Duplication and DNA segmental loss in the rice genome: implications for diploidization

Xiyin Wang[1,2,3]*, Xiaoli Shi[1,2]*, Bailin Hao[2,5], Song Ge[4] and Jingchu Luo[1]

[1]College of Life Sciences, National Laboratory of Plant Genetic Engineering and Protein Engineering, Center of Bioinformatics, Peking University, Beijing 100871, China, [2]Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China, [3]College of Mathematics, Hebei Polytechnic University, Tangshan, Hebei 063009, China, [4]Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; [5]Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China; *These authors contributed equally to this work

Authors for correspondence:
Jingchu Luo
Tel: +86 10 62757281
Fax: +86 10 62759001
Email: luojc@pku.edu.cn
Song Ge
Tel: +86 10 62591431x6097
Fax: +86 10 62590843
Email: gesong@ns.ibcas.ac.cn

Received: *25 August 2004*
Accepted: *27 September 2004*

Summary

• Large-scale duplication events have been recently uncovered in the rice genome, but different interpretations were proposed regarding the extent of the duplications.
• Through analysing the 370 Mb genome sequences assembled into 12 chromosomes of *Oryza sativa* subspecies *indica*, we detected 10 duplicated blocks on all 12 chromosomes that contained 47% of the total predicted genes. Based on the phylogenetic analysis, we inferred that this was a result of a genome duplication that occurred *c*. 70 million years ago, supporting the polyploidy origin of the rice genome. In addition, a segmental duplication was also identified involving chromosomes 11 and 12, which occurred *c*. 5 million years ago.
• Following the duplications, there have been large-scale chromosomal rearrangements and deletions. About 30–65% of duplicated genes were lost shortly after the duplications, leading to a rapid diploidization.
• Together with other lines of evidence, we propose that polyploidization is still an ongoing process in grasses of polyploidy origins.

Key words: diploidization, DNA loss, duplication, genome, polyploidy, rice (*Oryza sativa*).

New Phytologist (2005) **165**: 937–946

© *New Phytologist* (2005) **doi**: 10.1111/j.1469-8137.2004.01293.x

Introduction

The Gramineae (Poaceae) is a large angiosperm family, diverged from a common ancestor *c.* 55–70 million years ago (mya) (Jacobs *et al.*, 1999; Kellogg, 2001; Gaut, 2002). Many economically important crops belong to this family, including the best-characterized ones such as rice (Yu *et al.*, 2002; Zhao *et al.*, 2004), maize (Martienssen *et al.*, 2004) and wheat (Huang *et al.*, 2002). As extensive colinearity has been well maintained among divergent grass species (Moore *et al.*, 1995; Gale & Devos, 1998; Feuillet & Keller, 2002), grasses are taken as a single genetic system and genetic analyses in the grass family are likely to proceed with a greater level of efficiency because of potential cross-reference between different model systems such as rice and maize (Gale & Devos, 1998). With a small and sequenced genome, rice is most likely to take the central stage for the better understanding of genetic and evolutionary problems.

For decades, numerous large-scale duplications, probably resulting from polyploidy, have been documented in the grass family based on comparative mapping analysis of closely related grass species (Gale & Devos, 1998; Devos & Gale, 2000; Levy & Feldman, 2002). A large number of grasses, distributed in all main lineages of the family, are classified as polyploids based on cytological observation and restriction fragment length polymorphism (RFLP) markers (Stebbins, 1971; Levy & Feldman, 2002). These include some of the most important cereal crops such as bread wheat (*Triticum aestivum*), which was easily recognized as a hexaploid derived possibly from a hybridization between a tetraploid and a diploid progenitor *c.* 9500 yr ago (Ozkan *et al.*, 2001; Huang *et al.*, 2002). Maize (*Zea mays*) was also considered to have an ancient allopolyploid origin (Anderson, 1945; Helentjaris *et al.*, 1988) and was further proposed to have originated from a segmental allopolyploid event 11 mya (Gaut & Doebley, 1997).

The polyploidy origin of the rice genome has been a long-standing hypothesis with little supporting evidence (Nayar, 1973; Levy & Feldman, 2002). Based on the draft sequence of the genome of the subspecies *japonica* of rice, Goff *et al.* (2002) detected that approx. 59% of the cDNA markers had two or more copies and proposed that a whole-genome duplication occurred 40–50 mya, and another duplication involving chromosome 11 and 12 occurred *c.* 25 mya. To clarify the date and extent of the duplication in rice genome, Vandepoele *et al.* (2003) further analysed 2897 rice (ssp. *japonica*) bacterial artificial chromasome (BAC) sequences generated by the International Rice Genome Sequencing Project, and discovered that appox. 15% of all rice genes are in duplicated segments, with a major fraction of the duplication associated with chromosome 2. On this basis, they proposed that rice is an ancient aneuploid that has experienced the duplication at *c.* 70 mya. In their recent publication based on more completely assembled and annotated sequences, they further asserted the aneuploid hypothesis of rice genome although more duplicated genes (20%) are found in duplicated regions (Simillion *et al.*, 2004). Based on largely the same *japonica* genomic sequences, however, Paterson *et al.* (2003, 2004) found that much more extensive duplicated regions (approx. 61.9% of the rice transcriptome) existed in the rice genome, involving all rice chromosomes. Paterson *et al.*, 2003) suggested that an ancient polyploidy rather than aneuploidy occurred before the divergence of the major cereals, although they also dated the duplication event to *c.* 70 mya. In addition to the above inconsistency on the extent of the duplication, the frequency of duplications and the subsequent gene losses after duplication in the rice genome have not been fully investigated.

With the availability of the 370 Mb genome sequences assembled into 12 chromosomes of *Oryza sativa* subspecies *indica*, we made an extensive analysis of rice genome in order to test the hypotheses of the aneuploid vs. polyploidy origin of the rice genome. Our results showed that an ancient polyploidy occurred in the common ancestor of rice and maize *c.* 70 mya, strongly supporting the polyploid origin of the rice genome. In addition, we dated a segmental duplication involving chromosomes 11 and 12 to *c.* 5 mya. We also found that the diploidization, through large-scale losses of duplicated genes might have occurred shortly after the genome duplication.

Materials and Methods

Materials

A total of 53952 proteins of rice (ssp. *indica*; http://rise.genomics.org.cn/ GenBank Accession No. AAAA02000000) were predicted with the gene-finding program BGF developed by the Beijing Institute of Genomics (Zhao *et al.*, 2004; http://rise.genomics.org.cn/). Maize sequences were downloaded from EMBL (http://www.ebi.ac.uk/embl/) and rice cDNA from KOME (http://cdna01.dna.affrc.go.jp/cDNA/).

Duplicated genes and blocks

All-against-all BLASTP (Altschul *et al.*, 1997) was performed with rice proteins to reveal pairs of duplicated genes at the identity coverage (IC) > 60% which largely corresponds to the BLASTP E-value < 10^{-20}. The identity coverage of paired genes, similar to the criteria adopted by BLASTCLUST (Altschul *et al.*, 1997), was calculated as 2× (number of identical residues)/(sum of two protein lengths). With the gene pair information as input, a dynamic programming program DUPBLOCKSCAN was written and implemented to find the longest collinear regions, called duplicated blocks, among chromosomes. We arranged two chromosomes, referred to chromosome A and chromosome B and represented by all the genes positioned on them, along a gene-pairing information matrix, with chromosome A horizontally and chromosome B vertically. A cell of gene-pairing information matrix was filled with 1, if the corresponding gene on chromosome A and another on chromosome B constituted a gene pair according to the BLASTP identity coverage, otherwise the cell was filled with 0. We iteratively scanned throughout the matrix using a similar algorithm to the local alignment to reveal duplication blocks by a simple score scheme to count gene pairs. The extension of a block stopped if the present gene pair does not have another pair ahead in its field of vision at the distance of 1 Mb. After locating the longest block, the one corresponding to the largest score, we masked the path of the block in the gene-pairing information matrix by assigning the corresponding matrix cells with 0 and iterated the search until no block longer than 4 collinear gene pairs could be found.

For each putative duplicated block, we performed local permutation tests to check its significance. We reshuffled the paired genes on the duplicated block and ran DUPBLOCKSCAN to detect the longest random duplicated blocks in the local region. We calculated two ratios to reveal the significance of a block:

the length ratio = the number of collinearly paired genes in a putative block/the number of collinearly paired genes in the longest random block

and

the density ratio = the linear gene density in a putative block/the linear gene density in the longest random block

where the linear density was computed by the following formula

the linear gene density = the number of collinearly paired genes in a block/the sum of physical length of two block copies

The blocks were checked further if their lengths were longer and their collinearly paired genes were denser than all of the best random ones in 1000 permutation tests (the length ratio > 1 and the density ratio > 1). According to the two ratios, there

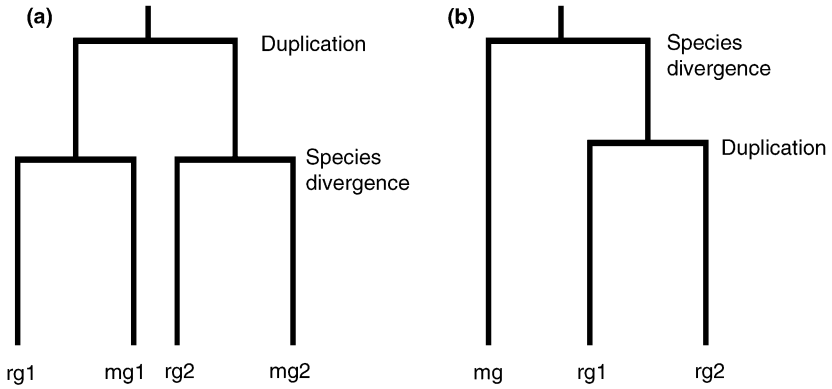

Fig. 1 Trees of duplicated genes produced by different duplications before and after the rice–maize divergence. (a) For a duplication event occurring before rice–maize divergence, there are two paralogous genes in a species and each of has an orthologue in the other species. (b) For a duplication occurring after the rice–maize divergence, the duplicated genes in rice have the same orthologue in maize.

was an obvious gap between 10 blocks and the others. These 10 blocks were taken as significant for further analysis.

Dating the polyploidy

To date the duplication events, we retrieved the maize homologues for the collinearly paired rice genes residing in the duplicated block. As the first step, we selected the collinearly paired rice genes with cDNA sequence information (Kikuchi *et al.*, 2003; http://cdna01.dna.affrc.go.jp/cDNA/). The genes with only two copies (at IC > 60%) in the rice genome were analysed to increase the chance of finding the true maize homologues. Second, we retrieved the best-matched maize genes from EMBL for the collinearly paired rice genes. These produced 433 clusters of the homologues of rice and maize genes. In two-thirds of the clusters the paired rice genes have the same maize homologues, while the rice genes in the other clusters have different maize homologues. To determine whether a duplication event happened before or after the rice–maize divergence, we checked each cluster whether there existed a maize homologue that was genetically more similar to one of the rice genes than the rice genes to each other (Fig. 1), and then tested the null hypothesis that the duplication occurred at the mean time of rice–maize divergence against the alternative that it occurred before or after the divergence.

After a genome duplication was dated before rice and maize divergence, we were sure that two collinearly paired rice genes in the blocks involved had different orthologues in maize. For further analysis, a possible maize orthologue was operationally defined, and for one of the paired rice genes if the maize gene was more similar to this rice gene than the maize gene to the other rice gene, and than the two rice genes to each other (Fig. 1a). We estimated the time of the genome duplication by the formula

$$1 + \left(\frac{median(d(rg1, rg2)) - median(d(o1, o2))}{median(d(o1, o2))} \right) \times T$$

(rg1 and rg2 are the collinearly paired rice genes which are paralogous; o1 and o2 are orthologous rice and maize genes; d(a,b) is the synonymous substitution rate (Ks) between homologous genes, the median of the Ks among paralogues or othologues were used here; T is the time elapsed since rice–maize divergence).

Another segmental duplication was proposed to have happened much later than rice–maize divergence. Then the two collinearly paired rice genes produced by this duplication have the same maize orthologue. We estimated the time of the segmental duplication by the formula (Fig. 1b)

$$\frac{median(d(rg1, rg2))}{median(d(o1, o2))} \times T$$

Estimation of gene loss rate

There are two copies for a duplicated block that located in different chromosomes and these were referred to as copy 1 and copy 2 for convenience. If a gene in copy 1 has no counterpart in copy 2, we assumed that the duplicated gene had been deleted in copy 2. By counting the single-copy genes in copy 1, we estimated the gene loss rate in copy 2 as follows:

$$\frac{S1}{S1 + N2} \times 100\%$$

($N2$ is the number of extant genes in copy 2; and $S1$ is the number of single-copy genes in copy 1).

Calculation of Ks

The Ks value (Nei & Gojobori, 1986) was calculated among collinearly paired rice genes and their maize homologues using the software package PAML (Yang, 1997) after sequence alignments using CLUSTALW (Thompson *et al.*, 1994).

Results

Detection of the duplicated blocks

A total of 25% (13885 out of 53952) predicted genes have duplicates in the present rice genome sequence. By scanning the possible duplicated blocks with at least five collinearly

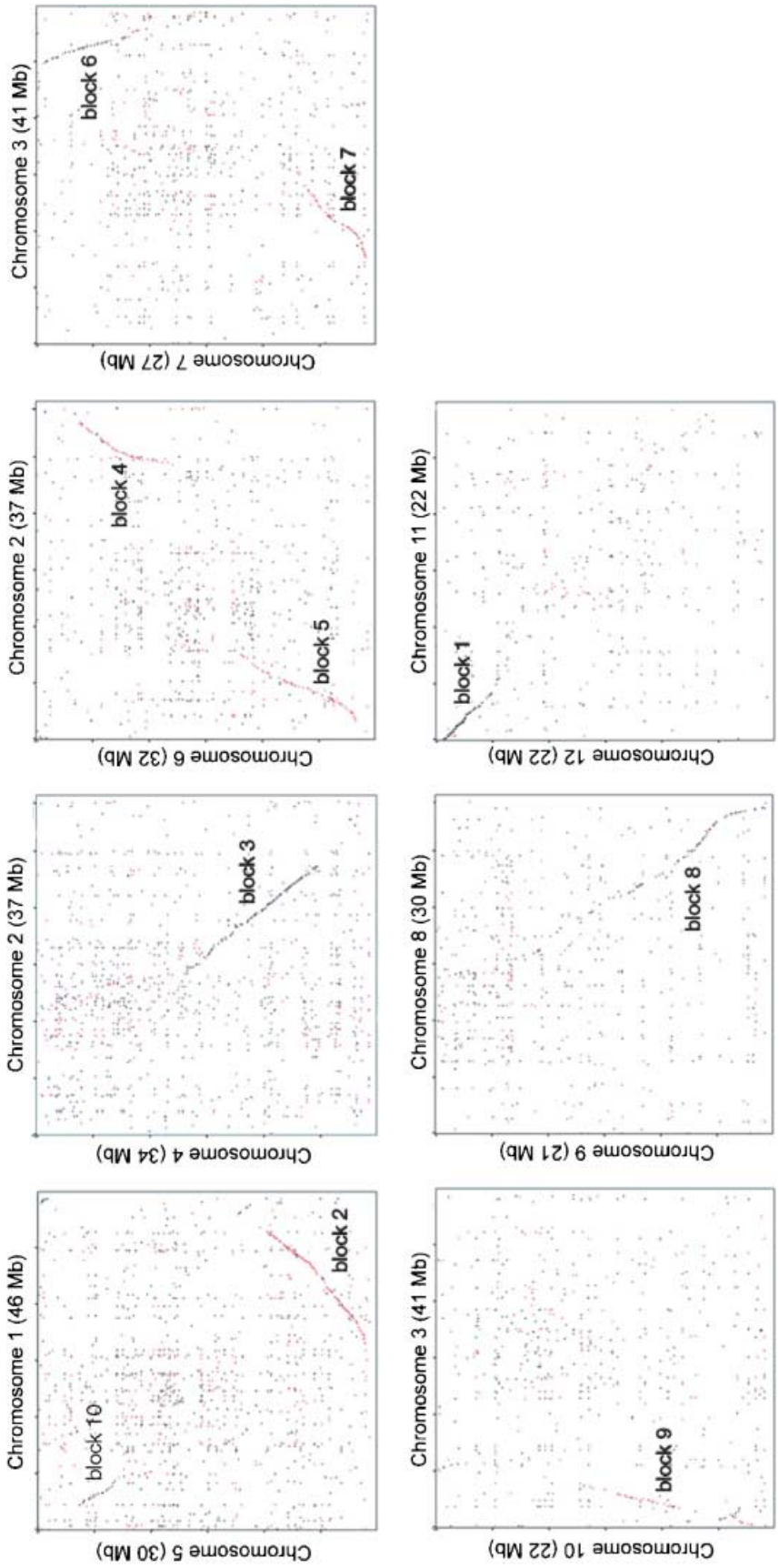

Fig. 2 Ten duplicated blocks between rice chromosomes. The gene pairs in the same transcription orientations on the two related chromosomes are marked with black dots, those in reverse orientations are marked with red dots.

Table 1 The duplicated genes in the 10 duplicated blocks

	Copy 1 of a duplicated block			Copy 2 of a duplicated block			Collinearly gene pairs in duplicated block			
Block	Chromosome	Length (Mb)	Number of genes	Chromosome	Length (Mb)	Number of genes	Number of gene pairs	Number of genes with cDNA	Orientation identity	Medians of Ks
1	11	5.44	895	12	4.27	702	148	70	95%	0.084
2	1	16.70	2600	5	9.50	1567	151	83	94%	0.743
3	2	14.62	2257	4	14.92	2288	125	66	84%	0.687
4	2	4.61	741	6	8.77	1225	55	37	91%	0.759
5	2	7.63	1100	6	11.98	1726	68	44	96%	0.818
6	3	4.15	712	7	8.27	1251	49	28	73%	0.797
7	3	8.97	1337	7	5.54	908	54	32	89%	0.713
8	8	10.37	1580	9	11.56	1833	65	39	89%	0.799
9	3	2.99	497	10	6.31	973	34	22	88%	0.804
10	1	4.18	625	5	4.67	644	23	12	91%	1.028
Total	–	79.7	12344	–	85.8	13117	772	433	–	–

Ks, synonymous substitution rate.

gene pairs, we obtained 2503 putative duplicated blocks and performed local permutation tests for each of them. We adopted two ratios, the length ratio and the density ratio, to measure the significance of a possible block and obtained 10 duplicated blocks (Fig. 2; Table 1). These 10 blocks contain 772 collinearly paired genes (73–95% identical transcription orientations) with 23–151 paired genes in each of them (Table 1). It is apparent from Table 1 that block 3 is the largest that involves chromosomes 2 and 4, whereas block 10 is the smallest involving chromosomes 1 and 5. Overall, the 10 blocks contain 165 Mb nucleotides that account for approx. 45% of the genome sequences (370 Mb) analysed in the present study (Fig. 2). It is important to note that the 10 blocks involve all 12 chromosomes (Table 1 and Fig. 2), indicating that large-scale duplications happened in the rice genome. Of the total of 53952 genes, 25461 (approx. 47%) were identified in the 10 blocks. The collinearly paired genes account for 6.1% of the total genes in the 10 blocks, whereas all the duplicated genes (2266) account for 9% of the total.

Level of Ks between duplicated blocks and genes

The Ks was calculated for each pair of the collinearly paired genes in the 10 blocks (Fig. 3). The medians of Ks of the duplicated genes in blocks 2–10 are much larger than that in block 1. We performed ANOVA with the null hypothesis that the means of Ks among the blocks were equal. The null hypothesis was rejected with a P-value 2.2×10^{-16} for all the 10 blocks, but was accepted at significance level 0.05 (P-value 0.06) when block 1 was excluded. This implies that blocks 2–10 were produced approximately at the same time and thus by the same evolutionary event, while block 1 was created by a different duplication event. Obviously, block 1 has a much smaller median of Ks and thus occurred more recently than the others (Fig. 3), which could be inferred to result from a segmental duplication. The blocks 2–10 involve 10 chromosomes and contain 156 Mb nucleotides and 23864 genes, accounting for approx. 42% and 43% of the present genome sequences and the total genes, respectively, indicating the possibility of a polyploid event. These results suggest that two different duplications occurred in the evolution of the rice genome (Fig. 3).

To further explore the nature of the ancient evolutionary events, we plotted the density curves of Ks of the collinearly paired genes in the blocks (Fig. 4). The first peak in the Ks density curve (on the left) for all blocks was produced by genes in block 1 (Fig. 4a), which have a unimodal Ks distribution (Fig. 4b). For the paired genes in blocks 2–10, the Ks density was plotted and a bimodal distribution was revealed with one peak at Ks = 0.46 and the other at 0.80 (Fig. 4c). Further exploration indicated that Ks was negatively correlated with GC content, predominantly the GC content at the third position of codons (GC3) with a Spearman correlation coefficient –0.455 and P-value 2.2×10^{-16}. According to GC3, the duplicated genes can be divided into two groups at GC3 = 0.52 with Euclidean distance using Ward's clustering method. The Ks bimodal distribution was split into two unimodal distributions (Fig. 4d,e) corresponding to the paired genes with GC3 < 0.52 and > 0.52, respectively. The density curve of the paired genes with smaller GC3 has a peak at Ks = 0.80, while that of genes with larger GC3 had a peak at Ks = 0.46. The peaks of these two curves (Fig. 4d,e) correspond perfectly to the two peaks of the density curve of all the genes (Fig. 4c). With the effect of GC3 considered, the ANOVA on Ks was redone for the two groups of genes in blocks 2–10. The null hypothesis was accepted with higher P-values: 0.52 for genes with GC3 > 0.52 and 0.27 for genes with GC3 < 0.52. The GC3 density curve has a large upper tail (Fig. 4f), which seems to be the cause for the heterogeneity of Ks. Thus, the bimodal distribution of Ks for all those paired

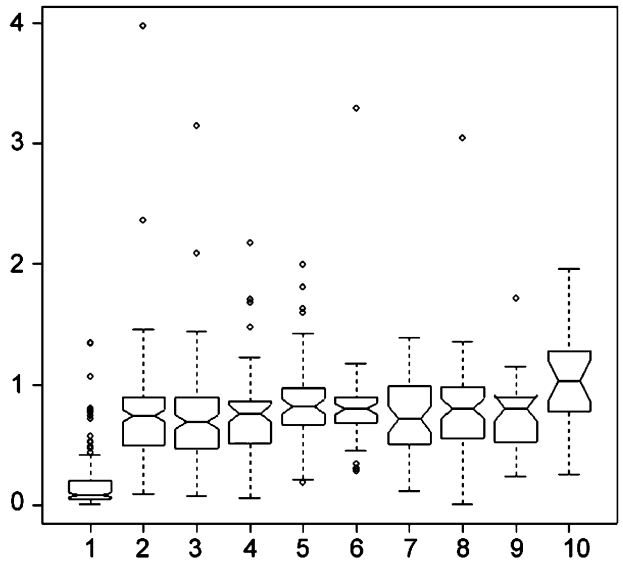

Fig. 3 The synonymous substitution rates (Ks) of genes in the 10 blocks. The median of block 1 is much smaller than those of other blocks, which are at roughly the same level.

genes is caused by GC3 rather than of the multiple duplications. Consequently, the Ks distribution patterns of the collinearly paired genes also indicated two duplication events in the evolution of the rice genome, in agreement with the above result.

Phylogenetic analysis of the duplication events

In order to clarify when (before or after the divergence of maize and rice) the duplication events took place, we analysed the collinearly paired genes with their maize homologues. We obtained 433 clusters of rice and maize homologues. While calculating the Ks with PAML, some clusters were discarded because the method adopted was not applicable to them or because stop codons were found in some of maize sequences. Finally, 341 clusters were used for further analysis. Of 289 clusters corresponding to the blocks 2–10, 234 support the duplication that occurred before rice–maize divergence with a *P*-value 2.2×10^{-16} in a χ^2 test. That is, in each of these clusters the maize gene is more similar to a rice gene than the rice genes are to each other (Fig. 1a). Of the other 52 clusters of rice and maize homologues corresponding to block 1, the values of Ks between the two paired rice genes in 48 clusters are considerably smaller than those between the rice genes and their maize orthologues (Fig. 1b), suggesting that a duplication associated with block 1 occurred after rice–maize divergence with a *P*-value 1.05×10^{-9} in a χ^2 test.

We dated the duplication by the formulas proposed and calculated the medians among the rice and maize paralogues or orthologues: medians of d(rg1,rg2), d(rg1,mg1), d(rg2,mg2) are 0.7663, 0.5527 and 0.5851, respectively, for the clusters corresponding to blocks 2–10; d(rg1,rg2), d(rg1,mg1), d(rg2,mg2) are 0.06370, 0.6937 and 0.6937, respectively, for the clusters corresponding to block 1. Assuming the divergence of rice and maize at *c.* 50 mya (Kellogg, 2001; Gaut, 2002), we inferred that the earlier duplication had occurred 16–20 mya before rice–maize divergence, i.e. *c.* 66–70 mya and the second duplication occurred only *c.* 4.6 mya.

DNA rearrangement and segmental loss

After the polyploidization, there were two nuclear genomes in one cell and two homologous copies for each chromosome.

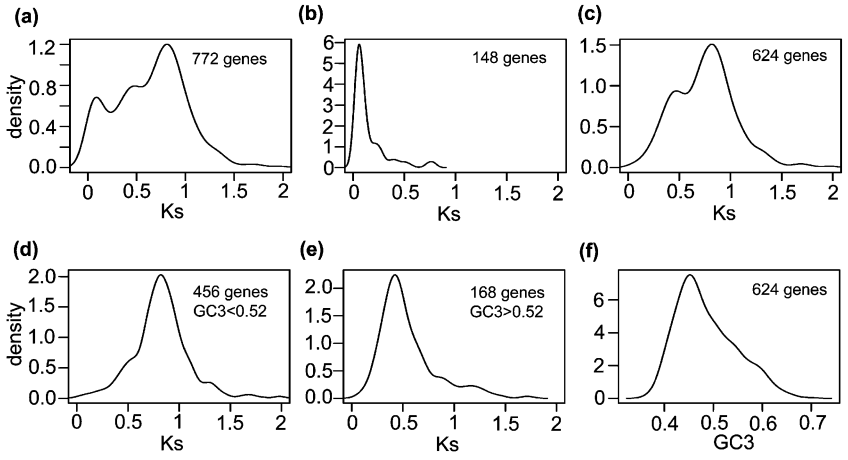

Fig. 4 The synonymous substitution rate (Ks) distribution and GC content at the third position of codons of duplicated blocks. (a) The Ks density curve of all collinearly paired genes in 10 blocks has three peaks. Genes in block 1 constitute the first peak on the left. (b) The Ks density curve of the paired genes in block 1 has a unimodal distribution. (c) The Ks density curve of all collinearly paired genes in blocks 2–10 has a bimodal distribution. Genes with GC contents at the third position of codons GC3 > 0.52 and < 0.52 correspond to the two peaks, respectively. (d) The Ks of the duplicated genes with GC3 < 0.52 in blocks 2–10 has a unimodal distribution. (e) The Ks of the duplicated genes with GC3 > 0.52 in blocks 2–10 also has a unimodal distribution. (f) Density curve of GC3 has a large upper tail, which explains the bimodal distribution of Ks for duplicated genes in blocks 2–10.

Table 2 Gene loss rates in 10 duplicated blocks

Block	Copy 1	Copy 2	Sum	Average
1	0.50	0.48	0.98	0.49
2	0.34	0.60	0.95	0.47
3	0.48	0.47	0.95	0.47
4	0.59	0.37	0.96	0.48
5	0.61	0.36	0.96	0.48
6	0.38	0.58	0.97	0.48
7	0.62	0.34	0.96	0.48
8	0.65	0.32	0.97	0.48
9	0.52	0.44	0.96	0.48
10	0.37	0.51	0.88	0.44

Although rearrangements and inversions followed (Fig. 2), the relocated blocks should have exactly two copies. By checking the rice block 3 on chromosomes 2 and 4, and blocks 4 and 5 on chromosomes 2 and 6, we can see a duplicated copy of chromosome 2 in the present rice genome (Fig. 2). This duplicated chromosome 2 was split into three segments, with one segment on chromosome 4 and the other two on chromosome 6. The most parsimonious explanation is that this duplication resulted from an interchromosomal rearrangement involving two ancient chromosomes.

In the duplicated blocks, segment DNA loss occurred at a considerably high level. Many genes in neighbouring locations in one copy do not have their counterparts in the other duplicated copy, indicating that some of the duplicated genes might be deleted after the polyploidization. It can be seen in Table 2, that the highest percentage of gene losses happened in one copy of block 8 on chromosome 8, in which 65% duplicated genes have been deleted, while in the duplicated counterpart on chromosome 9, only 32% duplicated genes have been deleted. Therefore, two copies of a duplicated block might have very different gene loss rates. The loss rates in the two copies of blocks 2, 7 and 8 vary up to twofold. Although the loss rates are very different among blocks and between two copies of the same block, the average loss rates of two copies of the same block vary little (44–50%) (Table 2), implying an unknown mechanism in controlling the losses of the duplicated genes. Similarly, we calculated the gene loss rate on block 1 that resulted from a much younger duplication. It is interesting to find that the gene loss rates are 50% on chromosome 11 (copy 1) and 48% on chromosome 12 (copy 2) with the mean of 49%, at the same level as those on the other nine blocks (Table 2).

Discussion

Rice is an ancient polyploid

The hypothesis that rice is of a polyploid origin is not novel because the first suggestion that rice is a secondary polyploid was provided based on cytological observations more than 70 yr ago (Lawrence, 1931). Later, many authors also proposed that rice was a secondary balanced allotetraploid that originated through hybridization between two species (Nayar, 1973). Recently, large-scale duplication events have been uncovered in the rice genome, which raised the old question of rice polyploidy, though different interpretations have been suggested regarding the scale and timing of those duplications (Goff *et al.*, 2002; Vandepoele *et al.*, 2003; Paterson *et al.*, 2004). By analysing *c.* 2000 cDNA sequences and locating them in the draft sequence of the genome of the japonica subspecies of rice, Goff *et al.* (2002) suggested that a genome duplication event shaped the rice genome. Using the public data emerging from the International Rice Genome Sequencing Project on the subspecies japonica of rice, two groups recently investigated the duplication of the rice genome, but reached different conclusions regarding the extent of rice genome duplication (Paterson *et al.*, 2003, 2004; Vandepoele *et al.*, 2003; Simillion *et al.*, 2004). Based on a set of approximately physically ordered BACs of rice, Paterson *et al.* (2003, 2004) found nine nonoverlapping duplicated blocks accounting for 61.9% of the rice transcriptome and indicated that substantial, perhaps genome-wide, duplication occurred in the rice genome. By contrast, Vandepoele *et al.* (2003) and Simillion *et al.*, 2004) demonstrated that a smaller percentage of the rice genome (15% and 20%, respectively) existed in duplicated blocks and a major fraction of the duplications detected involved only one or two chromosomes. Consequently, they concluded that rice was not an ancient polyploid but rather an ancient aneuploid.

In the present study, we analysed the most updated sequence data of the genome of the subspecies *indica* of rice, and detected 10 duplicated blocks accounting for 45% of the genome sequences. All the duplicated blocks except for block 1 have the same level of Ks values (Fig. 3) and they spread uniformly over 10 of all 12 chromosomes (Table 1 and Fig. 2). Our findings are consistent with those reported by Paterson *et al.* (2004) but in disagreement with the results of Vandepoele *et al.* (2003) and Simillion *et al.* (2004). Clearly, the observed duplication pattern in the rice genome can only be explained by an entire-genome duplication or polyploidization event because such a widespread distribution of the duplicated segments with the same divergence time would not be expected if only one or a few chromosomes had duplicated (Blanc *et al.*, 2003).

Vandepoele *et al.* (2003) and Simillion *et al.* (2004) detected the collinear gene pairs based on linear aggression analysis. This approach, on one hand, relied on a linear relationship among the true collinear gene pairs in the dot maps or in the gene homology matrix and thus might lead to some duplicated blocks undetectable because uneven losses of genes in different regions of a duplicated block would result in a nonlinear relationship represented by a curve rather than a straight line in the dot maps (Fig. 2) and in the gene homology

matrix (Vandepoele *et al.*, 2003). On the other hand, the linearity criteria adopted by Vandepoele *et al.* (2003) and Simillion *et al.* (2004) might not tolerate occasional outliers deviating from a line under evaluation, which are common for gene duplication and deletion. These probably explain why much fewer duplicated blocks were obtained in the analyses of Vandepoele *et al.* (2003) and Simillion *et al.* (2004). In our case, we adopted a more effective and efficient approach. Based the paired genes, dot maps were drawn to provide a direct illustration. We then developed a dynamic program to uncover the collinearly paired genes, most of which were certain to result from the large-scale duplication according to the permutation test. In comparison, Paterson *et al.* (2003, 2004) also made dot maps but performed the analysis based on the syntenic gene pairs rather than the collinear gene pairs used by us, which might lead to a larger coverage of duplicated regions.

An ancient duplication predating the divergence of rice and maize

According to the phylogenetic models and statistical analysis, we dated a whole-genome duplication in the rice genome before the divergence of rice and maize, *c.* 70 mya. This date is much earlier than 40–50 mya estimated by Goff *et al.* (2002). It should be noted, however, that the date estimated by Goff *et al.* (2002) was based on the rate of amino acid substitution of all possible paralogous protein pairs in the rice genome. As pointed out by many workers (Wolfe, 2001; Seoighe, 2003; Vandepoele *et al.*, 2003), protein distances are not very reliable for the large-scale dating of heterogeneous classes of proteins.

Given the fact that the grass family originated *c.* 55–70 mya (Stebbins, 1981; Kellogg, 2001; Gaut, 2002), the ancient duplication may have occurred around the period of the grass family origin. This implies that the polyploidization event should be shared by most, if not all, of the extant grass species, including important crops such as maize, rice, wheat, barley and sugarcane because the rice lineage (subfamily Ehrhartoideae) has diverged from the maize lineage (subfamily Panicoideae) as early as 50 mya, while wheat and barley diverged from rice and maize relatively later (Kellogg, 2001; Gaut, 2002). In another words, this ancient duplication has affected almost all the lineages in the grass family, though many of them appear to be diploids. In this sense, we agree with Levy & Feldman (2002) that many more, if not all, higher plant species, considered as diploids because of their genetic and cytogenetic behaviour, are actually ancient polyploids (paleopolyploids).

We propose that the polyploidy occurred *c.* 66–70 mya – almost the same as that estimated by Paterson *et al.* (2004). However, the approach adopted by Paterson *et al.* (2004) may be questionable. First, the evolutionary rates were always derived based on different genes and the estimation of the proposed synonymous rates vary greatly among plants, from 4.1×10^{-9} to 7.0×10^{-9} synonymous substitution per site per yr (Wolfe *et al.*, 1987; Gaut *et al.*, 1996; Li, 1997). To use different rates would result in different time estimations. Second, as what we proposed, the genes in rice were divided into two groups by the GC content, especially the GC content at the third codon site and they have divergent Ks values. This was supported by the finding that there are two classes of genes in plants according to GC content (Carels & Bernardi, 2000). Therefore, we used a different strategy by calculating the ratio of the medians of Ks (see the Materials and Methods section). The median is a more 'stable' quantification of the distribution's centre and insensitive to outliers. This may provide some advantages when dating the evolutionary events.

Genome duplication or polyploidy is an ongoing process

Recurrent formation of polyploidy has been well discussed in grasses (Levy & Feldman, 2002). Maize was widely regarded to originate from an allotetraploid (Gaut & Doebley, 1997; Wilson *et al.*, 1999) and a segmental allopolyploidy was inferred to have occurred 11.4 mya between two ancient species that diverged *c.* 20.5 mya (Gaut & Doebley, 1997). Bread wheat is the best-characterized allohexaploid resulting from a hybridization between a tetraploid and a diploid 9500 years ago (Ozkan *et al.*, 2001; Levy & Feldman, 2002). Obviously, these polyploid events identified in both maize and wheat (and probably many others) are irrelevant to the ancient duplication event detected in this study. Therefore, at least two rounds of large-scale duplications have occurred in the evolutionary history of both maize and wheat. Similarly, despite its diploid nature, the cultivated rice has also experienced at least two rounds of large-scale duplications, as revealed in this study. If more species are surveyed in the grass family, many more large-scale but relatively recent duplications will become apparent. For example, many wild relatives of the cultivated rice are polyploids that originated mainly through hybridization, followed by chromosome doubling (Nayar, 1973; Ge *et al.*, 1999). Although extant rice is a diploid, as many as nine allotetraploid species have been described in the rice genus (*Oryza*), where 23 species, including the cultivated rice, are currently recognized (Ge *et al.*, 1999; Vaughan *et al.*, 2003). Phylogenetic studies show that in this genus some of tetraploids (HHJJ and HHKK species) originated much earlier than the others (BBCC and CCDD species) (Ge *et al.*, 1999), suggesting that allopolyploid events in the rice genus have happened recurrently. Briefly, most, if not all, grasses underwent an additional cycle of chromosome duplication, and thus were considered as 'neopolyploids' (Levy & Feldman, 2002). As in the case of *Arabidopsis* lineage, where three ancient duplications have been inferred (Blanc *et al.*, 2003; Vandepoele *et al.*, 2003), it can be expected that many grass lineages have experienced more than one round of genome duplications and polyploidy is not only widespread but also an ongoing process in the grasses (Vandepoele *et al.*, 2003).

Segmental loss and its implication for diploidization

DNA segmental loss is one of the chromosomal changes leading to diploidization. Recently, by analysing five local homologous regions among maize, rice, sorghum and barley, small rearrangements of genes were considered to be a key factor differentiating the grass genomes (Bennetzen & Ramakrishna, 2002). Based on the comparative linkage maps, Ahn & Tanksley (1993) found that *c.* 28% of duplicated genes was lost or mutated so as to be undetectable in maize. It was proposed that only 13.1% of the duplicated genes resulted from a duplication event 300 mya were left and *c.* 29.7% of the duplicated genes from a much younger duplication (20–80 mya), were still present in the *Arabidopsis* genome (Bowers *et al.*, 2003; Paterson *et al.*, 2004). Although we observed that the duplication involving rice chromosomes 11 and 12 occurred only 5 mya, the gene loss rates estimated on the two chromosomes were at the same level as those in the other duplicated blocks with much more ancient origin (Table 2). This implies that the gene losses or diploidization process might take place at the early stage of the polyploid and the duplicated genome tended to be stable later. This was partly supported by experimental research in the synthetic polyploid species (Ozkan *et al.*, 2001; Levy & Feldman, 2002).

A possible result of diploidization is the suppression of intergenomic recombination structurally, which might increase fertility (Levy & Feldman, 2002). To what extent the duplicated DNA was lost so that the genome could recover its stability is still unclear. The availability of the rice genome sequences provided a unique chance to inspect DNA elimination after duplication. Although the loss rates vary greatly among the copies of different duplicated blocks in the rice genome, the average loss rates of all the blocks are essentially the same (Table 2). The sum of loss rates of the two copies of each block are > 88%, suggesting that most of the duplicated genes are single copy through elimination of their counterparts in the duplicated blocks. Therefore, we infer that stabilization involved the deletion of one copy of a majority of the duplicated genes in the duplicated blocks.

Acknowledgements

We thank Jun Yu, Liping Wei, Jun Wang, Fumin Zhang, Yanbin Yin, Qihui Zhu, Di Liu and Min Zhao for helpful discussions. We are also grateful to Tao Sang of Michigan State University, USA, for critically reading the manuscript. This study was supported by the National Key Basic Research Program of China (2003CB715900), the National Natural Science Foundation of China (30025005, 30170232) and the 863 project.

References

Ahn S, Tanksley SD. 1993. Comparative linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences, USA* **90**: 7980–7984.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Anderson E. 1945. What is *Zea mays*? A report of progress. *Chronicle of Botany* **9**: 88–92.

Bennetzen JL, Ramakrishna W. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Molecular Biology* **48**: 821–827.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research* **13**: 137–144.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

Carels N, Bernardi G. 2000. Two classes of genes in plants. *Genetics* **154**: 1819–1825.

Devos KM, Gale MD. 2000. Genome relationships: the grass model in current research. *Plant Cell* **12**: 637–646.

Feuillet C, Keller B. 2002. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Annals of Botany* **89**: 3–10.

Gale MD, Devos KM. 1998. Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences, USA* **95**: 1971–1974.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. *New Phytologist* **154**: 15–28.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences, USA* **94**: 6809–6814.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences, USA* **93**: 10274–10279.

Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences, USA* **96**: 14400–14405.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.

Helentjaris T, Weber D, Wright S. 1988. Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics* **118**: 353–363.

Huang S, Sirikhachornkit A, Su X, Faris J, Gill B, Haselkorn R, Gornicki P. 2002. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences, USA* **99**: 8133–8138.

Jacobs BF, Kingston JD, Jacobs LL. 1999. The origin of grass-dominated ecosystems. *Annals of the Missouri Botanical Garden* **86**: 590–643.

Kellogg EA. 2001. Evolutionary history of the grasses. *Plant Physiology* **125**: 1198–1205.

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M,

Ryu R, Ueda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301**: 376–379.

Lawrence WJC. 1931. The secondary association of chromosomes. *Cytologia* **2**: 352–384.

Levy AA, Feldman M. 2002. The impact of polyploidy on grass genome evolution. *Plant Physiology* **130**: 1587–1593.

Li WH. 1997. *Molecular evolution*. Sunderland, MA, USA: Sinauer Associates.

Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR. 2004. Sequencing the maize genome. *Current Opinion in Plant Biology* **7**: 102–107.

Moore G, Devos KM, Wang Z, Gale MD. 1995. Cereal genome evolution: grasses, line up and form a circle. *Current Biology* **5**: 1995.

Nayar NM. 1973. Origin and cytogenetics of rice. *Advances in Genetics* **17**: 153–292.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**: 418–426.

Ozkan H, Levy AA, Feldman M. 2001. Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops–Triticum*) group. *Plant Cell* **13**: 1735–1747.

Paterson AH, Bowers JE, Peterson DG, Estill JC, Chapman BA. 2003. Structure and evolution of cereal genomes. *Current Opinion in Genetics and Development* **13**: 644–650.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences, USA* **101**: 9903–9908.

Seoighe C. 2003. Turning the clock back on ancient genome duplication. *Current Opinion in Genetics and Development* **13**: 636–643.

Simillion C, Vandepoele K, Saeys Y, van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Research.* **14**: 1095–1106.

Stebbins GL. 1971. *Chromosomal evolution in higher plants*. London, UK: Edward Arnold.

Stebbins GL. 1981. Coevolution of grasses and herbivores. *Annals of the Missouri Botanical Garden* **68**: 75–86.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673–4680.

Vandepoele K, Simillion C, van de Peer Y. 2003. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* **15**: 2192–2202.

Vaughan DA, Morishima H, Kadowaki K. 2003. Diversity in the *Oryza* genus. *Current Opinion in Plant Biology* **6**: 139–146.

Wilson WA, Harrington SE, Woodman WL, Lee M, Sorrells ME, McCouch SR. 1999. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* **153**: 453–473.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* **2**: 333–341.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* **84**: 9054–9058.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Bioscience* **13**: 555–556.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92.

Zhao W, Wang J, He X, Huang X, Jiao Y, Dai M, Wei S, Fu J, Chen Y, Ren X, Zhang Y, Ni P, Zhang J, Li S, Wang J, Wong GK, Zhao H, Yu J, Yang H, Wang J. 2004. BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research* 32(Database issue): D377–D382.