LETTERS

Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts

Devin R. Scannell[1]*, Kevin P. Byrne[1]*, Jonathan L. Gordon[1], Simon Wong[1] & Kenneth H. Wolfe[1]

A whole-genome duplication occurred in a shared ancestor of the yeast species *Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata*. Here we trace the subsequent losses of duplicated genes, and show that the pattern of loss differs among the three species at 20% of all loci. For example, several transcription factor genes, including *STE12*, *TEC1*, *TUP1* and *MCM1*, are single-copy in *S. cerevisiae* but are retained in duplicate in *S. castellii* and *C. glabrata*. At many loci, different species have lost different members of a duplicated gene pair, so that 4–7% of single-copy genes compared between any two species are not orthologues. This pattern of gene loss provides strong evidence for speciation through a version of the Bateson–Dobzhansky–Muller mechanism, in which the loss of alternative copies of duplicated genes leads to reproductive isolation[1,2]. We show that the lineages leading to the three species diverged shortly after the whole-genome duplication, during a period of precipitous gene loss. The set of loci at which single-copy paralogues are retained is biased towards genes involved in ribosome biogenesis and genes that evolve slowly, consistent with the hypothesis that reciprocal gene loss is more likely to occur between duplicated genes that are functionally indistinguishable. We propose a simple, unified model in which a single mechanism—passive gene loss—enabled whole-genome duplication and led to the rapid emergence of new yeast species.

We used the Yeast Gene Order Browser (YGOB, ref 3) to compare six yeast species, three of which diverged after their common ancestor experienced a whole-genome duplication (WGD), and three of which diverged from this lineage before the WGD. The YGOB compares pairs of genomic regions from post-WGD species (*S. cerevisiae*[4,5], *S. castellii*[6] and *C. glabrata*[7]) to single genomic regions in pre-WGD species (*Kluyveromyces waltii*[8], *Kluyveromyces lactis*[7] and *Ashbya gossypii*[9]) (Fig. 1). We use the term 'ancestral locus' to describe a locus in a pre-WGD species, or the corresponding duplicated pair of loci in a post-WGD species (that is, a column in Fig. 1). Synteny conservation enabled us to determine unambiguously whether each of 2,723 ancestral loci was retained in one or two copies in each post-WGD genome. If only one copy remained, the syntenic context allowed orthologues to be distinguished from paralogues (Fig. 1).

The fate of an ancestral locus among the three post-WGD species can be classified into one of 14 possible patterns (Fig. 2). The most common pattern (Class 4, seen at 1,957 ancestral loci or 72% of the total) is that all three species have lost the same (orthologous) copy of the gene, such as in the *LYS2* column in Fig. 1. For clarity, we show this as three separate losses in Fig. 2, but a loss could have occurred in the ancestor of two or three of the species. Of the ancestral loci, 210 (8%) remain duplicated in all three post-WGD species (Class 0). The other 556 ancestral loci (20%) have had variable fates among the three post-WGD species, which indicates that the consequences of WGD were still being sorted out when these lineages diverged. A striking example is the set of 18 genes that are single-copy in *S. cerevisiae* but double-copy in both *S. castellii* and *C. glabrata* (Class 1B). Transcription factors are disproportionately over-represented in this group (it includes *STE12*, *TUP1*, *GAL11*, *GCR2*,

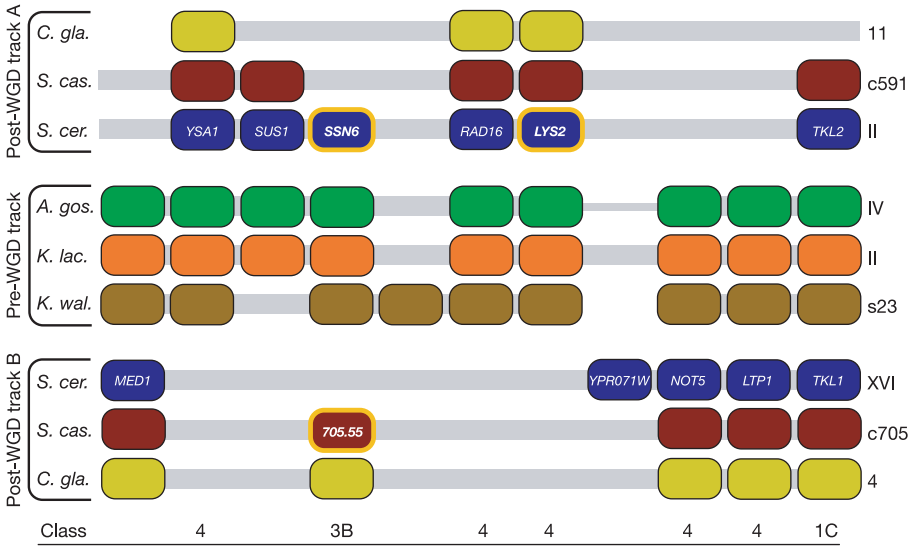

Figure 1 | Gene order relationships in the region around *S. cerevisiae SSN6* and its homologues. Relationships based on YGOB output (http://wolfe.gen.tcd.ie/ygob). Coloured boxes represent genes and are not drawn to scale. Chromosomal regions from each pre-WGD species are represented by one horizontal track each. The two corresponding regions in each post-WGD species are represented by two tracks (A and B) at the top and bottom. Chromosome, contig or scaffold numbers are indicated on the right. Homologous genes are arranged in columns. Thick grey horizontal bars connect genes that are immediate neighbours in the genome. Codes below columns indicate the gene loss class for that ancestral locus, as used in Fig. 2. Columns without codes did not meet the criteria for scoring.

[1]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland.
*These authors contributed equally to this work.

SFP1, *YAP3* and *TYE7*; $P = 0.001$ by Fisher's test), which suggests that the transcriptional regulatory network in *S. cerevisiae* is simpler than in the other yeasts (Supplementary Table S1). *MCM1* and *TEC1* are also in a 1:2:2 relationship among the post-WGD genomes, but these two loci were not counted in Fig. 2 because the syntenic context around them is not completely conserved.

S. cerevisiae SSN6 (also known as *CYC8*) and *S. castellii* gene *705.55* are an example of single-copy paralogues (Fig. 1). This situation arises when opposite members of a gene pair are lost in two daughter species. Between *S. cerevisiae* and *S. castellii*, 176 of the 2,723 loci we surveyed (6.4%; Classes 2E, 3A and 3B in Fig. 2) show this pattern of reciprocal gene loss. Reciprocal gene loss is a particular form of reciprocal silencing[1] or divergent resolution[2,10] of duplicated genes, and is a property of a pair of genomes. Similarly, there are 198 reciprocal gene-loss loci between *C. glabrata* and *S. castellii* (7.3%), and 100 between *S. cerevisiae* and *C. glabrata* (3.7%). Thus, a significant minority of genes that are mutual best BLASTP hits between the post-WGD genomes are not orthologues. More importantly, the process of reciprocal gene loss has the effect of changing the location of the functional copy of a gene[1,2]. For instance, *S. castellii* effectively carries a null allele at its locus orthologous to *SSN6*, and *S. cerevisiae* has a null allele orthologous to gene *705.55* (Fig. 1). If this were the only difference between these two species and they formed a hybrid, the hybrid would be likely to have low fitness because one-quarter of its spores would lack a functional copy of both *SSN6* and gene *705.55* (*S. cerevisiae ssn6* mutants are defective in respiratory growth and sporulation). In fact, 66 of the 176 loci that have undergone reciprocal gene loss between *S. cerevisiae* and *S. castellii* involve essential *S. cerevisiae* genes, so the spore viability of the hypothetical hybrid is reduced to approximately $(0.75)^{66}$ or 6×10^{-9} as a consequence of essential genes alone. Viability will be reduced further by reciprocal gene loss at loci that were not scored in Fig. 2 owing to inadequate synteny conservation (about half the genome), and at loci such as *SSN6* that are not essential but still contribute to fitness. The number of reciprocal losses observed among the post-WGD species is ample to account for their reproductive isolation, notwithstanding the contributions of mechanisms such as inter-chromosomal rearrangement[11,12] and mismatch repair[13,14].

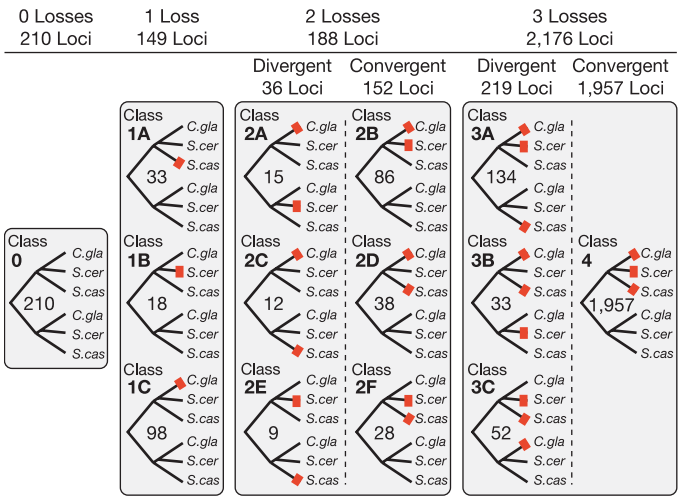

Figure 2 | Classes of gene loss pattern among 2,723 ancestral loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*, and their frequencies. Red marks denote gene absence and are used to group ancestral loci into 14 gene-loss classes, described by schematic trees showing the fates of orthologous and paralogous genes. The number of ancestral loci in each gene-loss class is shown in the centre of the tree. The two sets of species names in each tree denote tracks A and B in arbitrary order. In some cases, the absence of a gene copy in two or more species may be due to a single gene-loss event on a shared branch, but this does not affect classification. Convergent classes are those in which all genes lost are orthologues; divergent classes involve some loss of paralogues in different species.

The situation described above for the *SSN6* and *705.55* genes is a special case of Bateson–Dobzhansky–Muller (BDM) interspecific genomic incompatibility[15]. The BDM model proposes that negative epistatic interactions between two loci can reduce the fitness of a hybrid. Werth and Windham[1] and Lynch and Force[2] applied the BDM model to duplicated genes, hypothesizing that reciprocal loss (or silencing) of different copies in two species would create a BDM incompatibility, leading to reduced hybrid fitness. Reciprocal gene loss at multiple loci could lead to reproductive isolation, and where many duplicated genes exist (as in a polyploid) there is the potential for successive nested speciation events to occur[1,2,10].

To investigate whether reciprocal gene loss was involved in the establishment of reproductive isolation among the post-WGD lineages, we determined the timing of gene losses by estimating the number of duplicated genes surviving at each node on the lineage leading to *S. cerevisiae* (Fig. 3). To increase the resolution of this analysis we included data from *S. bayanus*[16], a close relative of *S. cerevisiae*. (Reproductive isolation between these two species is due to processes other than reciprocal gene loss[12].) We expressed the ages of the nodes as a proportion of the time (T) since the initial divergence of gene pairs created by WGD (see Supplementary Notes S2 and S3). We then estimated the numbers of genes still duplicated in the common ancestors of *S. cerevisiae* and each of *S. bayanus*, *C. glabrata* and *S. castellii* using two methods: parsimony (which gives the minimum number of genes that must have been retained in duplicate) and a model-based approach (Supplementary Note S4). We consider the latter to be more realistic because it allows for parallel gene losses in different lineages.

The parsimony and model-based methods both show a precipitous loss of duplicated genes in the time interval between WGD and the first speciation event (Fig. 3b, c). Both methods also show that the fraction of genes retained in duplicate declined appreciably (from 47% to 32% according to the model-based method) in the interval between the first (*S. castellii*) and the second (*C. glabrata*) speciation, even though this corresponds to a very short time period. From this we conclude that gene loss was still occurring rapidly during the emergence of the post-WGD lineages. Moreover, because reciprocal gene loss (by definition) cannot have occurred before *S. castellii* diverged from the other post-WGD lineages, and the number of gene losses on the right-hand side of the curve is very few (*S. bayanus* differs from *S. cerevisiae* at only two of the scored ancestral loci), the vast majority of reciprocal losses must have occurred at around the time of the two speciation events. In fact, we estimate that two-thirds of all reciprocal gene-loss events occurred between the time of *S. castellii* divergence and time $0.337T$ (Fig. 3b). The reproductive barriers imposed on these species by reciprocal gene loss are therefore not recent reinforcements but were erected contemporaneously with speciation.

The fate awaiting most gene pairs formed by WGD was that the duplication was resolved by deleting one gene copy (Fig. 2). If the two copies were functionally identical, we would expect the 'choice' of which copy to delete to be arbitrary. This hypothesis can be tested at ancestral loci that have been resolved independently in more than one post-WGD lineage. We find that in cases of two independent losses, the two retained genes are more often orthologues than paralogues (in Fig. 2, compare Class 2D to 2C, and 2F to 2E; χ^2-test of homogeneity, $P < 0.05$ for each). A possible explanation for the excess of convergent losses is that at some loci the two copies were not functionally identical, and that the same (better-functioning) copy was retained on both occasions. In contrast, the fact that divergent resolution is seen at some other loci suggests that the choice of survivor at those loci was arbitrary (Classes 2A, 2C, 2E and 3). These observations can be reconciled if some pairs of genes were functionally indistinguishable at the time the duplication was resolved (in which case either copy could be retained) but others were

functionally distinct (so that a particular copy was preferred by selection).

Differences in the performance of a function can only have been due to sequence differences between the gene copies themselves or between their *cis*-regulatory regions. This sequence divergence must have accumulated in the time between WGD and gene loss, or, if the WGD was an allopolyploidy, have been inherited from parental species. Therefore, neutral gene loss (which results in divergent resolution half of the time) is expected to be more frequent at ancestral loci that are slowly evolving or involved in highly conserved biological processes where the potential for functional divergence is low. We tested this prediction and indeed find that loci in Class 3 (all of which underwent reciprocal gene loss between two species) evolve on average 30% slower than Class 4 (for which no reciprocal gene loss occurred) (Supplementary Note S5; Wilcoxon rank-sum test $P < 10^{-14}$). Moreover, Gene Ontology terms such as 'ribosomal RNA processing', 'ribosome biogenesis' and 'RNA binding' are disproportionately over-represented among Class 3 loci, as are proteins that are localized to the nucleolus[17] and proteins in complexes that bind RNAs[18] (Supplementary Table S1 and Note S5). Finally, we also find that genes for small nucleolar (sno)RNAs, many of which function in ribosomal RNA processing, have undergone reciprocal gene loss unusually frequently (Supplementary Note S5). Thus, the set of reciprocal gene-loss loci seems biased towards those with functions most likely to be conserved between duplicates. This functional bias increases the potential contribution of reciprocal gene-loss loci to reproductive isolation, because 40% of the Class 3 loci are essential[19] in *S. cerevisiae*, compared with 20% of Class 4 loci ($P < 10^{-10}$, χ^2-test).

The passive loss of genes from genomes in which there is no selection to retain them is a familiar phenomenon in molecular evolution[20,21]. We suggest that passive gene loss is the likely mechanism of the original WGD event in yeast. Our model (Fig. 4) begins with two haploid cells fusing to form a diploid. If the haploids are from different species, or differ by a chromosomal rearrangement, or carry particular mutations, the resulting diploid may be unable to form viable spores but still able to divide mitotically. If the diploid cell lineage continues to divide mitotically for many generations, it can start to lose one allele from every locus that is not haploinsufficient. During this process, there is nothing to prevent an allele at the *MAT* locus from being deleted, in which case the cell will behave as a haploid. It can switch mating type, undergo mother–daughter mating, form a diploid, and so regain fertility[22]. Former alleles become separate loci, each of which is homozygous. Continuing loss of redundant gene copies will result in separate lineages that are self-fertile but reproductively isolated from one another by reciprocal gene loss (Fig. 4).

Our results provide evidence that reciprocal gene loss at multiple ancestrally duplicated genes may lead to speciation, as has previously been hypothesized (but not demonstrated) for polyploid plants[1,23] and fish[10,24]. Indeed, because we have shown that reciprocal gene loss is implicated in the emergence of three different lineages, our data support the feature of the modified BDM mechanism[1,2] that most distinguishes it from other theories of reproductive isolation: the ease with which it accounts for multiple speciation events. Finally, by showing that slowly evolving genes and those involved in very fundamental processes are the ones most likely to undergo reciprocal

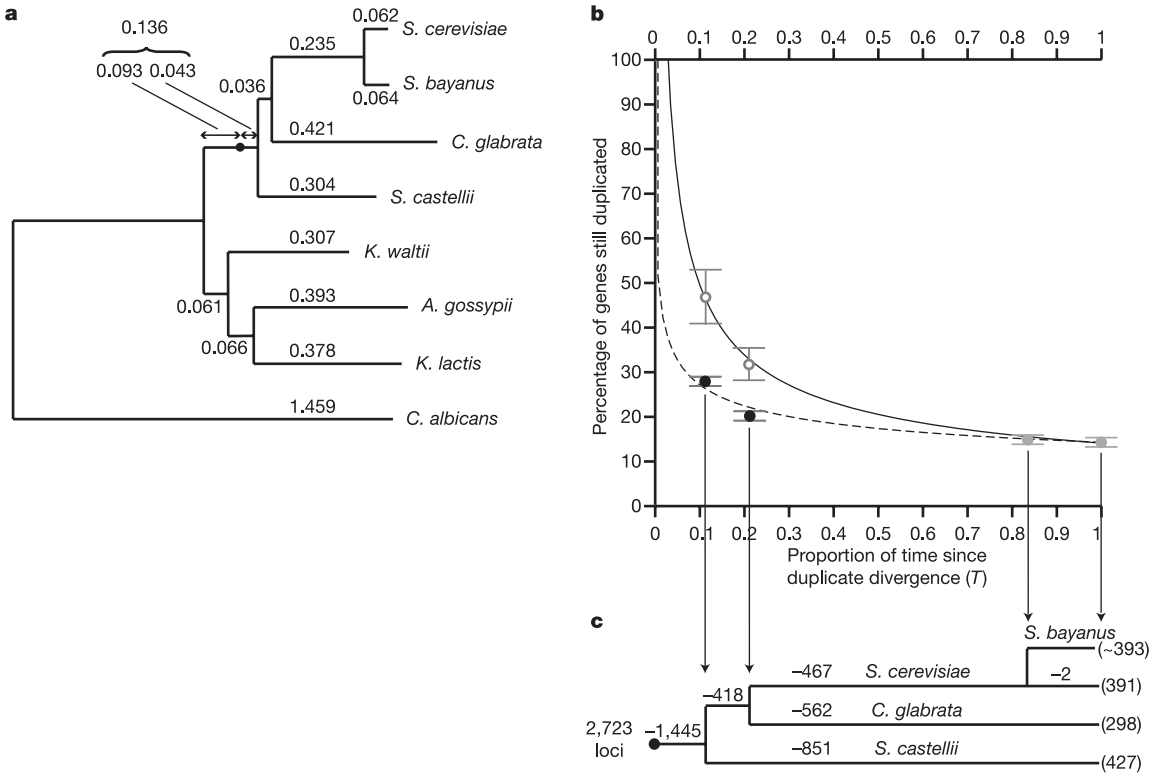

Figure 3 | Time course of duplicated gene loss following WGD. a, Tree reconstructed from 909 protein sequences using a constrained topology (Supplementary Note S2) and branch-length estimation by maximum likelihood. The black dot indicates the initial divergence of duplicates created by WGD (Supplementary Note S3). **b**, Gene-loss curves estimated by the model-based method (open circles and solid curve; Supplementary Note S4) and by parsimony (black circles and dashed curve). Filled grey circles are common to both methods and show the percentages of loci duplicated in *S. cerevisiae* and its common ancestor with *S. bayanus*. The horizontal scale represents the time from the initial divergence of duplicates created by WGD ($0T$) to the present ($1T$) and is derived from the tree in **a** assuming a molecular clock (Supplementary Note S3). Power-law curves were fitted to the data[25]. Standard errors for x (all <2%; omitted for clarity) and y values were estimated by bootstrapping. **c**, Numbers of genes lost on each branch leading to post-WGD species, as inferred by the model-based method. The current numbers of duplicates remaining in each post-WGD genome are shown in parentheses. All numbers refer to the 2,723 loci summarized in Fig. 2.

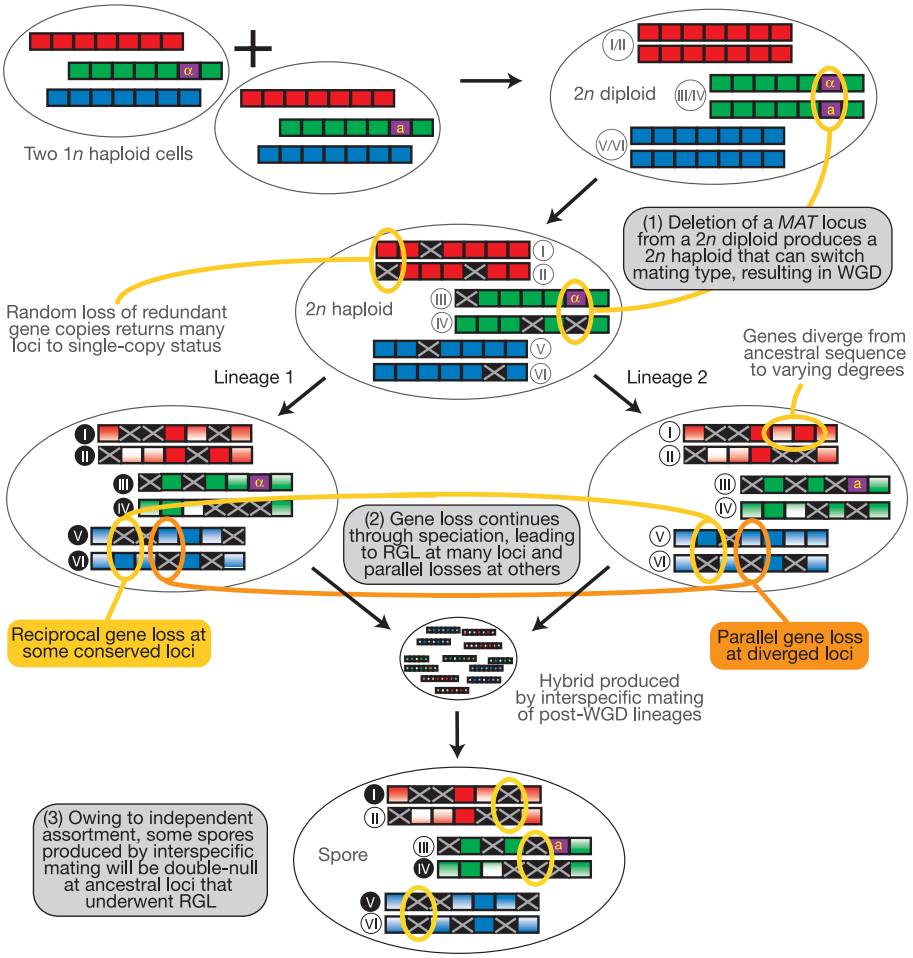

Figure 4 | Model of passive gene loss as a mechanism for WGD and establishment of reproductively isolated lineages. Individual steps are discussed in the main text. Ovals represent yeast cells. Genes are shown as red, green or blue boxes (except for the *MAT* locus, shown in purple), and are arranged horizontally as chromosomes. Grey 'X' symbols indicate genes that have been deleted. Roman numbering of chromosomes is used to indicate the parent of origin where relevant. Features relevant to each step are circled in yellow or orange. RGL, reciprocal gene loss.

gene loss, our study leads to the conclusion that these genes, which individually are among the most conservative in the genome, may collectively be responsible for the most radical of evolutionary events.

METHODS

We used the YGOB engine[3] to assess the status and syntenic conservation of loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*. Each ancestral locus (that is, a column in Fig. 1, corresponding to two genomic sites in post-WGD species and one site in pre-WGD species) was scored up to 18 times: on tracks A and B in each of the three post-WGD species, and comparing against each of the three pre-WGD genomes. On the basis of homology and syntenic context, the status of each of the six genomic sites in the post-WGD species was designated as one of (1) gene unambiguously present, (2) gene unambiguously absent, (3) gene present but with insufficient syntenic support, or (4) gene absent but with insufficient syntenic support. Loci were retained for further analysis if presence or absence could be determined unambiguously on both tracks in all three post-WGD species and if the scoring against all three pre-WGD genomes was not contradictory. This yielded reliable information for 2,723 ancestral loci, as summarized in Fig. 2. The scoring protocol and our implementation are described in ref. 3. We ignored a small number of ancestral loci for which one of the post-WGD species retained neither gene copy. *S. bayanus* was scored relative to the 2,723 ancestral loci in *S. cerevisiae* because their genomes are almost completely co-linear. In *S. bayanus*, 2,631 loci had conserved syntenic context (by the criteria above), and manual inspection of candidates generated by the YGOB engine revealed just two differences between *S. bayanus* and *S. cerevisiae* (*S. bayanus* has retained paralogues as well as orthologues of *HEK2* and *YAT1*).

Received 5 August; accepted 22 December 2005.

1. Werth, C. R. & Windham, M. D. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *Am. Nat.* **137,** 515–526 (1991).
2. Lynch, M. & Force, A. G. The origin of interspecies genomic incompatibility via gene duplication. *Am. Nat.* **156,** 590–605 (2000).
3. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15,** 1456–1461 (2005).
4. Goffeau, A. *et al.* The Yeast Genome Directory. *Nature* **387** (Suppl.), 5–105 (1997).
5. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
6. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
7. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
8. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).
9. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304,** 304–307 (2004).
10. Taylor, J. S., Van de Peer, Y. & Meyer, A. Genome duplication, divergent resolution and speciation. *Trends Genet.* **17,** 299–301 (2001).
11. Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res.* **11,** 2009–2019 (2001).
12. Delneri, D. *et al.* Engineering evolution to study speciation in yeasts. *Nature* **422,** 68–72 (2003).
13. Greig, D., Louis, E. J., Borts, R. H. & Travisano, M. Hybrid speciation in experimental populations of yeast. *Science* **298,** 1773–1775 (2002).
14. Hunter, N., Chambers, S. R., Louis, E. J. & Borts, R. H. The mismatch repair

system contributes to meiotic sterility in an interspecific yeast hybrid. *EMBO J.* **15,** 1726–1733 (1996).
15. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).
16. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
17. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
18. Krogan, N. J. *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell* **13,** 225–239 (2004).
19. Guldener, U. *et al.* CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.* **33,** D364–D368 (2005).
20. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc. Natl Acad. Sci. USA* **101,** 14144–14149 (2004).
21. Wolfe, K. H., Morden, C. W. & Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl Acad. Sci. USA* **89,** 10648–10652 (1992).
22. Greig, D., Borts, R. H., Louis, E. J. & Travisano, M. Epistasis and hybrid sterility in *Saccharomyces*. *Proc. R. Soc. Lond. B* **269,** 1167–1171 (2002).
23. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101,** 9903–9908 (2004).
24. Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y. L. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet.* **20,** 481–490 (2004).
25. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102,** 5454–5459 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to D. Bradley, G. Conant, J. Conery, B. Cusack, H. Fraser, N. Khaldi, W.-H. Li, A. Lloyd, A. McLysaght, G. Singer, R. Vilgalys and M. Woolfit for discussion. This study was supported by Science Foundation Ireland.

Author Contributions S.W. and K.H.W. did pilot studies. K.P.B. developed the YGOB and the scoring scheme. D.R.S. did all analyses except those of gene function (K.P.B.) and chromosomal rearrangements (J.L.G.). Manual editing of data in YGOB was done equally by all authors. D.R.S. and K.H.W. wrote the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.H.W. (khwolfe@tcd.ie).