Rate Asymmetry After Genome Duplication Causes Substantial Long-Branch Attraction Artifacts in the Phylogeny of *Saccharomyces* Species

Mario A. Fares,† Kevin P. Byrne,* and Kenneth H. Wolfe**

*Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland; and †Department of Biology, National University of Ireland, Maynooth, County Kildare, Ireland

Whole-genome duplication (WGD) produces sets of gene pairs that are all of the same age. We therefore expect that phylogenetic trees that relate these pairs to their orthologs in other species should show a single consistent topology. However, a previous study of gene pairs formed by WGD in the yeast *Saccharomyces cerevisiae* found conflicting topologies among neighbor-joining (NJ) trees drawn from different loci and suggested that this conflict was the result of "asynchronous functional divergence" of duplicated genes (Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421**:848–852). Here, we test whether the conflicting topologies might instead be due to asymmetrical rates of evolution leading to long-branch attraction (LBA) artifacts in phylogenetic trees. We constructed trees for 433 pairs of WGD paralogs in *S. cerevisiae* with their single orthologs in *Saccharomyces kluyveri* and *Candida albicans*. We find a strong correlation between the asymmetry of evolutionary rates of a pair of *S. cerevisiae* paralogs and the topology of the tree inferred for that pair. *Saccharomyces cerevisiae* gene pairs with approximately equal rates of evolution tend to give phylogenies in which the WGD postdates the speciation between *S. cerevisiae* and *S. kluyveri* (B-trees), whereas trees drawn from gene pairs with asymmetrical rates tend to show WGD pre-dating this speciation (A-trees). Gene order data from throughout the genome indicate that the "A-trees" are artifacts, even though more than 50% of gene pairs are inferred to have this topology when the NJ method as implemented in ClustalW (i.e., with Poisson correction of distances) is used to construct the trees. This LBA artifact can be ameliorated, but not eliminated, by using gamma-corrected distances or by using maximum likelihood trees with robustness estimated by the Shimodaira-Hasegawa test. Tests for adaptive evolution indicated that positive selection might be the cause of rate asymmetry in a substantial fraction (19%) of the paralog pairs.

Introduction

Anciently polyploid genomes offer a unique opportunity to study genome evolution because of the large sets of simultaneously duplicated genes they contain. The bakers' yeast *Saccharomyces cerevisiae* is a degenerate polyploid that underwent whole-genome duplication (WGD) and subsequent rearrangements and gene loss (Wolfe and Shields 1997; Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004). More than 500 pairs of genes (ohnologs) in the *S. cerevisiae* genome were formed simultaneously by the WGD event. In some cases, the duplicated pairs have highly similar sequences (e.g., duplicated genes for ribosomal proteins), and the reason why both copies of the gene were retained seems to be selection for increased expression and consequent rapid cell growth. In other cases, the ohnologs have undergone successful functional divergence, and several possible examples of neofunctionalization have been proposed where one sequence retains an apparently ancestral function and the other has a derived function (Kellis, Birren, and Lander 2004; but see Lynch and Katju 2004). In general, functional divergence in one gene copy after duplication is correlated with an increased rate of fixation of amino acid substitutions in that copy (Kellis, Birren, and Lander 2004).

Conflicting scenarios have been proposed regarding the timing and outcome of the genome duplication seen in *S. cerevisiae*. In particular, Langkjaer et al. (2003) proposed, based on phylogenetic analysis of the sequences of 38 randomly chosen ohnolog pairs and their homologs in other species, that a single polyploidy event took place *before* the divergence between *S. cerevisiae* and several distantly related yeasts such as *Saccharomyces kluyveri* and *Kluyveromyces lactis*. This proposal contradicted analyses based on comparisons of chromosomal gene order, which had led to the conclusion that the polyploidization event occurred in the *S. cerevisiae* lineage *after* the lineages leading to *S. kluyveri* and *K. lactis* had branched off (Keogh, Seoighe, and Wolfe 1998; Wong, Butler, and Wolfe 2002). That *K. lactis* is a "pre-WGD" species (i.e., one that diverged from the lineage leading to *S. cerevisiae* before the latter underwent WGD) has now been fully confirmed through complete genome sequencing, which shows a 2:1 intercalation of gene orders between pairs of chromosomal regions in *S. cerevisiae* and single regions in *K. lactis* (Dujon et al. 2004). We show here that gene order in the *S. kluyveri* draft genome sequence (Cliften et al. 2003) confirms it as a pre-WGD species too.

Although the placement of the WGD event in the yeast phylogenetic tree appears to have been largely settled by the new genome sequence data, we remained puzzled as to why the phylogenetic analyses of Langkjaer et al. (2003) showed so many trees where an *S. kluyveri* sequence appeared to be more closely related to one member of an *S. cerevisiae* ohnolog pair than to the other, instead of appearing as an outgroup to the pair. Because the WGD took place 100 or more MYA (Wolfe and Shields 1997; Friedman and Hughes 2001), it is likely that at least some of the preserved gene copies experienced strong changes in their selective constraints during the course of diverging from their ancestral functions. This poses the question of whether the conflict between some phylogenetic trees and the inferences from gene order are due to phylogenetic artifacts that may be correlated with changes in selective constraints after gene duplication.

In this study, we show that conflicting phylogenies from different ohnologs in the *S. cerevisiae* genome are

Key words: accelerated substitution rates, genome duplication, *Saccharomyces*, long-branch attraction.

E-mail: khwolfe@tcd.ie.

Mol. Biol. Evol. 23(2):245–253. 2006
doi:10.1093/molbev/msj027
Advance Access publication October 5, 2005

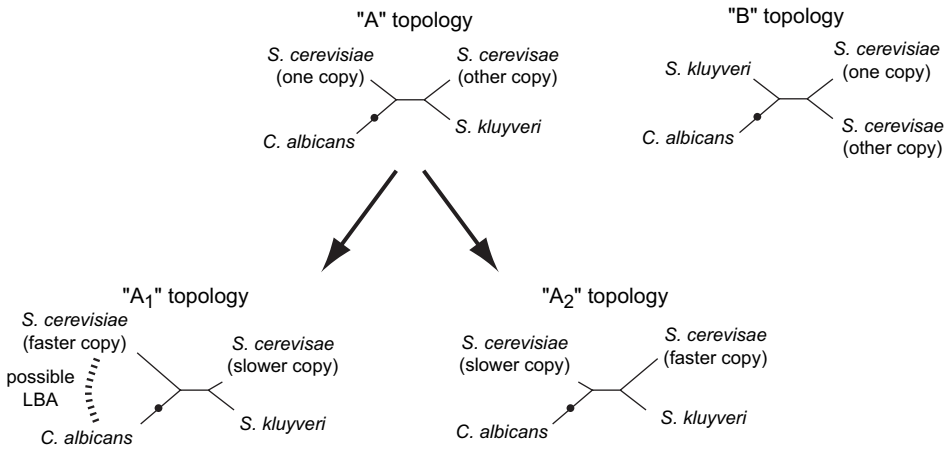

FIG. 1.—Nomenclature scheme for topologies obtained from sets of four sequences (two from *Saccharomyces cerevisiae* and one each from *Saccharomyces kluyveri* and *Candida albicans*). The "A" and "B" nomenclatures were used by Langkjaer et al. (2003). We further subdivided the "A" topology into "A_1" and "A_2" according to whether the *S. kluyveri* sequence clusters with the slower or the faster *S cerevisiae* sequence. The dot represents the position of the presumed root of the tree.

the result of strong long-branch attraction (LBA) effects in duplicated loci where the evolutionary rates are asymmetric and thus that there is no real discrepancy between the conclusions from gene order and molecular phylogenetic methods. LBA is one of the most important factors hampering accurate phylogenetic inference, particularly when both orthologous and paralogous sequences are used in the same tree (Brinkmann and Philippe 1999). The LBA effect consists of an attraction between distantly related sequences that lie on long branches, irrespective of the underlying phylogenetic relationship among them (Felsenstein 1978). We also demonstrate that selective constraints have changed after gene duplication and allowed the fixation of amino acid substitutions by positive selection in many of the gene copies. These changes are likely to have been involved in neofunctionalization of genes and the emergence of metabolic novelties in bakers' yeast.

Materials and Methods

Data

We began with a set of 455 pairs of duplicated *S. cerevisiae* genes that we had previously identified as having been formed by WGD (gene pairs on minor diagonals in the study of Wong, Butler, and Wolfe 2002). This set is very similar to the lists of paralog pairs recently published by Kellis, Birren, and Lander (2004) and Dietrich et al. (2004) (a reconciliation of these lists is given in Byrne and Wolfe 2005). We initially assembled quartets of sequences: an ohnolog pair from *S. cerevisiae* with its orthologs in *S. kluyveri* and *Candida albicans* (fig. 1). We considered only matches showing BlastP $E \leq 10^{-10}$ and aligned sequence lengths of at least 150 amino acids (ensuring enough phylogenetic signal), which led to 433 gene pairs being retained for further analysis. For each ohnolog pair in *S. cerevisiae*, there was only a single likely ortholog in *S. kluyveri*. These sequence sets were used to examine asymmetry of evolutionary rate. In searching for cases of positive selection, we included data from *Saccharomyces bayanus* (Kellis et al. 2003), *Saccharomyces castellii* (Cliften et al. 2003), and, where available, *K. lactis* and *Yarrowia lipolytica* (using data from Génolevures-1 (Souciet et al. 2000; this analysis was done before publication of their complete genome sequences). To compare gene order between *S. kluyveri* and other species, the available 4x genome sequence data from *S. kluyveri* was imported into the Yeast Gene Order Browser as described (Byrne and Wolfe 2005).

Sequence Alignments and Phylogenetic Tree Inference

Amino acid sequences for each gene set were aligned using T-COFFEE (Notredame, Higgins, and Heringa 2000) and carefully revised using the GENEDOC program (Nicholas, Nicholas, and Deerfield 1997). Nucleotide sequences were then aligned by concatenating triplets according to the amino acid sequence alignment. In order to replicate the study of Langkjaer et al. (2003), we first inferred phylogenetic trees by the neighbor-joining method (NJ; Saitou and Nei 1987) as implemented in MEGA3 (Kumar, Tamura, and Nei 2004) using Poisson correction for amino acid distance estimates. Then, we inferred a second set of NJ trees from gamma-corrected amino acid distances, using the empirical amino acid substitution matrix of Whelan and Goldman (2001). The reason for using the gamma correction is that Poisson correction of amino acid distances accounts for multiple hits but does not take into account variation of substitution rates among amino acid sites, which is one of the factors causing LBA effects. Moreover, retained gene copies are expected to show divergence from the ancestral function and hence are likely to have a heterogeneous distribution of substitution rates among sites.

Comparison of the Rates of Evolution Between Gene Copies and Detection of LBA

We first inferred phylogenetic trees for the complete set of *S. cerevisiae* gene pairs with their homologs in *S. kluyveri*, rooting the trees with homologous *C. albicans*

sequences and using Poisson-corrected amino acid distances. These trees were classified as either A-type or B-type topologies (fig. 1), following the nomenclature of Langkjaer et al. (2003). The second step was the construction of NJ trees with correction of amino acid distances by the gamma distribution. We classified the different gamma-corrected trees as A_1-, A_2-, or B-type trees (fig. 1). Types A_1 and A_2 are two subsets of the A-type trees of Langkjaer et al. (2003), the distinction being that in A_1 trees the *S. kluyveri* sequence clusters with the slower evolving of the two *S. cerevisiae* ohnologs, whereas in A_2 trees the *S. kluyveri* gene clusters with the faster *S. cerevisiae* ohnolog. As before, B type refers to trees where the two *S. cerevisiae* sequences group together, with *S. kluyveri* lying outside. We evaluated the likelihood of each tree by the CODEML program from the PAML package (Yang 1997), using the three types of trees (A_1, A_2, and B) as initial trees and assuming a gamma distribution model for the amino acid substitution rates. We decided which of the three trees is correct by choosing the one with the highest log-likelihood value. Third, we considered only loci where the tree topology with the highest likelihood was significantly better than the other topologies by the test of Shimodaira and Hasegawa (1999).

To test the hypothesis that LBA effects can occur after gene duplication, for each ohnolog pair we used the ratio between the amino acid distances of the two *S. cerevisiae* gene copies to *C. albicans* as an approximate measure of the asymmetry in the rates of evolution of the two *S. cerevisiae* copies. This ratio was calculated as

$$R = \frac{\mathrm{Max}(d_{Ca-Sc1}, d_{Ca-Sc2})}{\mathrm{Min}(d_{Ca-Sc1}, d_{Ca-Sc2})},$$

where the numerator and denominator are, respectively, the higher and the lower of the two amino acid distances from *C. albicans* to an *S. cerevisiae* gene. This measure of asymmetry was chosen because it is independent of the topology of the phylogenetic tree. If both gene copies evolved at the same rate, then we expect $R = 1$. We plotted the cumulative numbers of A-type and B-type trees against R for all the duplicated genes, sorted by their R values. Thereafter, we repeated the same procedure distinguishing A_1 and A_2 trees tested under the maximum likelihood (ML) method.

Analyzing Selective Constraints in *Saccharomyces* After Gene Duplication

We used ML-based models to test for adaptive evolution in the duplicated genes. We first applied models to detect adaptive evolution at single-codon sites. These models were the discrete models 0 (M0; Goldman and Yang 1994), neutral model 1 (M1), positive selection model 2 (M2), and positive selection model 3 (M3) (Yang et al. 2000). All models are implemented in the program CODEML from the PAML package version 3.0. M0 assumes an equal nonsynonymous-to-synonymous rates ratio ($\omega = d_N/d_S$) for all the branches of the phylogeny and sites of the nucleotide sequence alignment (a single estimated ω value). Model M1 is the neutral model and assumes two classes of sites in the protein: the conserved sites ($\omega = 0$) and the strictly neutral sites ($\omega = 1$). M2 adds a third class of sites to M1 where ω is estimated from the data and hence allows for detecting adaptive evolution. M3 allows the estimation of different categories of codon sites based on the estimation of ω (i.e., we fixed three main categories of codon sites with different substitution ratios).

Because we expect that adaptive evolution may have occurred in some genes after duplication, we also tested the free-ratio model for significance. This model allows the free estimation (from the data) of different ω values for different lineages in the tree, although it assumes one average ω estimate across all codons in the sequence alignment. Hence, the free-ratio model allows us to test whether adaptive evolution occurred on the faster or the slower *S. cerevisiae* branch. Nested models (models with nested number of parameters) can be compared using the likelihood ratio test (LRT; Huelsenbeck and Crandall 1997). This is because twice the difference between the log-likelihood values for nested models follows a χ^2 distribution, with the degrees of freedom being the difference in the number of parameters between the nested models. Under this assumption, we compared model M1 to M2, model M0 to M3, and the free-ratio model to M0, with the degrees of freedom being 2, 4, and the number of branches in the tree minus 1, respectively.

Results

LBA Is the Probable Cause of Conflicting Topologies

We examined 433 sets of four sequences, each consisting of a duplicated gene pair from *S. cerevisiae* and single homologs from *S. kluyveri* and *C. albicans*. All the *S. cerevisiae* duplicated gene pairs were pairs inferred to have been formed by WGD (ohnologs) because they are located in paired regions of the *S. cerevisiae* genome that are in a double-conserved synteny relationship with pre-WGD outgroup species (Wong, Butler, and Wolfe 2002; Dietrich et al. 2004; Kellis, Birren, and Lander 2004).

We first drew phylogenetic trees for each set using the NJ method with Poisson-corrected amino acid distances, replicating the method used by Langkjaer et al. (2003). As in their study, we classified the resulting tree topologies as either A type or B type. A-type trees are those in which one of the *S. cerevisiae* gene copies groups with the *S. kluyveri* homolog to the exclusion of the second *S. cerevisiae* gene copy. B-type trees are those where the two *S. cerevisiae* copies cluster together to the exclusion of their *S. kluyveri* homolog (upper part of fig. 1). The 433 genes produced 234 A-type trees (54%) and 199 B-type trees (46%), a mixed result similar to the finding of Langkjaer et al. (2003) of 19 A-type and 14 B-type trees for *S. kluyveri* in the set of genes they examined (ignoring trees with low bootstrap values in their study). Because the apparent WGD in an ancestor of *S. cerevisiae* was a single event (Wolfe and Shields 1997; Langkjaer et al. 2003; Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004), all the gene pairs formed by it are expected to have a homogeneous phylogenetic relationship to orthologous genes in other yeast species. They should either all be A-trees (if the WGD pre-dated the speciation between *S. cerevisiae* and *S. kluyveri*) or all be B-trees (if

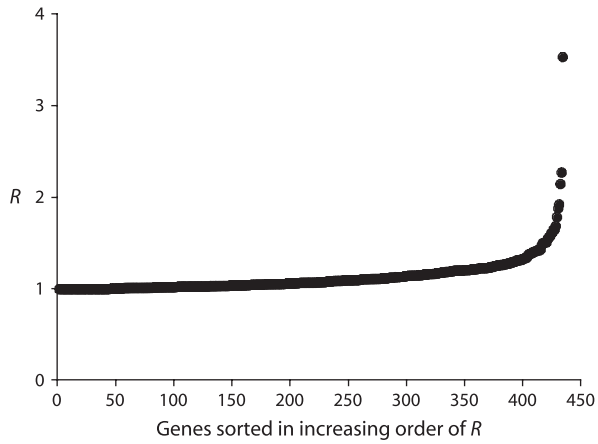

FIG. 2.—Distribution of values of the ratio (R) between the amino acid distances of the two *Saccharomyces cerevisiae* gene copies to the outgroup *Candida albicans*, among genes sorted in order of increasing R. The ratio was obtained by dividing the larger distance by the smaller distance, ensuring $R \geq 1$.

the WGD post-dated the speciation). Therefore, the fact that the NJ/Poisson method returns two different tree topologies (A and B) among the 433 WGD gene pairs can only be explained in one of two ways: either the NJ/Poisson method is performing poorly and returning artifactual trees for approximately half of the genes in this data set or the genes have been affected by an uncharacterized biological process (such as asynchronous functional divergence, as suggested by Langkjaer et al. 2003) that might cause the gene duplications to seem asynchronous, even though they were not.

We estimated the asymmetry of evolutionary rates in the two *S. cerevisiae* copies of each gene by dividing the amino acid distance from *C. albicans* to the faster *S. cerevisiae* copy by the distance to the slower one. This ratio (R) is an underestimate of the true level of inequality in rates between the two *S. cerevisiae* copies because the two distance terms include a shared component corresponding to the distance between *C. albicans* and the common ancestor of the two *S. cerevisiae* copies. However, it is a useful measure because it avoids making any assumption about whether the speciation between *S. kluyveri* and *S. cerevisiae* pre-dates or postdates the WGD. The rate ratio calculated in this way shows high variation among genes (fig. 2). In fact, only 179 of the 433 duplicated pairs (41%) have $R \leq 1.05$. Thus, most genes have had at least moderately asymmetrical divergence between the duplicated copies at the amino acid sequence level.

Using an approach similar to that of the analysis of Microsporidia of Thomarat, Vivarès, and Gouy (2004), we plotted the cumulative number of phylogenetic trees of each topology (A type or B type) against the gene pairs sorted in increasing order of their R values. Remarkably, the two types of trees do not accumulate in a linear fashion. Instead, they form curves whose slopes change as R increases (fig. 3*A*). The curve for B-trees is convex (i.e., it tends toward a plateau at high R values), whereas the curve for A-trees is concave (i.e., the rate of accumulation of A-trees increases as R increases). To test the significance of these relationships, we fitted linear and quadratic models

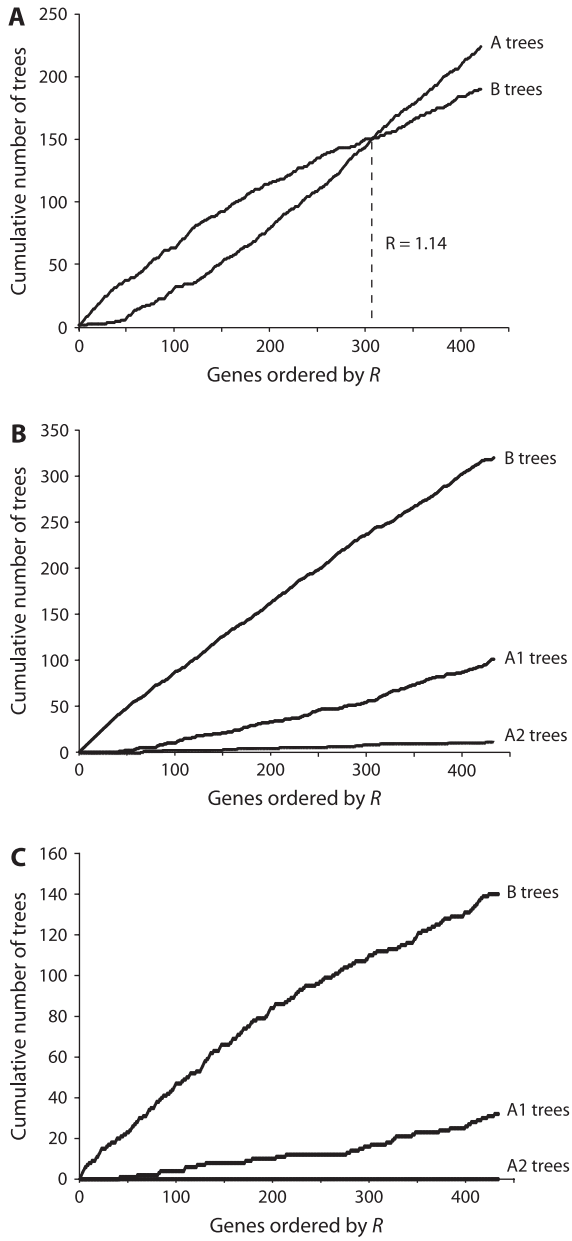

FIG. 3.—Cumulative numbers of A-type and B-type trees in genes, sorted in increasing order of R. (*A*) Trees drawn using Poisson-corrected amino acid distances, divided into those where *Saccharomyces kluyveri* clusters with one *Saccharomyces cerevisiae* ohnolog (A-trees) and those where the two *S. cerevisiae* ohnologs cluster together (B-trees). (*B*) Trees drawn using gamma-corrected amino acid distances, divided into B-trees and two subcategories of A-trees: those where *S. kluyveri* clusters with the slower *S. cerevisiae* ohnolog (A_1) and those where *S. kluyveri* clusters with the faster *S. cerevisiae* ohnolog (A_2). (*C*) Trees drawn using ML distances and that passed the SH test, also divided into topologies A_1, A_2, and B.

to the curves and compared the goodness of fit using analysis of variance. The comparison was performed using a *t*-test where

$$t_1 = \frac{(SS_Q - SS_L)}{\sqrt{SR_Q}}.$$

Here, t_1 stands for the Student t value using 1 degree of freedom; SS_Q and SS_L are the sums of squares explained

by the quadratic and linear regression models, respectively, and SR_Q is the sum of residuals for the quadratic regression model. For both types of trees, the quadratic model is a significantly better fit than the linear model (A-trees: $R_Q = 0.999$, $t_1 = 393$, $P < 0.001$; B-trees: $R_Q = 0.998$, $t_1 = 374$, $P < 0.001$). The curve for A-type trees has a positive slope whereas that of B-type trees has a negative slope, which implies that there are more A-trees and fewer B-trees than expected as *R* values increase. The rapid increase in the numbers of A-type trees at high *R* values is consistent with the hypothesis that LBA between the outgroup (*C. albicans*) and the faster evolving copy in *S. cerevisiae*, which produces topology type A, is occurring at high *R* values. The effect is especially evident when *R* values are higher than 1.14, which is the point where the cumulative number of A-type trees exceeds that of B-type trees (fig. 3*A*).

Phylogenetic trees based on amino acid distances that are corrected using a gamma distribution should be less sensitive to LBA effects. Using gamma-corrected distances instead of Poisson distances caused many genes to switch tree topologies from A type to B type, resulting in 113 A-trees and 320 B-trees overall. We further subdivided the gamma-corrected A-trees into types A_1 and A_2, where A_1 is the topology where the *S. kluyveri* sequence clusters with the slower of the *S. cerevisiae* copies and A_2 is the topology that groups *S. kluyveri* with the faster copy in *S. cerevisiae* (fig. 1). Of the 433 trees from duplicated genes, 74% (320 genes) were B type, 23% (101 genes) were A_1 type, and only 3% (12 genes) were A_2 type (fig. 3*B*). If some biological effect rather than LBA is responsible for the mixture of A-type and B-type trees, then we would expect an equal distribution of trees between A_1 and A_2 types, which is obviously not the case. The excess of A_1 over A_2 trees suggests that, in gene pairs where there is high asymmetry of rates, LBA is causing the faster *S. cerevisiae* copy to cluster with the outgroup, leaving the slower *S. cerevisiae* copy clustered with *S. kluyveri* and giving the A_1 topology (fig. 1). Only 12 duplicated genes gave the A_2 topology, and, notably, these genes were the ones where *S. kluyveri* also presented higher rates of substitutions compared to the mean branch length of the tree and to their *S. cerevisiae* homologs, which suggests that LBA between *S. kluyveri* and the faster *S. cerevisiae* copy may have been the cause of an artifactual tree.

Additional support for the hypothesis that unequal rates of evolution are causing artifactual trees comes from the observation that the relationship between *R* and bootstrap percentage (BP) support is different for different topologies. For B-trees there is a strong tendency for BP support to decline as *R* increases (for the NJ method with gamma correction, nonparametric Spearman's $\rho = -0.469$, $P \ll 0.01$), whereas there is no significant correlation between BP and *R* for A_1- or A_2-trees ($\rho = +0.068$ and $+0.095$, respectively). When the data is grouped into sets of 10 trees with the same topology and similar *R* values (fig. 4), it is apparent that as *R* values increase, not only do B-trees become rarer but those that are present tend to be more poorly supported. In contrast, the group of A_1-trees with the highest BP support is the group with the highest *R* values (fig. 4).

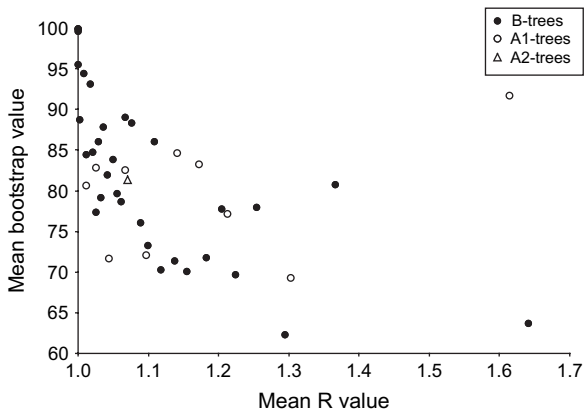

FIG. 4.—Relationship between *R* values and bootstrap support for phylogenetic trees. For each tree topology, genes were sorted in increasing order of *R* values, and the mean BP and *R* value were then calculated for successive groups of 10 genes. The numbers of groups of genes for the B, A_1, and A_2 topologies are 32, 10, and 1, respectively. Trees were reconstructed using the NJ method with gamma correction.

In a final approach to examining the relationship between *R* and support for tree topologies, we used ML and considered only genes where one topology was significantly better than the alternative ones by the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999). Of the ML trees for the 433 genes considered, 172 passed the SH test: 140 (81%) are B-trees, 32 (19%) are A_1-trees, and none are A_2-trees. An increase in the rate of accumulation of A_1-trees at high *R* is again apparent (fig. 3*C*).

In summary, using gamma-corrected instead of Poisson-corrected amino acid distances resulted in an increase in the proportion of B-type topologies from 46% to 74%. Using ML with estimation of robustness by the SH test was similarly effective, increasing the proportion of B-trees to 81%. We suspect that the trees that remain A type in these analyses are caused by the LBA artifact rather than any biological phenomenon because most of them have the particular topology (A_1) that is expected under LBA.

Gene Order Information Confirms That the *S. kluyveri*-*S. cerevisiae* Divergence Pre-dates the WGD

The results described above suggest that the A-type topologies that Langkjaer et al. (2003) reported for half of the 38 gene pairs they studied in *S. kluyveri* are artifacts caused by LBA. We compared the local gene order between species around the 38 gene pairs and found that in all cases *S. kluyveri* has a gene order similar to that in the pre-WGD species *Kluyveromyces waltii, Ashbya gossypii*, and *K. lactis*, which diverged from the lineage leading to *S. cerevisiae* before the WGD occurred in the latter lineage (Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004). Four examples of genes studied by Langkjaer et al. (2003) are shown on the right-hand side of figure 5. The 2:1 synteny relationship seen between *S. cerevisiae* and *S. kluyveri* at each of these loci, as with the other 34, strongly suggests that the WGD postdated the speciation between the *S. cerevisiae* and *S. kluyveri* lineages, again indicating that the A-type trees are in error.

Furthermore, comparisons throughout the *S. kluyveri* genome using the Yeast Gene Order Browser (a tool we

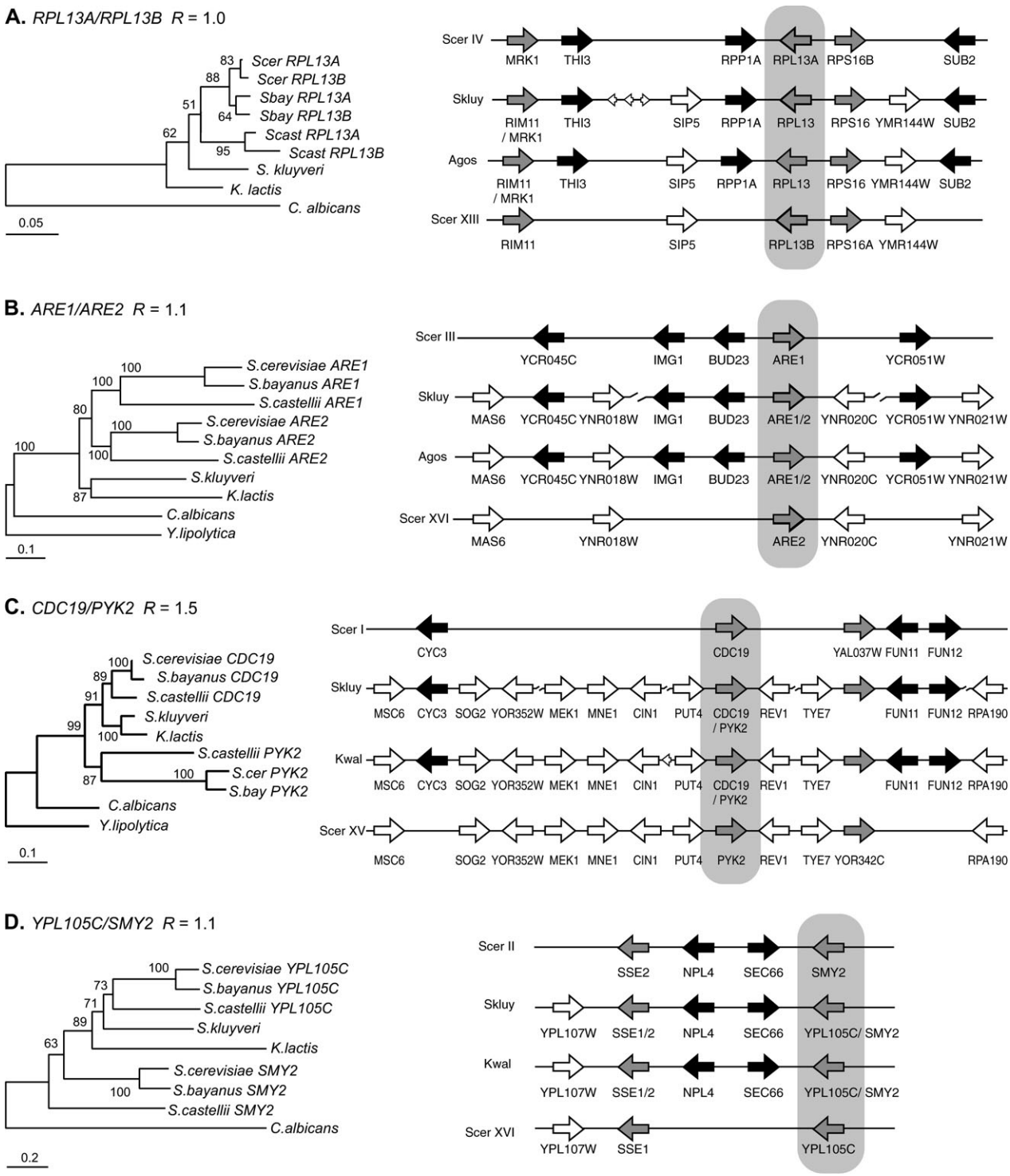

FIG. 5.—Gene order information and conflicting phylogenetic trees. Examples of four loci with varying values of rate asymmetry (R) are shown. For each, the phylogenetic tree (NJ method with gamma correction) is drawn beside an illustration (not to scale) of the local gene order relationships among *Saccharomyces kluyveri* contigs, pairs of chromosomal regions in *Saccharomyces cerevisiae*, and a representative pre-WGD species (*A. gossypii* or *K. waltii*). The loci used in the phylogenetic trees are highlighted. (*A*) *RPL13*, an example of tree topology B (the *S. kluyveri* sequence branches off before the two *S. cerevisiae* sequences). (*B*) *ARE1*/*ARE2*, another example of tree topology B. (*C*) *CDC19*/*PYK2*, an example of tree topology A_1 (the *S. kluyveri* sequence clusters with the slower evolving gene copy in *S. cerevisiae*). (*D*) *YPL105C*/*SMY2*, an example of tree topology A_2. Gaps in the horizontal lines for *S. kluyveri* show sequence gaps between contigs. Small arrows show species-specific genes without homologs in other species. The *S. kluyveri* gene orders are based on GenBank accession numbers AACE01000020.1; AACE01000332.1–AACE01000733.1–AACE01001815.1; AACE01000137.1–AACE01000584.1–AACE01000318.1–AACE01000370.1–AACE01000059.1; and AACE01000272.1.

Table 1
Extent of Conservation of Immediately Neighboring Gene Pairs Between *Saccharomyces kluyveri* and Other Yeast Genomes

Genome	Number of Neighboring Pairs Conserved	Percentage
Saccharomyces castellii	862	43.5%
Candida glabrata	807	40.7%
Saccharomyces cerevisiae	939	47.3%
Ashbya gossypii	1,396	70.4%
Kluyveromyces lactis	1,382	69.7%
Kluyveromyces waltii	1,438	72.5%

NOTE.—The 1,984 pairs of adjacent genes in *S. kluyveri* annotated by Cliften et al. (2003) were compared to the other genomes.

developed for comparing gene orders among yeast species; http://wolfe.gen.tcd.ie/ygob; Byrne and Wolfe 2005) show that its gene order is more similar to that in pre-WGD species (*K. waltii*, *K. lactis*, and *A. gossypii*) than to that in post-WGD species (*S. cerevisiae*, *S. castellii*, and *Candida glabrata*) and that there is a consistent 1:2 synteny relationship between single *S. kluyveri* genomic regions and pairs of genomic regions in each of the post-WGD species. The available genome sequence data from *S. kluyveri* (4x coverage; Cliften et al. 2003) include 1,984 pairs of genes that are immediate neighbors in *S. kluyveri*. About 70% of these pairs are also immediate neighbors in pre-WGD species, whereas in the post-WGD species the fraction is only about 45% (table 1).

We constructed phylogenetic trees for the same four gene pairs from the set of Langkjaer et al. (2003), which were chosen to reflect a range of *R* values (fig. 5). We included additional sequences from *S. bayanus* and *S. castellii* (both of which are post-WGD species; Langkjaer et al. 2003), as well as *K. lactis*, *C. albicans*, and in some cases, *Y. lipolytica*. We used the NJ method with gamma correction of distances and examined the relationship between tree topologies, gene order, and *R* values. For genes with relatively low *R* values (*RPL13A/RPL13B* and *ARE1/ARE2*; fig. 5*A* and *B*), B-type trees are obtained which is consistent with the local gene order information and the hypothesis that *S. kluyveri* split off before genome duplication (The *RPL13* tree has also been affected by gene conversion within each of the post-WGD species.) On the other hand, the *CDC19/PYK2* gene pair has $R = 1.5$ and gives an A_1 topology (fig. 5*C*). This contradicts the gene order information because the *S. kluyveri* genomic region shown in figure 5*C* contains homologs of all the genes in the corresponding regions of *S. cerevisiae* chromosomes I and XV, there is no other similar region in the *S. kluyveri* genome, and the *S. kluyveri* gene order in the region is almost identical to that in *K. waltii*. This leads us to conclude that the phylogenetic tree in figure 5*C* is wrong, despite the use of the gamma correction and the high bootstrap score (91%) supporting clustering of the *S. kluyveri* gene with *CDC19* to the exclusion of *PYK2*. The *YPL105C/SMY2* gene pair (fig. 5*D*) shows a rare example of an A_2 topology, where the *S. kluyveri* sequence groups with the faster copy in *S. cerevisiae* (*YPL105C*). In this tree, the *S. castellii SMY2* sequence appears to be misplaced due to LBA to the outgroup. Examination of *S. castellii* gene order in this region and throughout its genome shows unambiguously that *S. castellii* and *S. cerevisiae* speciated after the WGD. If *Y. lipolytica* is added as a second outgroup to the tree in figure 5*D* (data not shown), the three *SMY2* sequences form a monophyletic group and the branch length asymmetry between *S. cerevisiae YPL105C* and *SMY2* disappears, but the single *S. kluyveri* and *K. lactis* genes continue to group with *YPL105C*.

Adaptive Evolution After Gene Duplication Is Associated with LBA Effects on Phylogenies

One possible cause of unequal evolutionary rates in different gene copies is positive selection. We searched for evidence of positive selection on ohnolog sequences using CODEML (see *Materials and Methods*). To increase the power of the test for this analysis, we included sequences where possible from *S. bayanus* and *S. castellii* as well as *S. cerevisiae* and *S. kluyveri*. We did not use sequences from the outgroup *C. albicans* due to its high divergence from *Saccharomyces* species. In this way, we also avoided forcing the hypothesis about whether the genome duplication pre-dates or postdates the split between *S. kluyveri* and the other species.

Comparison of discrete models (M0 to M3) by the LRT showed significantly better log-likelihood values for M3 than for M0 in all the genes examined, supporting heterogeneous distribution of ω values among amino acid sites. Among the 433 genes studied, we identified 34 genes showing a proportion of amino acid sites under positive selection (data not shown). These genes were more or less equally distributed among those genes with *R* values <1.14 (4.6% of the 300 gene pairs in this set) and those showing *R* values ≥ 1.14 (3.1% of the 133 gene pairs in this set). Although positive selection was detected in these genes, indicating functional divergence and fast evolution over the complete phylogenetic tree, this result did not help to determine whether any sequence in particular is responsible for this positive selection.

The application of the free-ratio model identified 83 of the 433 gene pairs as having fixed amino acid substitutions by positive selection in one or both of the *S. cerevisiae* branches of the tree (19.2% of the duplicated genes; Table 1 of Supplementary Material online). This proportion is far beyond what we would expect by chance and shows the high level of divergence between ohnologs. If we assume that adaptive evolution in these genes is just the product of chance, then the maximum number we would expect is 5% of the tests carried out. When we divided the positively selected genes into two categories according to whether their *R* values were above or below 1.14 (the threshold value above which we have indications of LBA in fig. 3*A*), we found that the proportion of positively selected genes in the high-*R* category is nearly twice as high as in the low-*R* category (1.81 times higher; table 2).

These results suggest that some of the acceleration in substitution rates that leads to LBA effects on the phylogeny may have been associated with adaptive amino acid changes. To test this hypothesis, we further classified the positively selected genes into three categories, according to whether the free-ratio model identified positive selection

Table 2
Distribution of Positively Selected Genes (identified by the free-ratio model) Among Different Categories of Evolutionary Rate Asymmetry

		Positively Selected Branch			
Rate Asymmetry Category	Number of Gene Pairs	F	S	F and S	Total
$R < 1.14$	300	14 (4.7%)	14 (4.7%)	18 (6.0%)	46 (15.3%)
$R \geq 1.14$	133	19 (14.3%)	6 (4.5%)	12 (9.0%)	37 (27.8%)
Ratio of categories		3.94	0.96	1.50	1.81

NOTE.—Gene pairs are divided into two categories of rate asymmetry, depending on the value of the ratio (R) between the amino acid distances of the two ohnologs to the outgroup *Candida albicans*. Within each category, we further distinguish among pairs showing positive selection only in the faster evolving *Saccharomyces cerevisiae* ohnolog (F), only in the slower evolving ohnolog (S), or in both ohnologs (F and S). The ratio between the percentages of genes in the rate categories is shown at the bottom.

only in the faster *S. cerevisiae* ohnolog (F), only in the slower ohnolog (S), or in both (F and S). As shown in table 2, the distribution of the number of positively selected genes is highly heterogeneous among the three branch categories and between the two R-value groups. For the category of genes with reasonably symmetrical rates of evolution ($R < 1.14$), if only one branch shows positive selection it is equally likely to be either the faster or slower branch (4.7% each). In contrast, in the asymmetric category (R values ≥ 1.14), positive selection is three times more frequent in the faster copy than in the slower copy. Moreover, the total incidence of positive selection is higher in the LBA-prone asymmetric category than in the symmetric category (27.8% vs. 15.3%), although it should be noted that the absolute levels of sequence divergence are also higher in the genes with asymmetric rates, which may facilitate the detection of positive selection (The mean amino acid substitution estimate for the low-R and high-R groups are 0.52 ± 0.35 and 0.81 ± 0.36, respectively. The difference between these means is not significant; $P = 0.213$.) Hence, a correlation seems to exist between LBA effects and the fixation of amino acid substitutions by positive selection in the faster evolving gene copies.

Discussion

Despite the many models developed to infer accurate phylogenetic relationships among organisms, artifactual or incompletely resolved trees still pose a major problem in research (Delsuc, Brinkmann, and Philippe 2005). For example, LBA has been detected in previous studies that attempted to resolve the controversial phylogenetic relationships among rodents (Reyes, Pesole, and Saccone 2000), among angiosperms (Stefanovic, Rice, and Palmer 2004), and the placement of Microsporidia within eukaryotes (Thomarat, Vivarès, and Gouy 2004). Moreover, it has been suggested that the basal topology of the eukaryotic rRNA tree could be affected by LBA effects among faster evolving eukaryotes (Stiller and Hall 1999). A groundbreaking study showed that the sister grouping of eukarya and archaea observed with the ancient duplicated gene pair *SRP54*/*SRα* is mainly due to fast-evolving positions and that this clustering is unsustainable using slower evolving amino acid sites (Brinkmann and Philippe 1999).

LBA is the consequence of fast evolution in some lineages of the tree compared to the average tree length. Several studies have shown that duplicated genes often have asymmetrical rates of evolution (Van de Peer et al. 2001; Conant and Wagner 2003; Kellis, Birren, and Lander 2004), and in our data set more than half of the ohnolog pairs show more than a 5% difference in evolutionary rates. This inequality of branch lengths, together with the general acceleration of evolutionary rates that often accompanies gene duplications due to the relaxation of evolutionary constraints (Lynch and Conery 2000), makes phylogenetic trees that include mixtures of orthologous and paralogous sequences particularly prone to LBA artifacts.

LBA provides a likely explanation for the conflicting phylogenetic trees obtained in the analysis of Langkjaer et al. (2003) of the WGD event during the evolution of *Saccharomyces* species. Gene order comparisons show that the WGD undoubtedly postdates the divergence of *S. cerevisiae* from *S. kluyveri*, which implies that the trees with topology type A are in error. The observation that there are far more trees of type A_1 than A_2 renders remote the possibility that there is a real biological explanation for the A-trees. We found that the proportion of artifactual trees could be reduced, but not completely eliminated, by using gamma-corrected distances (instead of Poisson correction) with the NJ method or by using ML with robustness being estimated by the SH test.

What causes duplicated genes to have asymmetric rates of evolution? Rate differences can be the result of altered levels of negative selection on gene copies or of different numbers of positively selected changes in the two copies. Our analysis of selective constraints revealed that a surprisingly high proportion (19%) of duplicated genes have fixed amino acid substitutions by positive selection and that in genes showing rate asymmetry, positively selected changes were more frequent on the faster branch than on the slower one. However, it is important to note that positive selection was only detected in a minority of genes, even among those whose rates of evolution were asymmetric ($R \geq 1.14$). Whether this is because current methods for detecting positive selection are not sufficiently sensitive or because the asymmetry in the other genes is simply due to different levels of constraints on the two copies remains to be determined.

Supplementary Material

Table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This study was supported by Science Foundation Ireland. We thank Gavin Conant for comments on the manuscript.

Literature Cited

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. **16**:817–825.

Byrne, K. P., and K. H. Wolfe. 2005. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. **15**:1456–1461.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science **301**:71–76.

Conant, G. C., and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. **13**:2052–2058.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. **6**:361–375.

Dietrich, F. S., S. Voegeli, S. Brachat et al. (14 co-authors). 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science **304**:304–307.

Dujon, B., D. Sherman, G. Fischer et al. (67 co-authors). 2004. Genome evolution in yeasts. Nature **430**:35–44.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

Friedman, R., and A. L. Hughes. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. **11**:373–381.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428**:617–624.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**:241–254.

Keogh, R. S., C. Seoighe, and K. H. Wolfe. 1998. Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. Yeast **14**:443–457.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5**:150–163.

Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421**:848–852.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

Lynch, M., and V. Katju. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet. **20**:544–549.

Nicholas, K. B., H. B. Nicholas Jr., and D. W. Deerfield II. 1997. GeneDoc: Analysis and visualization of genetic variation. EMBNET News **4**:14.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**:205–217.

Reyes, A., G. Pesole, and C. Saccone. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene **259**:177–187.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

Souciet, J., M. Aigle, F. Artiguenave et al. (24 co-authors). 2000. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett. **487**:3–12.

Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol. Biol. **4**:35.

Stiller, J. W., and B. D. Hall. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol. Biol. Evol. **16**:1270–1279.

Thomarat, F., C. P. Vivarès, and M. Gouy. 2004. Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of Microsporidia and reveals a high frequency of fast-evolving genes. J. Mol. Evol. **59**:780–791.

Van de Peer, Y., J. S. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol. Evol. **53**:436–446.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**: 691–699.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

Wong, S., G. Butler, and K. H. Wolfe. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc. Natl. Acad. Sci. USA **99**:9272–9277.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Herve Philippe, Associate Editor

Accepted September 19, 2005