The Plant Cell, Vol. 15, 2192–2202, September 2003, www.plantcell.org © 2003 American Society of Plant Biologists

Evidence That Rice and Other Cereals Are Ancient Aneuploids

Klaas Vandepoele, Cedric Simillion, and Yves Van de Peer[1]

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Detailed analyses of the genomes of several model organisms revealed that large-scale gene or even entire-genome duplications have played prominent roles in the evolutionary history of many eukaryotes. Recently, strong evidence has been presented that the genomic structure of the dicotyledonous model plant species Arabidopsis is the result of multiple rounds of entire-genome duplications. Here, we analyze the genome of the monocotyledonous model plant species rice, for which a draft of the genomic sequence was published recently. We show that a substantial fraction of all rice genes (~15%) are found in duplicated segments. Dating of these block duplications, their nonuniform distribution over the different rice chromosomes, and comparison with the duplication history of Arabidopsis suggest that rice is not an ancient polyploid, as suggested previously, but an ancient aneuploid that has experienced the duplication of one—or a large part of one—chromosome in its evolutionary past, ~70 million years ago. This date predates the divergence of most of the cereals, and relative dating by phylogenetic analysis shows that this duplication event is shared by most if not all of them.

INTRODUCTION

Large-scale duplication events have been considered important for the evolution of many organisms because they provide a way to considerably increase the genetic material on which evolution can work (Stephens, 1951; Ohno, 1970; Sidow, 1996; Holland, 2003). Because duplicated genes are redundant, one of the copies is, at least theoretically, freed from functional constraint and can evolve a new function (Van de Peer et al., 2001; Prince and Pickett, 2002). The search for traces of (ancient) large-scale gene duplications has received much attention of late, and hypotheses about the number and age of polyploidy events in eukaryotes are actively discussed (Wolfe, 2001; Durand, 2003). This is partly attributable to the fact that the detection of homologous (or paralogous) regions in genomes is not self-evident (Gaut, 2001; Vandepoele et al., 2002a).

Identifying duplicated regions at the gene level is based on a within-genome comparison that aims at delineating regions of conserved gene content and order (such regions are said to be colinear) in different parts of the genome. In general, one tries to identify a number of homologous gene pairs (usually referred to as anchor points) in relatively close proximity to each other between two different segments in the genome, either on the same chromosome or on different chromosomes. When such a candidate colinear region is detected, usually some sort of permutation test is performed in which a high number of randomized data sets are sampled to calculate the probability that the observed colinearity could have been generated by chance (Gaut, 2001). When it can be shown that the similarity between two genomic segments is unlikely to be the result of chance and therefore is statistically significant, the conclusion is reached that the duplicated genes are the result of a single segmental (block) duplication. The statistics that determine colinearity depend on two factors: the number of anchor points and the distance over which these are found, which usually depends on the number of "single" genes that interrupt colinearity. The high level of gene loss—together with phenomena such as translocations and chromosomal rearrangements—often renders it very difficult to find statistically significant homologous regions in the genome, particularly when the duplication events are ancient.

In plants, the systematic analysis of the Arabidopsis genome sequence has shown that this genome contains a large number of duplicated regions and that up to ~90% of the Arabidopsis genes occur in genomic segments that have been duplicated at one time or another (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003). By applying novel techniques to detect heavily degenerated block duplications in Arabidopsis, we showed recently that the genome of this dicotyledonous model plant has been reshaped by not one but three large-scale gene, and probably even entire-genome, duplication events (Simillion et al., 2002).

Apart from Arabidopsis (Arabidopsis Genome Initiative, 2000), rice is currently the only plant species for which draft sequences of the nuclear genome have been published (Goff et al., 2002; Yu et al., 2002). In addition, more complete versions of chromosomes 1, 4, and 10 have been published by the International Rice Genome Sequencing Project (Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003). Rice is one of the most important cereal crops in the world and also is an excellent plant model system, as a result of its small genome size (430 Mb) and the high level of synteny with other cereals. Comparative mapping analyses of genomes of closely related grass species revealed a remarkably good conservation of markers within large chromosomal segments

[1] To whom correspondence should be addressed. E-mail yves.vandepeer@psb.ugent.be; fax 32-9-331-3809.

Article, publication date, and citation information can be found at www.plantcell.org/cgi/doi/10.1105/tpc.014019.

(for review, see Keller and Feuillet, 2000). Soon after the detection of colinearity based on genetic maps, detailed sequence analyses confirmed the existence of microcolinearity (i.e., conserved gene content and order at the gene level) between orthologous loci from closely related grass genomes, which varied extensively in size (Chen et al., 1997; Tikhonov et al., 1999; Paterson et al., 2000; Tarchini et al., 2000). Consequently, grasses can be studied as a single genetic system, allowing the transfer of biological information from a well-studied model grass genome, such as that of rice, to related plant species (Gale and Devos, 1998). Although several studies that crossed the monocot-dicot boundary also identified numerous microcolinear segments between Arabidopsis and rice (Paterson et al., 1996; Liu et al., 2001; Mayer et al., 2001; Salse et al., 2002; Vandepoele et al., 2002a), the small size of these regions seems to seriously limit their value for comparative analysis of dicotyledonous and grass genomes.

In strong contrast to Arabidopsis, in which the initial sequencing of the genome sequence already revealed numerous duplicated segments (Terryn et al., 1999; Blanc et al., 2000; Paterson et al., 2000), very few studies have reported possible evidence for large-scale gene or complete-genome duplications in rice (Kishimoto et al., 1994; Nagamura et al., 1995), although a polyploid origin for rice has been suggested on several occasions (Goff et al., 2002; Levy and Feldman, 2002). Here, we report the detailed analysis of the rice genome, focusing on large-scale gene duplications. We show that large-scale gene duplication events did occur in the evolutionary past of rice but that the duplication history and magnitude are considerably different from those of its dicotyledonous counterpart Arabidopsis.

RESULTS

Detection of Nonhidden Block Duplications in the Rice Genome

Because one preferentially wants to use large genomic regions for the detection of duplicated segments in a genome, we built a data set of assembled rice genomic BAC sequences that were obtained from the International Rice Genome Sequencing Project (Sasaki and Burr, 2000). Where traditional sequence assembly programs are designed mainly to assemble large sets of individual sequence reads into larger contigs, the construction of large genomic scaffolds starting from already assembled genomic BAC clones is far from trivial. Because no publicly available assembly program was found that could handle and assemble genomic BAC clones, which range in size from 10 to 250 kb, we applied a newly developed assembly routine. The automatic sequence-to-genome assembly routine (ASGAR) is a conservative method that physically merges BAC clones with significant overlap (see Methods).

After applying two rounds of assembly using ASGAR to the initial data set, the number of genomic sequences was reduced from 2897 BACs to 1025 genomic scaffolds (498 supercontigs and 527 singleton BACs). The total size of these scaffolds is 330.47 Mb, with an average size of 322 kb per scaffold. Gene annotation was retrieved from RiceGAAS (Sakata et al., 2002) and yielded 39,096 genes after filtering. This filtering step removed potential falsely predicted genes, based on the absence of homology for a predicted gene with a rice EST or any other protein present in the public protein databases (Vandepoele et al., 2002a). In addition, all predicted genes with similarity to transposable elements were removed. On average, 32 genes were present per genomic scaffold, which corresponds with an average gene density of one gene per 10 kb. An overview of the gene density and the length distribution of the scaffolds is shown in Figure 1.

By applying the ADHoRe (automatic detection of homologous regions) algorithm to an assembly covering ~70% of the annotated rice genome sequence, 193 statistically significant duplicated segments were identified ($P < 0.001$), of which 150 contain three or four paralogous gene pairs (so-called anchor points) and 43 contain five or more gene duplicates. The complete set of block duplications, omitting tandem duplications, contains 862 anchor points and includes nearly 15% of all rice proteins in our annotated nonredundant data set. Approximately two-thirds of the duplicated blocks (i.e., 129 of all detected duplicated blocks) are located at the beginning or the end of a genomic scaffold (i.e., the first or last five genes), which can be explained by the incomplete assembly of our data set. Regarding the 43 large block duplications (more than five anchor points), 34% of the total number of genes in these segments are retained duplicates. The largest block duplication in our assembled scaffold data set is formed by a 0.96-Mb segment with 107 genes on chromosome 1 and a 0.69-Mb segment with 62 genes on chromosome 5, governing 33 retained gene duplicates. Apart from the set of paralogous genes located in duplicate blocks, 1609 tandem duplications were detected involving 4308 individual genes. This number corresponds with 16.9% of all genes in our data set, which is very similar to what is found in Arabidopsis (Vision et al., 2000; Simillion et al., 2002). The largest tandem repeat was formed by 16 genes.

Hidden and Ghost Duplications

Apart from the large set of block duplications identifiable by direct comparisons of different genomic segments (so-called "nonhidden" duplications), an additional number of block duplications in the rice genome could be identified by indirect comparisons (so-called "hidden" and "ghost" duplications; see Methods) (Figure 2). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments; rather, they are observed only through comparison with a third segment. Consequently, hidden duplications are important to consider for determining the actual number of duplication events that have occurred over time, as we demonstrated previously for Arabidopsis (Simillion et al., 2002). Reconstruction of multiplicons (i.e., sets of homologous segments; Simillion et al., 2002) for rice through the identification of hidden duplications revealed only two cases in which a chromosomal segment was involved in more than one duplication event.

Considering all 157 colinear regions detected between rice and Arabidopsis, another five ghost duplications were identified. The largest rice ghost duplication was found between ge-

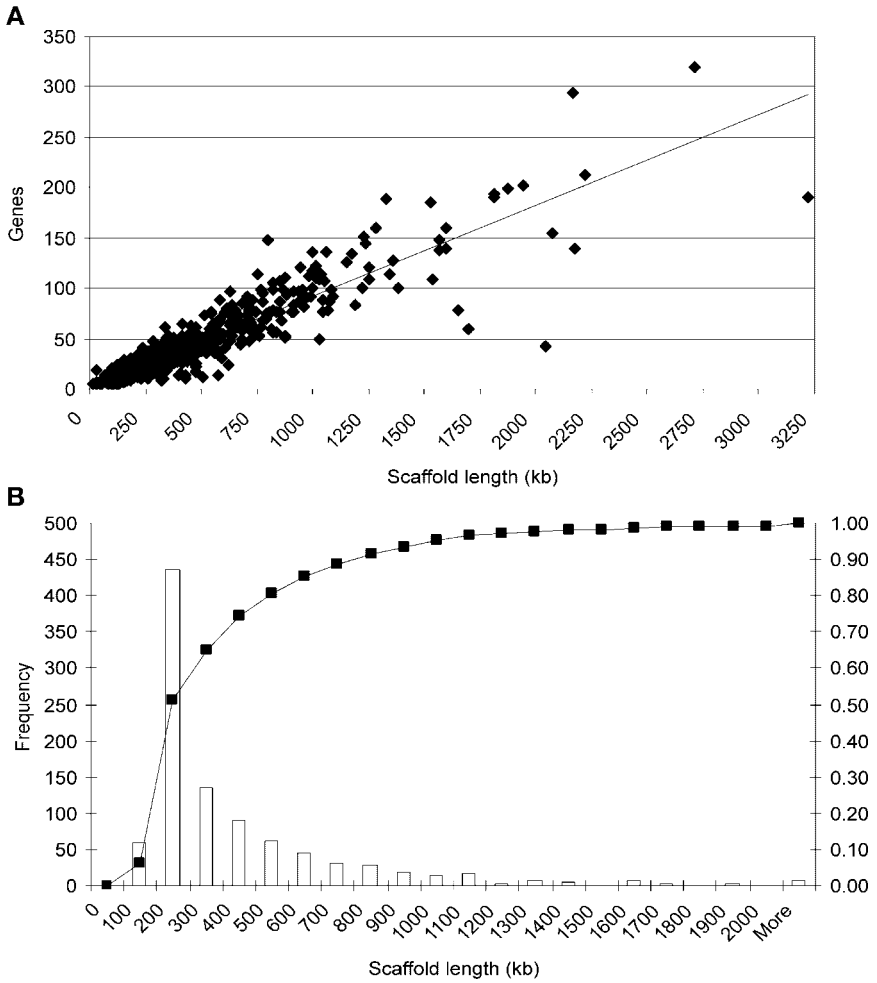

Figure 1. Overview of the Genomic Scaffolds Generated by ASGAR.

(A) Scatterplot showing the number of genes versus the scaffold length for all 966 genomic scaffolds that were used for the detection of duplicated blocks. The best-fit line, which shows a quite homogeneous gene density for the scaffolds ($R^2 = 0.85$), represents a gene density of 1 gene per 10 kb.
(B) Length distribution of all genomic scaffolds that were subjected to block detection. The line indicates the relative (cumulative) contribution of the scaffolds assigned per bin (i.e., length segment) in the histogram.

nomic segments of chromosome 4 (46 genes spanning 477 kb) and chromosome 10 (64 genes spanning 761 kb), both colinear with chromosome 2 of Arabidopsis. More detailed analysis of these duplicated segments showed that each genomic segment has lost a different set of genes and that only a subset of the initial number of gene duplicates is retained (data not shown). Therefore, the combination of a limited number of gene duplicates with different types of rearrangements subsequent to the original duplication event does not allow the detection of this degenerated paralogous region using only the rice genome.

Age Estimation of Duplicated Blocks

For reasons of statistical significance (see Methods), only the set of block duplications with five or more anchor points (377 anchor points in total) was used to date the duplication events. Briefly, for a duplicated block, all anchor points were subjected to a dating method based on the number of synonymous substitutions per silent site (K_s), and all values obtained were used subsequently to calculate the mean K_s for each block duplication after removing outliers (Simillion et al., 2002). Although large variation in K_s estimates among contemporaneously duplicated genes in Arabidopsis has been reported (Zhang et al., 2002), removal of outliers greatly reduces the variation of the final K_s estimate for a duplicated block. Nearly half of all anchor points (i.e., 47%) have K_s values of between 0.6 and 1.1 (Figure 3), corresponding with duplication dates of 46 and 85 million years ago, respectively. The median, a K_s value of 0.87, corresponds with 67 million years ago.

Because absolute dating of duplication events has been criticized and may rely heavily on obtained K_s values and the esti-

mated rate of synonymous substitutions for the organism of interest, which may not be very accurate (Li, 1997; Zeng et al., 1998; Blanc et al., 2003), we also applied relative dating by phylogenetic means (see Methods). In short, for a given pair of gene duplicates that is part of a duplicated block, homologous genes of related monocotyledonous plants were selected together with an appropriate outgroup sequence, and the evolutionary relationships between these different organisms were inferred based on the topology of the phylogenetic tree obtained. In total, 170 phylogenetic trees with bootstrap support were generated, representing a set of 99 block duplications (i.e., 1.7 trees per duplicated block on average). Fifty-four percent of these trees clearly supported the duplication event having occurred before the divergence of the cereals (Figure 4) (Kellogg, 2001).

Regarding the 18 large (more than five anchor points) block duplications with K_s values between 0.6 and 1.1, 74% of the topologies clearly supported duplication having occurred before the divergence of cereals. When more than one anchor point in the same block duplication could be used for tree construction (as was the case for 39 block duplications), 78% of the inferred trees within one duplicated block were congruent with one another. For all of the remaining tree topologies, no conclusions could be reached, for different reasons, such as the absence of real orthologs or sequences being too conserved. However, none of the trees was in clear conflict with a duplication event shared between rice and other cereals. Supplemental data on the block duplications detected, along with more detailed results from the dating analyses, are available at http://www.psb.ugent.be/bioinformatics/.

DISCUSSION

Block Duplications in the Rice Genome

The grass family has been the subject of many detailed comparisons of genome structure and gene order. Based on the presence of large colinear regions between different grass genomes, the creation of a grass consensus map clearly revealed the structural similarity between related grass genomes (Gale and Devos, 1998). Although large chromosomal rearrangement events can be determined with the current resolution of these maps, information regarding large-scale duplication events within rice is scarce (Kishimoto et al., 1994; Nagamura et al., 1995).

Based on a BAC assembly covering >70% of the genome sequence of rice, we applied the ADHoRe algorithm to detect block duplications at the gene level. Subsequent to the detec-

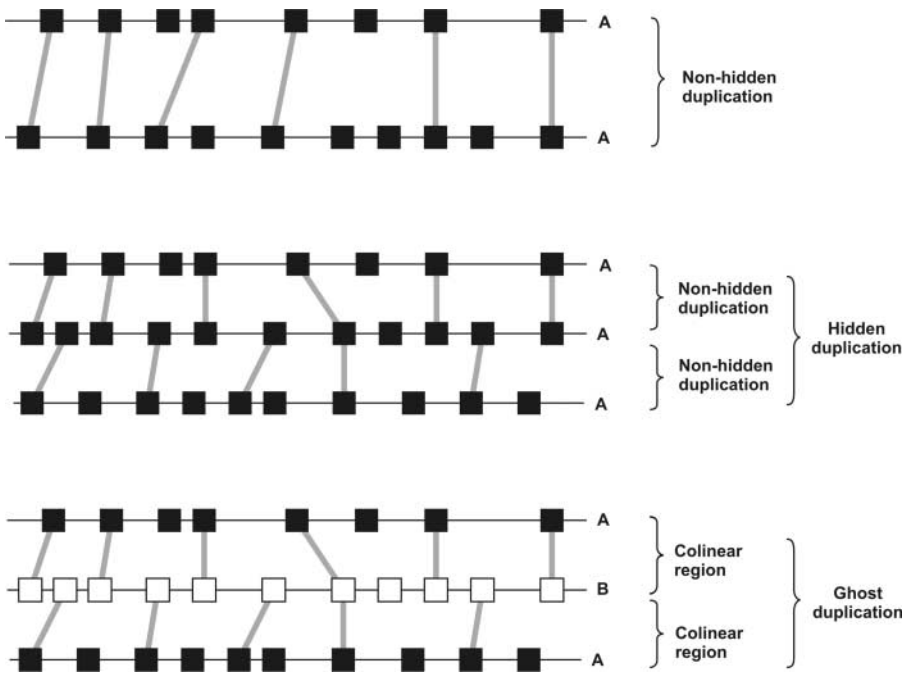

Figure 2. Scheme of Nonhidden, Hidden, and Ghost Duplications.

Boxes represent the genes on chromosomal segments of genomes A and B, whereas connecting lines indicate the anchor points (i.e., homologous or duplicated genes). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments; rather, they are observed only through comparison with a third segment from the same genome. Because nonhidden duplications are used to infer hidden duplications, no additional genomic segments are assigned to a duplication event, although the number of duplication events for a given segment increases. Ghost duplications are hidden block duplications that can be identified only through colinearity with the same segment in a different genome. In contrast to hidden duplications, the identification of ghost duplications increases the fraction of the genome involved in a duplication event.

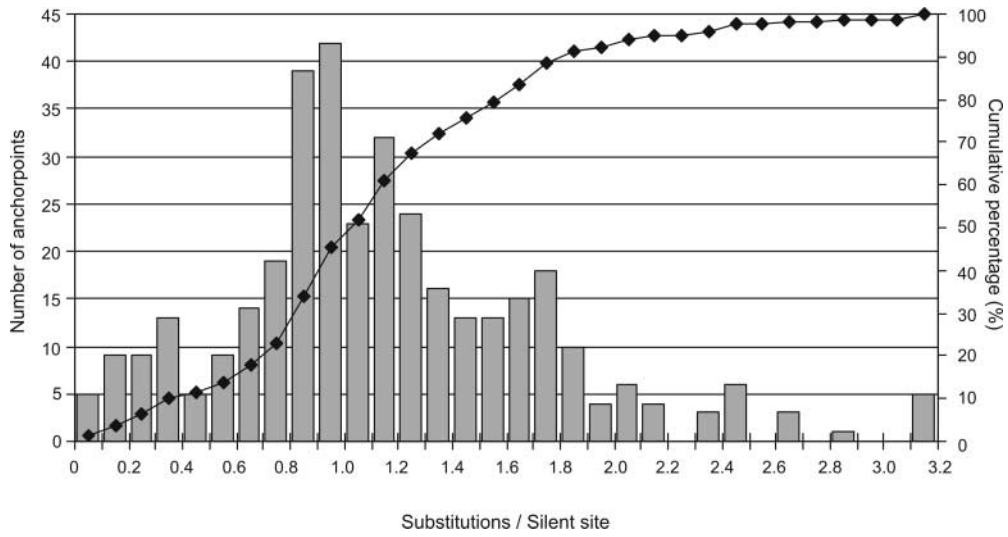

Figure 3. Absolute Dating of Block Duplication Events in the Rice Genome.

Age distribution of all gene duplicates that are part of large (more than five anchor points) duplicated segments in the rice genome. The line indicates the relative (cumulative) contribution of the anchor points assigned per bin (i.e., age segment) in the histogram.

tion of a large number of duplicated segments by direct comparison of all rice genomic scaffolds, a comparative approach using the genome sequence of Arabidopsis also yielded a set of ghost duplications, reflecting heavily degenerated duplicated segments. Regarding the 43 large (more than five anchor points) block duplications, 34% of the total number of genes in these segments are retained duplicates. This fraction of retained gene duplicates, when the estimated time of duplication is considered (see below), is very similar to what has been observed in Arabidopsis and yeast (28 and 25%, respectively)

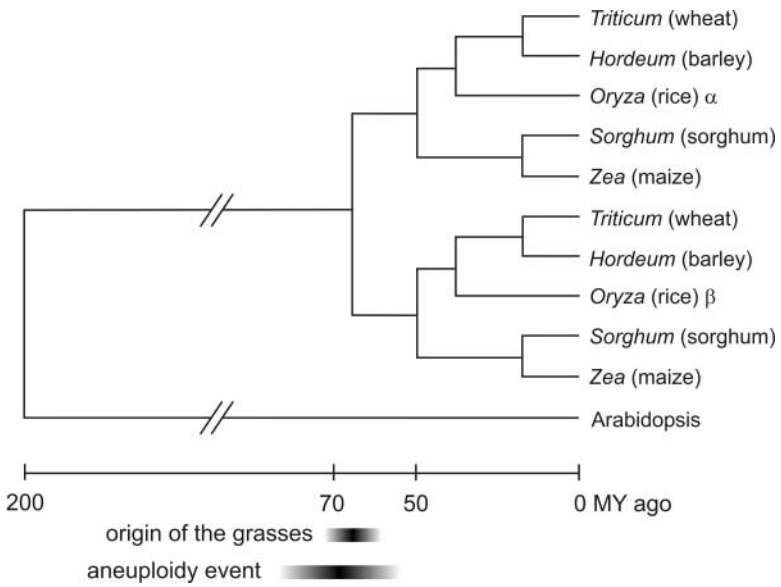

Figure 4. Dating of Duplication Events in the Rice Genome by Phylogenetic Means.

Expected tree topology and date of origin for genes of the cereals wheat, barley, rice, maize, and sorghum if these genes have duplicated before the divergence of rice and other cereals. The large majority of tree topologies obtained in this study, including those of two copies of rice (i.e., the retained duplicates found in large duplicated segments) and at least one copy of another cereal, are congruent with this tree topology, in which one rice gene branches off before the divergence of rice and other cereals. Such topologies suggest a duplication before the divergence of rice, barley, wheat, maize, and sorghum, estimated at ~50 million years ago (Kellogg, 2001), and may have occurred just before the origin of the grasses, as suggested by the K_s-based dating (see text for more details).

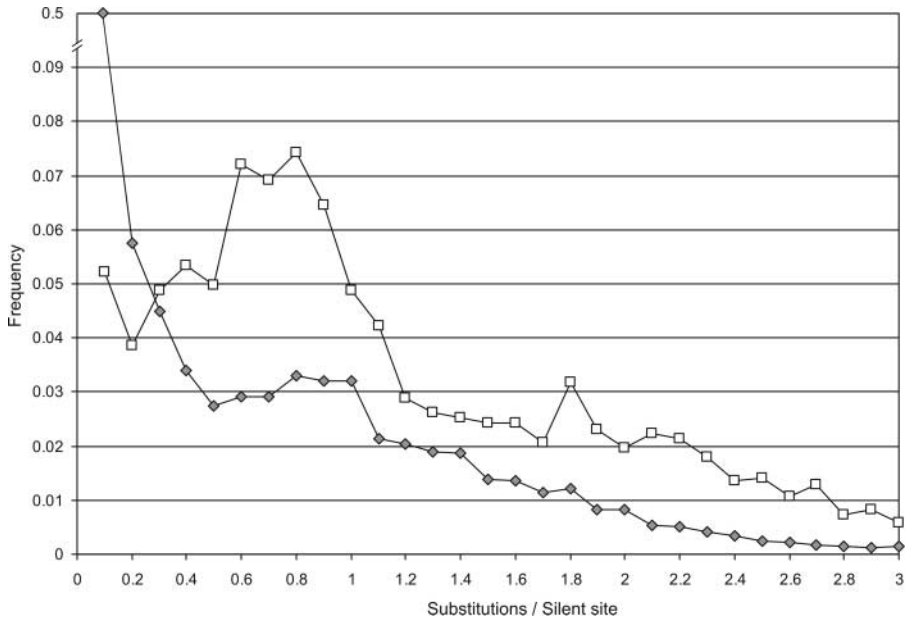

Figure 5. Frequency Distribution of Duplicated Genes in Arabidopsis and Rice as a Function of the Number of Silent Substitutions per Silent Site.

All frequencies were corrected for the total number of dated gene duplicates per genome, which were 4928 for Arabidopsis (white squares) and 7698 for rice (gray diamonds). The fact that the total number of duplicated genes is higher in the rice than in the Arabidopsis gene family is attributable to the facts that the rice genome contains more predicted genes and that in Arabidopsis more gene families with >10 members have been omitted from the analysis.

(Wolfe and Shields, 1997; Simillion et al., 2002), which seems to indicate similar rates of gene loss after duplication events.

When inferring the multiplication levels for all multiplicons (sets of homologous segments) present in the rice genome through nonhidden, hidden, and ghost duplications, ~1.3% of the genome resides in multiplicons with multiplication levels greater than two. This finding demonstrates that, given the quality of the current rice genomic data, a very small number of chromosomal regions seems to have been involved in multiple duplication events, in strong contrast to the findings in Arabidopsis, in which the majority of chromosomal regions have been involved in multiple duplication events (Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003).

Is Rice an Ancient Polyploid?

It has been suggested that many polyploidy and/or aneuploidy events in the evolutionary history of the grasses are required to explain the current distribution of chromosome numbers among grass taxa (for review, see Gaut, 2002). Although an apparent whole-genome duplication, ~40 to 50 million years ago, was reported based on the rate of amino acid substitution of all possible paralogous protein pairs in the rice genome (Goff et al., 2002), there is good evidence that protein distances are not very reliable for the large-scale dating of heterogeneous classes of proteins (Li, 1997; Wolfe, 2001; Raes et al., 2003). To answer the question of whether rice is an ancient polyploid, we compared the duplication history of Arabidopsis and rice by plotting the total number of gene pairs in both species against their genetic distance inferred from the nucleotide substitutions at silent sites (Figure 5).

When all duplicated gene pairs in Arabidopsis and rice were plotted as a function of K_s, the shape and height of both curves were quite different. In Arabidopsis, the number of duplicates with K_s values between 0.6 and 0.9 increased dramatically, which corresponds with a genome duplication ~40 to 75 million years ago, as reported previously (Lynch and Conery, 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003). Although overall, an exponential decay of the number of retained gene duplicates over time can be observed (Lynch and Conery, 2000), a small but significant increase also was observed for rice duplicates with K_s values between 0.6 and 1.1. However, because the increase in the number of duplicates, relative to the total number of duplicates, is much smaller in rice than in Arabidopsis (Figure 5), a complete genome duplication in rice seems highly unlikely.

In Arabidopsis, in which at least three rounds of large-scale gene duplication have been suggested (Vision et al., 2000; Simillion et al., 2002), 80% of the genome resides in duplicated blocks, 60% of which can be attributed to the most recent duplication event (data not shown). In the yeast *Saccharomyces cerevisiae*, also supposedly an ancient tetraploid, ~50% of the genome is found in duplicated segments (Wolfe and Shields, 1997). Therefore, if similar rates of gene loss are assumed during diploidization (the process whereby a tetraploid species be-

comes a diploid) between different eukaryotes, the fact that only ~15% of the rice genome is found in duplicated blocks also disagrees with the notion of whole-genome duplication.

Mapping the locations of all block duplications on the different chromosomes provides an alternative way to estimate the distribution and overall impact of the duplicated blocks (Table 1). The physical size of all duplicated segments between two chromosomes was determined by comparing the fraction in our data set with the estimated chromosome sizes described by Chen et al. (2002). If a complete-genome duplication or polyploidization event had occurred in the evolutionary past of rice, it would be expected that all duplications would be spread uniformly over all chromosomes (i.e., null hypothesis). On the contrary, if only one chromosome, or a larger segment of a chromosome, had duplicated, such a uniform distribution would not be expected. Clearly, the observed distribution differs significantly from the null hypothesis (Table 1), which strongly suggests that the observed duplication landscape is not the result of an entire-genome duplication. Instead, a major fraction of the detected duplications involve chromosome 2, suggesting that this chromosome, or at least parts thereof, might have been involved in an aneuploidy event, followed by a number of chromosomal rearrangements. Both the dating based on synonymous substitutions and the dating by phylogenetic means support the notion that this event occurred before the divergence of cereals.

Because of the incomplete and fragmented nature of the current data set and the conservative approach used, our results only partially confirm previously reported duplicated segments based on marker analysis (Kishimoto et al., 1994; Nagamura et al., 1995). Therefore, we expect that the total number of duplications will be slightly higher once more complete chromosomal sequences become available.

The presence of a small number of rice genomic regions that seem to have experienced multiple duplication events suggests that additional older block duplications occurred in the evolutionary past of rice. Indeed, analysis of mixed multiplicons (Figure 6), which represent all homologous relationships between genomic segments from Arabidopsis and rice, shows that additional information regarding genome evolution and duplication events within these plant model systems can be inferred. Careful investigation of colinear segments between Arabidopsis and rice shows that a number of very degenerated block duplications still can be recovered for both organisms, allowing a more realistic estimation of the number of duplication events that a homologous genomic segment in both species has undergone. Because for a number of mixed multiplicons the colinearity between homologous segments of rice and Arabidopsis still can be determined in a statistically significant way, which is not the case for paralogous segments within the genomes of rice and Arabidopsis (Figure 6), this pattern of conserved gene content and order could represent the remnants of a duplication event

Table 1. Coverage of Block Duplications between All 12 Rice Chromosomes

		0.003																					
1	0.003																						
	N.D.			0.011																			
2		N.D.	0.011																				
				0.003	N.D.																		
3			0.003			N.D.																	
		6E-04		**0.083**		0.003		**0.021**															
4	0.001		**0.096**		0.002		**0.021**																
		0.108		0.002		0.005		0.001	N.D.														
5	**0.157**		0.002		0.003		5E-04			N.D.													
		0.002		**0.081**	N.D.			0.004		0.006		**0.019**											
6	0.001		**0.115**			N.D.	0.007		0.003		**0.019**												
		0.004		**0.018**		**0.032**		0.008		0.007		0.009		0.011									
7	0.008		**0.016**		**0.021**		0.013		0.001		0.017		0.011										
		0.006		0.006	N.D.			0.035		3E-04		**0.018**		0.012		0.006							
8	0.011		0.004			N.D.	0.011		2E-04		**0.017**		0.018		0.006								
	N.D.			7E-04		0.01		0.002	N.D.			0.004		0.003		0.007	N.D.						
9		N.D.	8E-04		0.023		0.007			N.D.	0.012		0.006		0.014			N.D.					
		0.003		**0.03**		0.009		**0.056**	N.D.			0.003		0.005		0.001	N.D.		N.D.				
10	0.012		**0.038**		0.007		**0.063**			N.D.	0.002		0.01		0.001			N.D.		N.D.				
	N.D.			5E-04	N.D.			0.002	N.D.			0.003		0.002	N.D.			0.007		0.002	N.D.			
11		N.D.	0.001			N.D.	0.006			N.D.	0.018		0.004			N.D.	0.005		0.005			N.D.		
		0.005		0.002				0.005	N.D.			0.011		7E-04		0.01		0.009	N.D.			0.008		0.009
12	0.006		0.004				0.012			N.D.	0.015		0.001		0.015		0.005			N.D.	0.002		0.009	
		1		2		3		4		5		6		7		8		9		10		11		12

Each pair of values in a particular cell describes the fraction covered by duplicated segments between both chromosomes (e.g., all duplicated segments between chromosome 1 and 5 cover 10.8% of chromosome 1 and 15.7% of chromosome 5). All values that differ significantly from the null hypothesis (i.e., that duplicated segments are distributed uniformly over all chromosomes) are indicated in boldface ($P < 0.001$). N.D. (not determined) refers to the fact that no block duplications could be detected between the chromosomes.

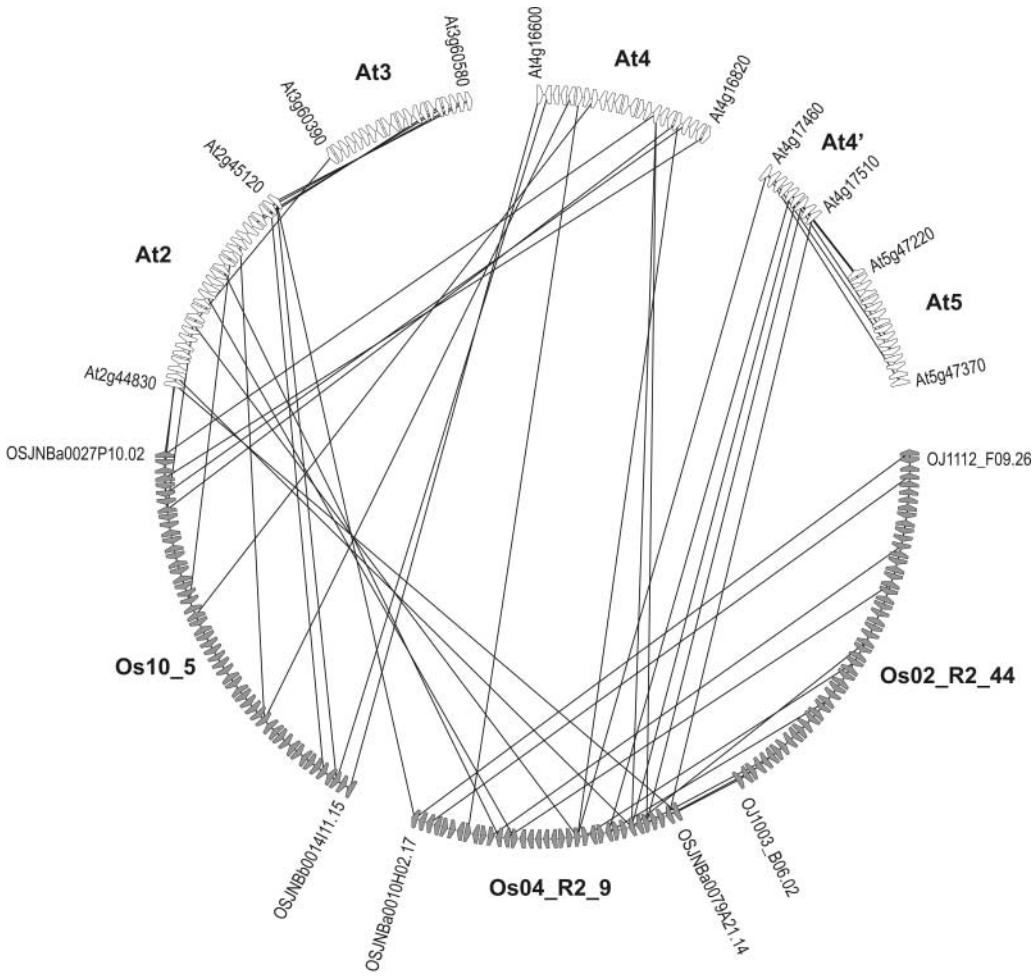

Figure 6. Set of Homologous Chromosomal Segments (Multiplicon) of Arabidopsis and Rice.

Arrows represent the genes on the chromosomal segments, and connecting lines indicate the anchor points (i.e., homologous or duplicated genes) that are part of a significant colinear relation determined by the ADHoRe algorithm. For each genomic segment, the names of the two genes delineating the segment are shown. Chromosomal segments of rice and Arabidopsis are shown in gray and white, respectively. By considering the colinearity between Arabidopsis and rice, a set of seemingly unrelated Arabidopsis segments can be joined into a multiplicon with a multiplication level of five, confirming the three duplication events in Arabidopsis described previously (Simillion et al., 2002). This colinearity also reveals that all three rice segments are linked with each other by two duplication events. Scaffold Os04_R2_9 includes BACs with accession numbers AL663006, AL662998, AL606459, AL607006, AL606728, AL606695, AL606587, AL606647, AL606633, AL663000, AL731613, AL606682, AL606687, AL606694, AL606628, AL607001, AL663003, and AL662954; scaffold Os10_5 includes BACs with accession numbers AC084763, AC079890, AC079874, AC069300, AC037426, and AC026758; and scaffold Os02_R2_44 includes BACs with accession numbers AP005108, AP004037, AP004883, AP005072, AP005289, AP005006, and AP004676.

predating the monocot-dicot divergence, as was suggested recently (Bowers et al., 2003; Raes et al., 2003).

METHODS

Rice Data Set

A total of 2897 rice (*Oryza sativa*) BAC sequences of the International Rice Genome Sequencing Project were retrieved from GenBank (September 2002). The total size of these genomic sequences amounts to 406.66 Mb, with an average size of 140 kb per BAC. Because both the sequence quality and the average length of genomic scaffolds from whole-genome shotgun approaches (fourfold to sixfold coverage and ~6 to 10 kb) (Goff et al., 2002; Yu et al., 2002) are inferior compared with BAC data, the former are less suited for the detection of block duplications. In addition, gene annotations for both whole-genome shotgun approaches are not publicly available.

Automatic Sequence-to-Genome Assembly Routine

A newly developed assembly routine for BAC sequences called the automatic sequence-to-genome assembly routine (ASGAR) was applied to

merge significantly overlapping BAC sequences into larger contigs (so-called supercontigs). For each genomic BAC sequence, ASGAR determines the BAC with the most significant overlap and creates a linked BAC pair. In the next step, either a new BAC pair is formed with no relation to the existing pair or a BAC pair that can be linked to an existing pair is formed. Afterward, all overlapping BAC sequences that are linked and thus represent a tiling path are merged into supercontigs using the EMBOSS program megamerger (Rice et al., 2000). A significant overlap between two BAC sequences is defined by an overlap of at least 1500 nucleotides with minimum 99% sequence identity. In addition, the overlap must be located at the end of one of the BAC sequences (i.e., the first or last 20% of the sequence). Sequence similarity searches were performed with BLASTN (Basic Local Alignment Search Tool; Altschul et al., 1997). Because both the input and output of ASGAR are a set of genomic sequences, multiple rounds of assembly can be performed until no more BAC sequences can be merged.

Detection of Nonhidden Block Duplications, Hidden Block Duplications, and Ghost Block Duplications

All rice scaffolds covering five or more genes (966 scaffolds, or 286.01 Mb) were used for the detection of block duplications using ADHoRe, a recently developed tool for the automatic detection of homologous regions. Homologous gene pairs for the two genomic fragments compared were determined using BLAST and homology-derived secondary structure prediction (Rost, 1999). The ADHoRe parameters were set to Q = 0.9 and G = 25 (Vandepoele et al., 2002a). Only block duplications that had a probability of being generated by chance of <0.1% (or a significance level of 99.9%) were retained in our analysis. For the determination of the number of tandem duplications within the rice genome, only homologous genes with five or fewer unrelated intervening genes were considered.

Apart from block duplications that can be recognized clearly (so-called obvious or nonhidden block duplications) and tandem duplications, we also discerned hidden and ghost duplications (Figure 2). Hidden duplications are heavily degenerated block duplications that cannot be observed by directly comparing the duplicated segments with each other; rather, they are observed only through comparison with a third segment (Simillion et al., 2002). Ghost duplications are defined as hidden duplications between different genomes. Thus, two genomic segments in the same genome form a ghost duplication when their homology can be inferred only through comparison with the genome of another species (Vandepoele et al., 2002b). To detect ghost duplications, initially, all colinear regions between rice and Arabidopsis were determined using ADHoRe (Q = 0.9, G = 25, and 99.9% significance level). Subsequently, all duplicated segments within Arabidopsis (Simillion et al., 2002) and all colinear regions between rice and Arabidopsis were mapped to infer networks of colinearity between both model plants and to detect ghost duplications in rice. Only nonhidden duplications and colinear regions with at least five anchor points were considered.

Age Estimation of Block Duplications

For all nonhidden block duplications that were shown to be statistically significant, the time of duplication (age in million years) was determined using a dating method based on the fraction of synonymous substitutions per silent site (K_s), as described previously (Simillion et al., 2002). In short, the mean K_s value (average of the estimates obtained by three methods) was derived for each anchor point. These values then were used to calculate the mean K_s for each block duplication, excluding outliers. The mean rate of synonymous substitutions for rice was considered to be 6.5 synonymous substitutions per 10^9 years (Gaut et al., 1996; Li, 1997).

Age Estimation of Individual Gene Pairs

First, the complete set of rice and Arabidopsis genes was used to determine all gene families based on sequence similarity. In this procedure, an all-against-all sequence comparison is performed at the protein level for the complete set of genes in a genome. Subsequently, the alignable region and sequence identity between two similar proteins are validated to infer genuine paralogous relationships (Li et al., 2001). Finally, a single-linkage clustering procedure is applied to assign individual genes to a gene family, given all paralogous relationships. For each gene family, the number of K_s was determined for all paralogous gene pairs by the method of Li (1993). Gene families with >10 members were excluded to reduce the number of gene family–specific pairwise comparisons.

Phylogenetic Reconstruction

Phylogenetic trees were constructed with the neighbor-joining algorithm as implemented in LinTree (Takezaki et al., 1995), based on Poisson distances inferred from amino acid sequence alignments. Bootstrap analysis involving 1000 resamplings was performed to test the significance of the internodes. For each pair of duplicated rice genes, a sequence similarity search (BLASTP; Altschul et al., 1997) was performed to detect homologous monocotyledonous gene sequences and an appropriate dicotyledonous outgroup. The detection of a suitable outgroup was performed by selecting the best hit with an E value of <1e-50 for one of the gene duplicates among a set of dicotyledonous proteins (i.e., all Arabidopsis proteins from TIGR combined with all other dicotyledonous proteins present in SWISS-PROT [Boeckmann et al., 2003]).

To detect homologous monocot gene sequences that contain sufficient information to reconstruct a reliable phylogenetic tree, two selection criteria were applied. First, all hits for both rice gene duplicates had to have an E value of <1e-10. Second, only sequences that had an alignable region of >150 amino acids with the query rice sequences were selected for the final phylogenetic analysis. The total set of monocotyledonous protein sequences contains 18,885 proteins, which were obtained by selecting SWISS-PROT proteins for *Triticum*, *Sorghum*, *Hordeum*, *Zea*, and *Avena*, translation of coding sequences from the National Center for Biotechnology Information (NCBI) Unigene collection for *Hordeum vulgare*, *Triticum aestivum*, and *Zea mays*, and the construction of open reading frames that show sequence similarity to rice proteins (BLASTX with E values of <1e-05) for all publicly available monocotyledonous ESTs and NCBI Unigenes lacking coding sequence information. Phylogenetic trees clearly in disagreement with the established grass phylogeny (Kellogg, 2001) or showing nonsignificant bootstrap values (<70%) were removed from further analysis.

Upon request, materials integral to the findings presented in this publication will be made available in a timely manner to all investigators on similar terms for noncommercial research purposes. To obtain materials, please contact Yves Van de Peer, yves.vandepeer@psb.ugent.be.

ACKNOWLEDGMENTS

We thank Stephane Rombauts for helpful discussions regarding the development of the ASGAR program, Yvan Saeys for developing the visualization tools, and Martine De Cock for help in preparing the manuscript. The authors also greatly acknowledge the valuable comments of two anonymous referees. K.V. and C.S. are indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk–Technologisch Onderzoek in de Industrie for predoctoral fellowships.

Received May 26, 2003; accepted June 26, 2003.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25,** 3389–3402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796–815.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12,** 1093–1101.

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13,** 137–144.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31,** 365–370.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422,** 433–438.

Chen, M., et al. (2002). An integrated physical and genetic map of the rice genome. Plant Cell **14,** 537–545.

Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A., and Bennetzen, J.L. (1997). Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. Proc. Natl. Acad. Sci. USA **94,** 3431–3435.

Durand, D. (2003). Vertebrate evolution: Doubling and shuffling with a full deck. Trends Genet. **19,** 2–5.

Feng, Q., et al. (2002). Sequence and analysis of rice chromosome 4. Nature **420,** 316–320.

Gale, M.D., and Devos, K.M. (1998). Comparative genetics in the grasses. Proc. Natl. Acad. Sci. USA **95,** 1971–1974.

Gaut, B.S. (2001). Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res. **11,** 55–66.

Gaut, B.S. (2002). Evolutionary dynamics of grass genomes. New Phytol. **154,** 15–28.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. USA **93,** 10274–10279.

Goff, S.A., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296,** 92–100.

Holland, P.W.H. (2003). More genes in vertebrates? J. Struct. Funct. Genomics **3,** 75–84.

Keller, B., and Feuillet, C. (2000). Colinearity and gene density in grass genomes. Trends Plant Sci. **5,** 246–251.

Kellogg, E.A. (2001). Evolutionary history of the grasses. Plant Physiol. **125,** 1198–1205.

Kishimoto, N., Higo, H., Abe, K., Arai, S., Saito, A., and Higo, K. (1994). Identification of the duplicated segments in rice chromosomes 1 and 5 by linkage analysis of cDNA markers of known functions. Theor. Appl. Genet. **88,** 722–726.

Levy, A.A., and Feldman, M. (2002). The impact of polyploidy on grass genome evolution. Plant Physiol. **130,** 1587–1593.

Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36,** 96–99.

Li, W.H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. (2001). Evolutionary analyses of the human genome. Nature **409,** 847–849.

Liu, H., Sachidanandam, R., and Stein, L. (2001). Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. Genome Res. **11,** 2020–2026.

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

Mayer, K., et al. (2001). Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. Genome Res. **11,** 1167–1174.

Nagamura, Y., et al. (1995). Conservation of duplicated segments between rice chromosome-11 and chromosome-12. Breed. Sci. **45,** 373–376.

Ohno, S. (1970). Evolution by Gene Duplication. (Berlin: Springer-Verlag).

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.X., Katsar, C.S., Lan, T.H., Lin, Y.R., Ming, R., and Wright, R.J. (2000). Comparative genomics of plant chromosomes. Plant Cell **12,** 1523–1540.

Paterson, A.H., et al. (1996). Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat. Genet. **14,** 380–382.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: The diverging fates of duplicated genes. Nat. Rev. Genet. **3,** 827–837.

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., and Van de Peer, Y. (2003). Investigating ancient duplication events in the *Arabidopsis* genome. J. Struct. Funct. Genomics **3,** 117–129.

Rice Chromosome 10 Sequencing Consortium (2003). In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300,** 1566–1569.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. **16,** 276–277.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng. **12,** 85–94.

Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., Sasaki, T., and Higo, K. (2002). RiceGAAS: An automated annotation system and database for rice genome sequence. Nucleic Acids Res. **30,** 98–102.

Salse, J., Piegu, B., Cooke, R., and Delseny, M. (2002). Synteny between *Arabidopsis thaliana* and rice at the genome level: A tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res. **30,** 2316–2328.

Sasaki, T., and Burr, B. (2000). International Rice Genome Sequencing Project: The effort to completely sequence the rice genome. Curr. Opin. Plant Biol. **3,** 138–141.

Sasaki, T., et al. (2002). The genome sequence and structure of rice chromosome 1. Nature **420,** 312–316.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genet. Dev. **6,** 715–722.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van De Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99,** 13627–13632.

Stephens, S.G. (1951). Possible significance of duplication in evolution. Adv. Genet. **4,** 247–265.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12,** 823–833.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12,** 381–391.

Terryn, N., et al. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the *APETALA2* locus on chromosome 4. FEBS Lett. **445,** 237–245.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., and Avramova, Z. (1999). Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. USA **96,** 7409–7414.

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

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and Van de Peer, Y. (2002a). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res. **12,** 1792–1801.

Vandepoele, K., Simillion, C., and Van de Peer, Y. (2002b). Detecting the undetectable: Uncovering duplicated segments in *Arabidopsis* by comparison with rice. Trends Genet. **18,** 606–608.

Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000). The origins of genomic duplications in *Arabidopsis*. Science **290,** 2114–2117.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2,** 333–341.

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

Yu, J., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296,** 79–92.

Zeng, L.W., Comeron, J.M., Chen, B., and Kreitman, M. (1998). The molecular clock revisited: The rate of synonymous vs. replacement change in *Drosophila*. Genetica **103,** 369–382.

Zhang, L., Vision, T.J., and Gaut, B.S. (2002). Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. **19,** 1464–1473.