copies (a particular copy was preferentially retained), but it did not indicate whether these selective differences were present at the time of the WGD or emerged gradually afterward. By including data from *K. polysporus* it now becomes possible to study how the patterns of gene loss changed over time.

A Likelihood Model of Gene Loss After WGD That Incorporates Partisan Gene Loss. Our model of gene pair evolution (Fig. 2*A*) proposes that after WGD, all gene pairs are initially in a state U ("undecided") where the two copies are functionally equivalent and either of them could be lost. Over time, the pair can transition into one of three other possible states: F ("fixed") where the duplication has been fixed; S ("single-copy"), where one member of the pair has been lost; or C ("converging"), a state where both gene copies remain in the genome but there are selective differences such that the loss of one copy (copy 1, for instance) would be deleterious whereas loss of the other (copy 2) would be neutral. We included state C in our model to account for the aforementioned excess of convergent losses over reciprocal losses at loci where two independent losses had occurred (6). Note that loci cannot remain in states C or U indefinitely. As a hypothetical example, state C could include a pair of genes coding for a two-domain protein, but where one of the domains has been inactivated in gene copy 2, with the result that copy 1 is essential but copy 2 is not (Fig. 2*A*). This situation can be resolved either by inactivation of the other domain in copy 1 (subfunctionalization and transition to state F), or by complete loss of gene copy 2 (transition to state S). We refer to the latter as partisan gene loss (as distinct from neutral gene loss) because the identity of the lost gene copy is not arbitrary. If a speciation occurs when the C-state pair is still duplicated, any subsequent losses in the descendant species must be of gene copy 2 and so will be convergent. Inclusion of state C in the likelihood model significantly improves the fit to the data (*SI Appendix*, section 5). Moreover, when we compare the likelihoods of the model across all possible branching orders of the post-WGD species, the tree with the highest likelihood (Fig. 2*B*, *y* axis) has the expected topology (6) and places a significant number of gene losses on the shared branch between the WGD and first speciation (of *K. polysporus* from the other post-WGD species), which is evidence against the unparsimonious possibility that *K. polysporus* and *S. cerevisiae* might be descended from two independent WGD events.

The Pattern of Gene Loss from Duplicated Loci Changes with Time. In our model, gene pairs gradually move out of state U and into other states (Fig. 2*A*). Because state U is the only one that can

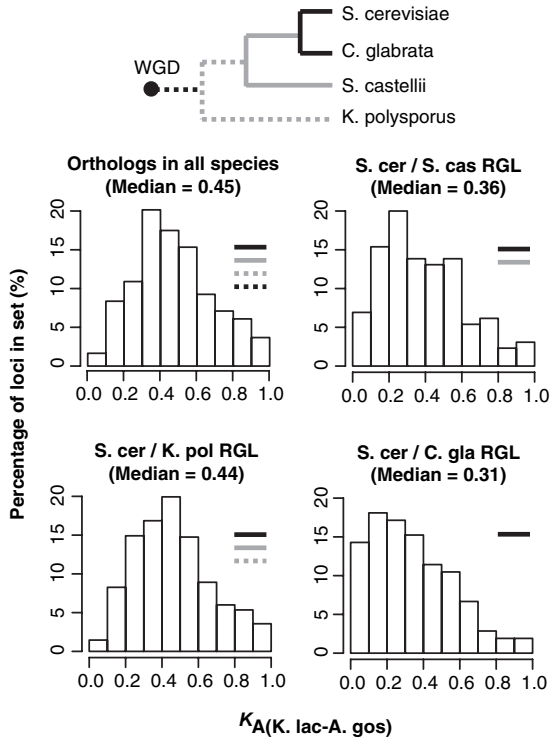


Fig. 4. RGL is restricted to slower-evolving loci at later time points. Histograms show the distribution of levels of nonsynonymous substitution (K_A) between *K. lactis* and *A. gossypii* (a proxy for rate of sequence evolution) for orthologs and sets of loci that have undergone RGL during different time intervals. The patterned lines beside each histogram show the branches of the phylogenetic tree (top) on which RGL could have occurred. RGL loci were always assigned to the most recent category possible. All data sets contain at least 100 loci, and all K_A distributions, except the two on the *Left*, differ significantly from one another ($0.0001 < P < 0.05$ by Wilcoxon rank-sum tests).

give rise to neutral gene losses, it is the only state that can lead to reciprocal gene loss (RGL, where two species lose alternative copies of the gene). Therefore we expect that the proportion of duplicated loci that are amenable to RGL will decrease as time elapses after WGD. Furthermore, because the accumulation of sequence divergence presumably tends to make gene pairs leave state U, we expect that the set of loci that remain in state U will gradually become enriched in slower-evolving loci. The model therefore predicts that loci that underwent RGL soon after WGD will tend to be a random subset of the genome, whereas more recent instances of RGL will tend to have been at more slowly evolving loci. We tested this hypothesis by partitioning RGL events into different time periods during the evolution of the post-WGD species, and indeed find that RGL events have become increasingly restricted to the slowest-evolving loci (Fig. 4). The loci that underwent RGL in the most recent interval, after *C. glabrata* and *S. cerevisiae* diverged, have a median rate of amino acid substitution that is only 70% of the median for loci that underwent RGL between *K. polysporus* and *S. cerevisiae*. A separate direct comparison between loci that underwent RGL and those that underwent convergent loss indicates that the former evolve significantly more slowly than the latter, thus excluding the possibility that this effect can be explained solely by a tendency for slower evolving loci to be resolved at later time points ($P = 0.006$ by Wilcoxon rank-sum test; *SI Appendix*, section 6). Furthermore, the loci that underwent RGL between *K. polysporus* and *S. cerevisiae* do not show any significant differences in GO categories compared with single-copy orthologs, contrary to what is seen for later RGL events (6).

We estimate that the proportion of gene losses that were partisan (i.e., losses from state C as opposed to state U) rose from 1% immediately after WGD to 40% for losses that occurred after the *S. bayanus–S. cerevisiae* speciation (Fig. 2*C* and *SI Appendix*, section 7). This increase can be explained by the accumulation of sequence divergence between the two gene copies, which will inevitably introduce selective differences between them and may cause them to have different deletion phenotypes (state C). The answer to our second question is therefore that initially there was little or no selective difference between the two gene copies, but that differences emerged quite quickly as the sequences diverged, which then caused particular gene copies to be favored for retention at single-copy loci. We note also that the fact that only low levels of partisan gene loss are estimated for the earliest time points after WGD indicates that the gene pairs were initially very similar in sequence. This inference in turn shows that the WGD event must have been an autopolyploidization or an allopolyploidization between two parental lineages with only minimal sequence divergence between them.

Conclusion

Our results show that the most recent common ancestor of *K. polysporus* and *S. cerevisiae* must have had >9,000 protein-coding genes. The two species show markedly convergent subsequent evolution, with both genomes shrinking to ≈5,600 protein-coding genes, and both retaining similar functional categories of genes in duplicate. That such similarities exist despite the fact that almost half of their single-copy genes are paralogs is remarkable and suggests that WGD provides unique evolutionary opportunities that can be capitalized upon in relatively predictable ways.

Materials and Methods

Draft Genome Sequence of *K. polysporus* DSMZ 70294. The type strain of *Kluyveromyces polysporus* (DSMZ 70294; synonym: *Vanderwaltozyma polyspora*) was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen and used to create genomic DNA libraries. A total of 101,838 sequence reads (79,976 reads from a plasmid library and 21,862 reads from a fosmid library) were assembled into 546 initial contigs by using the Phred (30) and Phrap (www.phrap.org) software. Assembly of scaffolds and annotation are described in the *SI Methods*.

YGOB and GO Analysis. We imported the *K. polysporus* genome annotation into our YGOB database, which also includes genome data from the post-WGD species *S. cerevisiae*, *S. bayanus*, *S. castellii*, and *C. glabrata*, and the non-WGD species *A. gossypii*, *K. lactis*, and *K. waltii* (7). The YGOB engine was then used to classify ancestral loci into different categories of gene loss or retention status, similar to ref (6). In this study, we worked with two data sets: 3,252 ancestral loci that can be reliably scored as either present or absent in both *K. polysporus* and *S. cerevisiae*, and 2,299 ancestral loci that can be reliably scored among *K. polysporus*, *S. cerevisiae*, *S. castellii*, *C. glabrata* and *S. bayanus*. GO terms associated with *S. cerevisiae* genes were downloaded from the *Saccharomyces* Genome Database (www.yeastgenome.org) in March 2006 and mapped to the 3,252 ancestral loci in the former data set to identify GO terms that are under- or overrepresented among double-copy relative to single-copy loci. Full details are provided in the *SI Methods*.

Phylogenetics. To estimate the timing of speciation events among *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *S. castellii*, and *K. polysporus* relative to the WGD (Fig. 2*C*) we constructed a superalignment from 11 loci that have been retained in duplicate in all five yeasts.

We used this superalignment and the WAG+G (8)+I+F model to evaluate branch-lengths under the known topology (6, 12) and calculated the relative distances from duplicate divergence to each speciation event. The selection of the 11 double-copy loci and additional details are provided in the *SI Methods*.

We thank D. Lundin and G. Manning for comments and the Science Foundation Ireland (SFI)/Higher Education Authority (HEA) Irish Centre for High-End Computing (ICHEC) for computational facilities. A.C.F. is a Swedish Research Council (Vetenskapsrådet) Fellow. This project was supported by SFI.

1. Wolfe KH, Shields DC (1997) *Nature* 387:708–713.
2. Kellis M, Birren BW, Lander ES (2004) *Nature* 428:617–624.
3. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, *et al*. (2004) *Science* 304:304–307.
4. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, *et al*. (2004) *Nature* 430:35–44.
5. Piskur J, Langkjaer RB (2004) *Mol Microbiol* 53:381–389.
6. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) *Nature* 440:341–345.
7. Byrne KP, Wolfe KH (2005) *Genome Res* 15:1456–1461.
8. Paterson AH, Bowers JE, Chapman BA (2004) *Proc Natl Acad Sci USA* 101:9903–9908.
9. Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, *et al*. (2005) *PLoS Biol* 3:e38.
10. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) *Proc Natl Acad Sci USA* 102:5454–5459.
11. Schranz ME, Mitchell-Olds T (2006) *Plant Cell* 18:1152–1165.
12. Kurtzman CP, Robnett CJ (2003) *FEMS Yeast Res* 3:417–432.
13. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al*. (1996) *Science* 274:546:563–567.
14. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) *Nature* 423:241–254.
15. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) *Science* 301:71–76.
16. Cliften PF, Fulton RS, Wilson RK, Johnston M (2006) *Genetics* 172:863–872.
17. Bon E, Neuveglise C, Lepingle A, Wincker P, Artiguenave F, Gaillardin C, Casaregola S (2000) *FEBS Lett* 487:42–46.
18. Casaregola S, Lepingle A, Bon E, Neuveglise C, Nguyen H, Artiguenave F, Wincker P, Gaillardin C (2000) *FEBS Lett* 487:47–51.
19. Wong S, Fares MA, Zimmermann W, Butler G, Wolfe KH (2003) *Genome Biol* 4:R10.
20. Wong S, Butler G, Wolfe KH (2002) *Proc Natl Acad Sci USA* 99:9272–9277.
21. van der Walt JP (1956) *Antonie van Leeuwenhoek* 22:265–272.
22. Roberts CJ, van der Walt JP (1959) *Compt Rend Lab Carlsberg* 31:129–148.
23. Sheeman B, Carvalho P, Sagot I, Geiser J, Kho D, Hoyt MA, Pellman D (2003) *Curr Biol* 13:364–372.
24. Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, *et al*. (2006) *Plant Cell* 18:1348–1359.
25. Lynch M, Force AG (2000) *Am Nat* 156:590–605.
26. Werth CR, Windham MD (1991) *Am Nat* 137:515–526.
27. Seoighe C, Gehring C (2004) *Trends Genet* 20:461–464.
28. He X, Zhang J (2005) *Curr Biol* 15:1016–1021.
29. Hughes AL, Friedman R (2003) *Genome Res* 13:794–799.
30. Ewing B, Hillier L, Wendl MC, Green P (1998) *Genome Res* 8:175–185.