Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Incongruent evolution of chromosomal size in rice

X. Guo[1], G. Xu[2], Y. Zhang[1], X. Wen[2], W. Hu[2] and L. Fan[1,2]

[1]Institute of Bioinformatics and [2]Institute of Crop Science,
Zhejiang University, Hangzhou, China
Corresponding author: L. Fan
E-mail: fanlj@zju.edu.cn

Genet. Mol. Res. 5 (2): 373-389 (2006)
Received December 20, 2005
Accepted May 5, 2006
Published June 22, 2006

ABSTRACT. To investigate genome size evolution, it is usually informative to compare closely related species that vary dramatically in genome size. A whole genome duplication (polyploidy) that occurred in rice (*Oryza sativa*) about 70 million years ago has been well documented based on current genome sequencing. The presence of three distinct duplicate blocks from the polyploidy, of which one duplicated segment in a block is intact (no sequencing gap) and less than half the length of its syntenic duplicate segment, provided an excellent opportunity for elucidating the causes of their size variation during the post-polyploid time. The results indicated that incongruent patterns (shrunken, balanced and inflated) of chromosomal size evolution occurred in the three duplicate blocks, spanning over 30 Mb among chromosomes 2, 3, 6, 7, and 10, with an average of 20.3% for each. DNA sequences of chromosomes 2 and 3 appeared to had become as short as about half of their initial sequence lengths, chromosomes 6 and 7 had remained basically balanced, and chromosome 10 had become dramatically enlarged (~70%). The size difference between duplicate segments of rice was mainly caused by variations in non-repetitive DNA loss. Amplification of long terminal repeat retrotransposons also played an important role. Moreover, a relationship seems to exist between the chromosomal size differences and the nonhomologous combination in corresponding regions in the rice genome. These findings help shed light on the evolutionary mechanism of genomic sequence variation after polyploidy and genome size evolution.

Key words: Chromosomal size evolution, *Oryza sativa*, Polyploidy, Genome size evolution, Incongruent patterns

INTRODUCTION

Genome size varies tremendously in plant species, and this variation is not likely to correlate with the extent of complexity of the organism, known as the "C-value paradox" (Thomas, 1971). For example, the size of the barley (*Hordeum vulgare*) genome is 11- and 35-fold larger than that of rice and *Arabidopsis*, respectively, although they are at a similar level of complexity (Bennett and Leitch, 1997). Numerous processes (mechanisms) determine the increase or decrease in genome size (Bennetzen, 2002). In the early 1930s, gene duplication was regarded as the first proposed mechanism for the increase of genetic content (Betran and Long, 2002). In plants, many mechanisms are now known to be responsible for increasing gene number by the duplication of genes, DNA segments or whole genomes (Ohno, 1970). Polyploidy was one of the most prominent forces to expand the genome (Wendel, 2000; Grover et al., 2004). Another key factor involved in genome expansion is transposable element amplification. It has been suggested that in the grass family long terminal repeat (LTR)-retrotransposon amplification in the recent 10 million years contributed to most of the genome size expansion (SanMiguel et al., 1996, 1998; Vicient et al., 1999; Shirasu et al., 2000; Wicker et al., 2001; Vitte and Panaud, 2003; Ma et al., 2004; Ma and Bennetzen, 2004). Apparently, this "one-way ticket to genomic obesity" (Bennetzen and Kellogg, 1997) was not the simple end of the evolutionary scenario. The return ticket was provided later; illegitimate recombination and unequal homologous recombination are the prominent mechanisms in the reduction of LRT-retrotranspon sequences counteracting genome expansion (Vicient et al., 1999; Shirasu et al., 2000; Devos et al., 2002; Ma et al., 2004). Recent LTR-retrotransposon sequence losses have been observed in rice and *Arabidopsis* genomes (Devos et al., 2002; Vitte and Panaud, 2003; Ma et al., 2004). In rice, at least 190 Mb of LTR-retrotransposon sequences were estimated to have been removed from the genome in the recent 8 million years (Ma et al., 2004). Petrov and his colleagues found that an imbalance of small deletions and insertions caused a reduction in genome size evolution in insects based on non-LTR-retrotransposon studies (Petrov et al., 1996, 2000; Petrov, 2002). However, it is yet to be determined whether a similar mechanism affected the transposable elements or whether another mechanism was responsible for the loss of non-repetitive DNA in plants.

To extrapolate the potential mechanisms for genome size evolution, it is desirable to compare genome sizes of closely related species that show a dramatic variation in genome size. A comparison was made between the compact *Drosophila* genome (165 Mb) and two extremely large genomes of the related insects *Laupala* crickets (1910 Mb) and *Podisma* grasshoppers (18,150 Mb) to determine the mechanisms of DNA loss (Petrov, 2002). Recently, ~105-kb contiguous orthologous sequences from two co-resident genomes (AA and DD) of the allopolyploid cotton species (*Gossypium hirsutum*) were employed to perform a pattern analysis of cotton genome size evolution (Grover et al., 2004).

Whole genome duplication took place in a more widespread manner in the flowering plants and is believed to play an important role in species evolution and divergence (Wendel, 2000). After the doubling of chromosomes, ancient polyploid or paleopolyploids appeared to have undergone a rapid genome evolution, i.e., a process of “diploidization” along with extensive DNA sequence elimination and chromosomal rearrangements (Wendel, 2000; Eckardt, 2001). The nearly complete genome sequences of rice (Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003) provide an unprecedented chance for investigating the evolutionary history of the rice genome. An ancient polyploid origin for the rice genome has been well documented (Paterson et al., 2004; Guyo and Keller, 2004; Zhang et al., 2005; Wang et al., 2005; Yu et al., 2005). The polyploidization event was estimated to have occurred ~70 million years ago before the divergence of the cereals from a common ancestor. This finding resulted from the observation of many large-scale, non-overlapping duplicated blocks almost covering the whole rice genome. For example, chromosome 2 was completely covered by the syntenic duplicated segments from chromosomes 4 and 6, while chromosome 3 by chromosomes 7, 10 and 12 (Zhang et al., 2005). In the duplicated blocks, there were perfect conservations in the order and orientation of genes in two counterpart segments. The duplicated blocks were found to be so clear and intact; no such case has ever been observed in other organisms. This finding thus provides an excellent opportunity for studying post-polyploid evolution of duplicate genes in the rice genome.

Two duplicate segments or chromosomes from a duplication event often had the same size at the beginning when the event occurred. The size of two paralogous segments might have become different after their long evolution. Duplicated blocks produced by the process of genome duplication (ohnologous blocks) underwent a sharp evolutionary process of “diploidization” along with extensive DNA sequence elimination, etc. (Eckardt, 2001). Similar to closely related species, the ohnologous segments that vary dramatically in size are useful for the elucidation of genome size evolution. In this study, an effort is made to model genome size evolution after polyploidy, and three duplicated blocks with distinct sizes among chromosomes 2, 3, 6, 7, and 10 were selected for comparative analysis of DNA sequence changes between the two ohnologous segments picked out from a recent whole genome duplicate in the rice genome. Our results suggest that incongruent patterns of chromosomal size evolution had occurred in rice during the last 70 million years in the post-polyploid evolutionary time, and that the size difference between duplicated segments in rice was mainly caused by variations in non-repetitive DNA loss, with LTR amplification also playing an important role.

MATERIAL AND METHODS

Sequence data sources

Twelve assembled chromosome sequences (pseudomolecules) of *japonica* rice Nipponbare and their annotation were downloaded from The Institute for Genomic Research (TIGR; www.tigr.org, osa1, version 2.0). The 12 pseudomolecules (virtual contigs) for each of the 12 rice chromosomes spanned a total 364.9 Mb with 59,712 annotated coding sequences in this release.

Genomic sequence analysis

Repeat sequences were first filtered by online RepeatMasker (http://www.repeatmasker.org; July 9, 2004), and the duplicated homologs from target duplicated blocks were then aligned (osa1, version 2.0) with each other to determine their syntenic and un-syntenic regions using Blastz program (Schwartz et al., 2003). An un-syntenic region >10 kb was taken as a segmental insertion when over 50% of its entire sequence could be found in several chromosomes. To ensure that all putative segmental insertions were found in the duplicated blocks, the whole genomic sequences of duplicated blocks were also aligned with all chromosomes and chloroplast (X15901) or mitochondria (AB076665, AB076666) genome sequences of rice using MUMmer (Delcher et al., 2002) with default values.

Gene analysis

Syntenic genes of TIGR's annotated genes in the target duplicated blocks were identified based on the synteny lines of duplicated blocks in Figure S1. Other genes were globally aligned to their syntenic or non-syntenic genes using the needle program of EMBOSS. A gene was regarded as a newborn gene when its amino acid substitution rate (d_A) relative to one of the syntenic genes in their corresponding duplicated region or genes in other regions (needle identity >50%) was less than 0.2, otherwise it was regarded as a unique gene, or retained gene after the ancient whole genome duplication. The d_A values among protein pairs were estimated using the aaml program of the PAML package (Yang, 1999) with the Dayhoff matrix.

Analysis of LTR-retrotranspon structures

Determination of LTR-retrotranspon structures in the three target genomic regions was performed as described by Ma et al. (2004). Elements of 77 annotated LTRs in Repbase database (http://www.repeatmasker.org; version March 6, 2004) were identified in the genomic regions using BLASTN searches. An intact element is defined as one that contains two relatively intact LTRs (more than 90% length coverage of the annotation LTR and a terminal with TG/CA), polypurine tract and primer binding sites, and that is flanked by short target site duplications. Solo LTR refers to any relatively intact LTR flanked by target site duplications.

RESULTS

Three duplicated blocks that vary dramatically in size

A total of 9 clear and large duplicated blocks produced by genome duplication were detected in rice using the first assembly of the rice genome prepared by TIGR (Paterson et al., 2004). Of the 9 blocks, three from chromosomes 2, 3, 6, 7, and 10 were distinct in size: for each block, at least one duplicated segment remained intact (no sequencing gap) (Yuan et al., 2003) and was less than half the length of its syntenic or duplicated partner. This raises an interesting question: what mechanism caused them to become different in size during the period of ~70 million years after the whole genome duplication event?

Based on an updated TIGR assembly of the rice genome (osa1, version 2.0) (Figure S1) (Zhang et al., 2005), genomic regions of three blocks (2-6b, 3-7b and 3-10b) were determined strictly and used in this study. The three blocks spanned a total 30.25 Mb, which covered 8.3% of the 364.9-Mb rice genome, and 12.9, 11.8, 30.9, 24.3, and 21.8% of chromosomes 2, 3, 6, 7, and 10, respectively. Ratios of two duplicated segment sequence lengths were 0.49:1 (2-6b), 0.48:1 (3-7b) and 0.17:1 (3-10b), respectively (Table 1), i.e., in the target regions, chromosome 10 was about 6-fold larger than chromosome 3 in size, and chromosome 6 and 7 were about 2-fold larger than chromosomes 2 and 3. The three duplicated blocks corresponded to the 3F, 4F and 5F blocks as reported by Paterson et al. (2004). A summary of annotated genes in our three blocks is listed in the supporting information (Table S1).

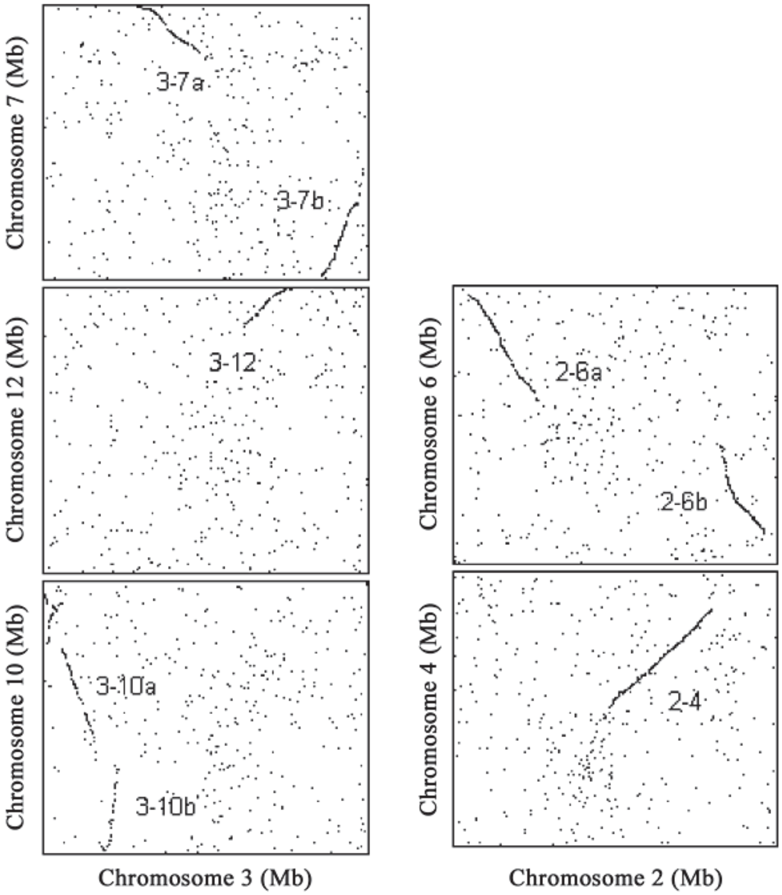

Figure S1. Selected duplicated blocks of a whole genome duplication in rice genome. Dot plots of inter-chromosome one-to-one paralogous or duplicated gene pairs in the rice genome based on the annotated coding sequences (cds) of TIGR's pseudomolecules (osa1, version 2.0). One dot represents a one-to-one paralogous gene pair. A total of 59,712 annotated coding sequences of *Oryza* (www.tigr.org, osa1, version 2.0) encoded by their chromosomal order were compared using reciprocal BLASTN searching (E<e-14) for any two chromosomes. Two sequences were defined as one-to-one paralogous or duplicated gene pairs when each of them was the best hit of the other. A pair was removed if the two duplicated genes synchronously BLASTN matched (<1e-10) members of the rice repeat database by TIGR. Also see Zhang et al. (2005).

Table 1. Evolutionary scenarios of genomic sequences in the three duplicated blocks after polyploidy.

Duplicated blocks	Current sequences (Mb)									Lost region (Mb)[¶]	Percent of initial length[§]	Sequence changes (Mb)[¥]	Percent of current length	Percent of initial length
	Genomic position	Length (gap number)	Synteny*	Increased[†]	Repeat (of current)	New genes	Organellar inserts	Inter-inserts (number)	Remnant[‡]					
2-6b	Chr2	33.86 - 29.30	4.56 (0)	0.375 (8.2%)	0.881 (19.3%)	0.488	0.001	0.030 (1)	2.786	5.283	62.6%	-3.884	-85.2%	-46.0%
	Chr6	3.22 - 12.50	9.28 (2)	0.396 (4.3%)	2.462 (26.4%)	1.052	0.035	0.051 (3)	5.283	2.786	32.9%	0.781	+8.4%	+9.2%
3-7b	Chr3	31.31 - 34.75	3.44 (0)	0.211 (6.1%)	0.617 (17.9%)	0.476	0.023	0	2.113	4.380	65.3%	-3.286	-95.5%	-49.0%
	Chr7	0.91 - 8.11	7.20 (0)	0.212 (2.9%)	1.774 (24.7%)	0.805	0.001	0.028 (2)	4.380	2.113	31.5%	0.494	+6.9%	+7.4%
3-10b	Chr3	7.44 - 8.26	0.82 (0)	0.053 (6.5%)	0.163 (19.9%)	0.103	0.002	0	0.498	2.369	81.1%	-2.103	-256.4%	-72.0%
	Chr10	1.03 - 5.98	4.95 (2)	0.056 (1.1%)	1.857 (37.5%)	0.663	0.005	0	2.369	0.498	17.0%	2.022	+40.9%	+69.2%

*Syntenic regions were identified using the BLASTZ program;

[†] In the increased regions, repeat parts were scanned by RepeatMask; parts of the new genes were summed based on their DNA sequence lengths in Table S1; inter-inserts refer to over 10-kb segmental insertions from other chromosomes (Chr) or other regions of the same chromosome;

[‡] Remnant length equals the difference between the increased length and the current length of the syntenic segment;

[¶] Lost region is the retained region (remnant) in its duplicated segment;

[§] Initial length is the sum of the lengths of the syntenic, retained and lost regions;

[¥] Sequence changes refer to the added regions minus lost region. “+” and “-” refer to increase and decrease of DNA sequence, respectively.

Table S1. Summary of current annotated genes in the three duplicated blocks.

Duplicated blocks	Current total number (No./kb)	Synteny*	Increased[†]				Retained (unique)[‡]	Syntenic homologs		Lost gene number[¶]	Of initial total[§]	Gene number changes of initial total[¥]
			Repeat	New genes	Inserts	Of current total		Of current total	Of initial total			
2-6b												
Chr2	742 (0.16)	124	42	158	4	27.5%	414	16.7%	9.5%	761	58.6%	-42.9%
Chr6	1427 (0.15)	124	162	371	9	38.0%	761	8.7%	9.5%	414	31.9%	+9.9%
3-7b												
Chr3	617 (0.18)	84	20	156	0	28.5%	357	13.6%	7.5%	679	60.6%	-44.9%
Chr7	1162 (0.16)	84	88	305	6	34.3%	679	7.2%	7.5%	357	31.9%	+3.8%
3-10b												
Chr3	132 (0.16)	16	8	36	0	33.3%	72	12.1%	3.2%	413	82.4%	-73.7%
Chr10	808 (0.16)	16	120	259	0	46.9%	413	2.0%	3.2%	72	14.4%	+61.3%

*Syntenic genes refer to one-to-one paralogs on the synteny line by reciprocal BLASTN.

[†] In increased genes, repetitive genes were BLASTN matched (<1e-10) with members of the TIGR rice repeat database. New genes refer to those newborn genes after the large-scale duplication; inserted genes refer to genes with organellar or segmental insertions in Table 1.

[‡] Retained gene number equals the current total number minus the syntenic and increased gene number.

[¶] Lost gene number equals the retained gene number in its duplicated homolog.

[§] Initial total gene number is the sum of the syntenic, retained and lost gene numbers.

[¥] Gene number changes refer to added gene number minus lost gene number. "+" and "-" refer to increase and decrease of gene number, respectively.

Chr = chromosome.

Synteny

In two duplicate segments from a duplication event, synteny would become detectable if they were not exposed to extremely sharp evolutionary forces. The syntenic regions within the three duplicate blocks were detected based on genomic sequence homology using the BLASTZ program, an independent implementation of the Gapped BLAST algorithm, which was specifically designed for aligning two long genomic sequences (Schwartz et al., 2003) and which has been successfully used in human-mouse-rat alignments (Kent et al., 2003; Rat Genome Sequencing Project Consortium, 2004). Three clear synteny lines could be observed in the three duplicated blocks (Figure 1). The proportions of syntenic regions in corresponding duplicated segments were 6.1-8.2% for short segments and 1.1-4.3% for long ones (Table 1). The “paths” of synteny based on genomic sequences were similar to that based on genes (Figure S1). The retained syntenic duplicated genes within the three blocks have been well documented using annotated genes in them and TIGR’s first (Paterson et al., 2004) and second release (Zhang et al., 2005), which served as evidence for genome duplication. A detailed list of syntenic duplicate genes based on TIGR’s second release is given in Tables S1 and S2.

Table S2 shows a summary of information about duplication (or lack thereof) of every gene in the duplicated blocks that were used in this study. See Figure S1 and Table 1 for the genomic positions of the three duplicated blocks. The genes were listed by their syntenic positions in a duplicated block. The following details are given for some items. 1) Best_hit pair: the one-to-one paralogous pair was indicated as syntenic when they occurred on the synteny line of the duplicated blocks, otherwise, non-syntenic. 2) Classes: repeat genes were BLASTN matched (<1e-10) with members of the TIGR rice repeat database. Newgene, which had >50% identity of global alignment and a d_A value <0.2 with the syntenic genes (newgene1) in the corresponding duplicated region or genes (newgene2) in other regions, refers to those newborn genes after the large-scale duplication. Inserted genes refer to genes with the segmental insertions shown in Table 1. Deleted genes correspond to those unique genes with their duplicated homologs. 3) Annotation follows TIGR’s annotation (osa1, version 2.0).

Table S2 is available at http://www.funpecrp.com.br/gmr/year2006/vol2-5/pdf/gmr0199tableS2.pdf.

DNA expansion

Numerous processes exist for increasing genome size. First, the repeat element is the most important one. In general, more repeat sequences appeared in the long-syntenic duplicate segments (from 24.7 to 37.5%) than in the short segments (7.9-19.9%) (Table 1). The ratios of repeat length in two syntenic segments were about 1:3 for blocks 2-6b and 3-7b and 1:11 for block 3-10b, which were higher than that of their current sequence lengths (about 1:2 for 2-6b and 3-7b and 1:6 for 3-10b, see above). The results suggest that repeat sequences contributed predominantly to the differences in sequence size differences within the blocks, particularly for the block of 3-10b. Of the repeat elements, LTR-retroelements and DNA transposons (En-spm, TC1-IS630-Pogo, Tourist/Harbinger, etc.) played major roles in the repeat-derived extension of the rice genome (Table S3). In the above two elements, the contribution of LTR-retroelements

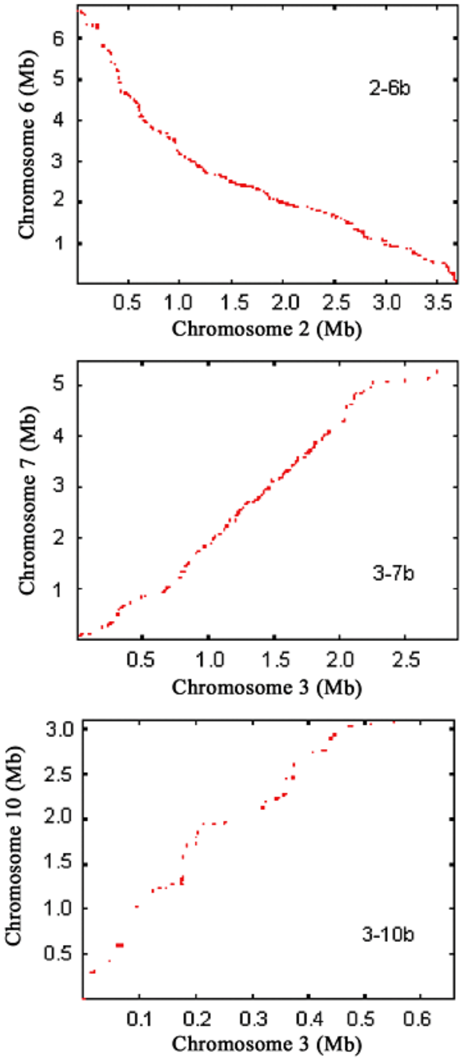

Figure 1. BLASTZ-based syntenic regions in three duplicated blocks of rice genome. TIGR's pseudomolecules of rice chromosomes (osa1, version 2.0) were used and repeat sequences were filtered by online RepeatMasker before BLASTZ alignment. For detailed genomic positions of the three duplicated blocks, see Table 1 and Figure S1.

had a little higher percentage than that of DNA transposons in all segments, except for chromosome 10 segment of block 3-10b. The 3-10b presented the most significant difference in size between its two syntenic segments, and repeat elements occupied nearly 40% in the chromosome 10 segment, and about 3/4 of repeat elements were due to LTR-retroelements (Table S3). The results indicated that LTR-retroelements played a particularly important role in DNA expansion in these regions, just as in the case of the maize genome (SanMiguel et al., 1996).

Genes produced by small-scale duplication after large-scale (genome) duplication were sought in whole genome range. A strict evolutionary distance (time) was set for the selection of the "new" genes relative to the genome duplication (for details, see Discussion). Apparently,

Table S3. Detailed list of masked repeats (%) in the three duplicated blocks using online RepeatMasker (http://www.repeatmasker.org).

Structures	Duplicated homologs					
	3-10b		3-7b		2-6b	
	Chr10	Chr3	Chr7	Chr3	Chr6	Chr2
Length (Mb)	4.95	0.82	7.2	3.44	9.28	4.56
Retroelements	26.04	8.35	12.72	7.65	12.88	9.1
SINEs	0.23	0.2	0.33	0.28	0.34	0.27
LINEs	0.23	0.01	0.31	0.12	0.14	0.24
Penelope	0	0	0	0	0	0
CRE/SLACS	0	0	0	0	0	0
L2/CR1/Rex	0	0	0	0	0	0
R1/LOA/Jockey	0	0	0	0	0	0
R2/R4/NeSL	0	0	0	0	0	0
RTE/Bov-B	0	0	0	0	0	0
L1/CIN4	0.23	0.01	0.31	0.12	0.14	0.24
LTR	25.58	8.15	12.08	7.25	12.4	8.59
BEL/Pao	0	0	0	0	0	0
Ty1/Copia	3.58	2.18	2.19	1.48	1.98	1.13
Gypsy/DIRS1	18.26	5.23	8.01	4.59	8.66	6.1
Retroviral	0	0	0	0	0	0
DNA transposons	9.64	9.67	9.95	8.29	11.28	8.67
hobo-Activator	0.5	0.68	0.45	0.16	0.42	0.43
Tc1-IS630-Pogo	1.55	1.94	2.16	2.4	2.01	1.73
En-Spm	3.09	1.4	2.1	0.78	2.77	1.3
MuDR-IS905	0.42	0.51	0.45	0.36	0.6	0.3
PiggyBac	0	0	0	0	0	0
Tourist/Harbinger	1.18	1.83	2	2.04	2.28	2.19
Other	0	0	0	0	0	0
Rolling-circles	0	0	0	0	0	0
Unclassified	0.12	0.05	0.06	0.08	0.1	0.04
Total	35.8	18.07	22.74	16.02	24.26	17.8
Small	0	0	0	0	0	0
Satellites	0	0	0	0	0	0
Simple	0.87	1	0.96	0.98	0.99	0.77
Low	0.82	0.83	0.88	0.92	1.08	0.72

Chr = chromosome.

although more new genes were in the long duplicate segments, gene densities were basically similar in all segments (Table 1 and Table S1). A similar distribution trend of tandemly duplicate gene arrays was also detected in the three duplicated blocks using the same method for *Arabidopsis* (Arabidopsis Genome Initiative, 2000) (data not shown). The results imply that new genes did not seem to be an evolutionary force causing genomic size in these blocks to vary.

An abundance of plastid and mitochondrial DNA insertions in nuclear genomes of rice has been reported (Notsu et al., 2002; Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003). The insertions would also cause an upward trend toward an enlarged genome. The results indicate that the insertions of organellar (chloroplast and mitochondria) genome fragments into the nuclear genome, including inter- or intro-chromosomal DNA sequence insertions, did not seem to be an important factor causing the size difference (Table 1).

DNA loss

After an exhaustive search, a large proportion of genomic sequences from the two duplicated segments in the three duplicated blocks was found to have no fundamental similarity to each other, even with repeat elements, etc. Here, these unique parts are regarded by the authors as the retained regions in two duplicated counterparts after the genome duplication event. In other words, they are remnants of the duplicated blocks. After an evolutionary period, the retained part in one duplicated segment from a duplicated block implies that this syntenic part was lost or deleted in its syntenic duplicated segment of the block.

Based on the above views, extensive DNA loss was clearly observed in three duplicated blocks (Table 1). The short syntenic segments appear to have lost most (over 80%) of their DNA sequence (such as chromosome 3 at block 3-10b), and even the most tenacious long segment lost 17.0% of its DNA sequences (chromosome 10 at 3-10b block) (Table 1). Two syntenic segments from a duplicated block showed significantly different rates of retained regions. Accordingly, long segments had more remnant sequences and lost less DNA sequences than short ones. The three short segments had lost most of their initial sequence lengths (62.6-81.1%), while their long partners lost less (17.0-32.9%). The results indicate that DNA loss had greatly contributed to the genome size differences of the blocks. DNA elimination often showed a mosaic pattern, i.e., short pieces were cut randomly from their genomic sequences (Figure 1).

Scenarios of chromosomal size evolution

Unbalanced DNA expansion and loss between syntenic segments of duplicated blocks caused a dramatic scenario of chromosomal size evolution: some of them gained more DNA and became enlarged, while others lost more and became shrunken. Based on the examination of the three duplicated blocks (Table 1), apparently, the duplicated segments from chromosomes 2 and 3 had experienced a constrictive process of genomic DNA sequence, and they shrank to almost half (46.0 and 49.0%) as much as their initial sizes from the whole genome duplication to now; segments from chromosomes 6 and 7 basically remained balanced with a slight increase, and the size of chromosome 10 increased substantially (69.2%). Meanwhile, a similar trend in gene number was observed in the three duplicated blocks (Table S1). The results show that incongruent patterns of genome size evolutions may have occurred for chromosomes of the rice genome. That is, some chromosomes seemed to have experienced a reduction process competing with amplification of some chromosomes, while others were likely to have regulated themselves to maintain a stable genome size over long evolutionary timescales. Moreover, the evolutionary scenario could be an ongoing process. For example, chromosomes 2 and 3 in rice are getting smaller whereas chromosome 10 is growing larger.

Nonhomologous recombination has been considered one of the key processes in the deletion of LTR-retrotransposons sequences in the rice genome, and a hallmark of nonhomologous recombination is the ratio of intact elements to solo LTRs (Ma et al., 2004). The average ratio of intact elements to solo LTRs in the three duplicated blocks was ~1:2, close to the previous results described in rice based on 1000 elements (~2:3, or 0.68; Ma et al., 2004). In the six duplicated segments, the ratios of intact elements to solo LTRs for short segments (0.65-1.0) were all greater than that of its long syntenic segments (0.28-0.42). Particularly, the chromosome 10 region of block 3-10b, which was suggested to have expanded dramatically (see above), had the lowest ratio (0.28) (Table 2). The results suggest that the forces slowing down the genome expansion through nonhomologous recombination in the duplicated segments seemed to be different, and that relatively weak forces were involved in the long segments, particularly in the chromosome 10 segment, which may account for the diversity of repetitive DNA content for the two duplicated segments in the three duplicated blocks.

Table 2. Structures of long terminal repeat retrotransposons in the three duplicated blocks of the rice genome.

	Number of elements in each class						
	Block 2-6b		Block 3-7b		Block 3-10b		Total
	Chr2	Chr6	Chr3	Chr7	Chr3	Chr10	
Intact elements	13	19	7	13	3	11	66
Intact elements without TSDs	0	1	0	2	1	7	11
Solo LTRs	20	45	10	25	3	40	143
Solo LTRs witout TSDs	1	23	0	15	0	6	45
At least one LTR partially/completely deleted	10	30	7	27	2	22	98
Ratio of intact elements to solo LTRs	0.65	0.42	0.70	0.52	1.00	0.28	0.46

LTR = long terminal repeat; TSDs = target site duplications; Chr = chromosome.

It has been reported that intron size contributes to genome size difference on a large evolutionary scale (Deutsch and Long, 1999; Vinogradov, 1999). However, a striking example contrary to the latter reports was witnessed in cotton (Grover et al., 2004). Various intron sizes of annotated genes by TIGR (excluding repeat genes) in the six duplicated segments were compared to determine their potential contribution to size variation. Larger than average intron sizes, which were observed in all three long duplicated segments relative to their short ones, resulted in a marginal gain of 44.8, 35.3 and 9.4 bp per intron in blocks 2-6b, 3-10b and 3-7b, respectively. This result suggests that there seemed to be a relationship between intron size variation and genomic size variation in duplicated segments in the rice genome.

DISCUSSION

Incongruent patterns of genome size evolution for chromosomes in rice

In the three distinct duplicated blocks, which resulted from a whole genome duplication event in the rice genome and covered over 20% of the length of the five corresponding chromosomes (chromosomes 2, 3, 6, 7, and 10), three different patterns of genome size evolution (shrunken, balanced and inflated) were observed (Figure 2). Chromosomes 2 and 3 were distinct from other chromosomes in their sequence length, the former two being significantly shortened (Pattern 1), while chromosome 10 was expanded substantially (Pattern 3). Based on our results, we propose for the first time that incongruent patterns of local genome size evolution among chromosomes or bidirectional evolution of chromosomal size gave rise to the modern rice genome. Wendel et al. (2002) suggested that DNA content increase and decrease occurred repeatedly during evolution based on the application of the phylogenetic method to genome size data. This observation helped shed light on the evolutionary mechanism of genomic sequences after polyploidy and genome size evolution.

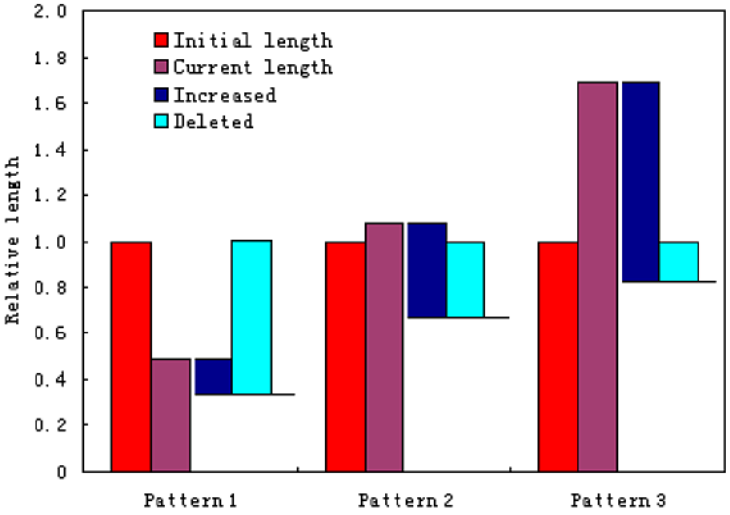

Figure 2. Three patterns (shrunken, balanced and inflated) of chromosomal size evolution in rice. Pattern 1 was averaged from the data of chromosomes 2 and 3, while Pattern 2 was from chromosomes 6 and 7, and Pattern 3 from chromosome 10 in Table 1. Active lengths were normalized by their initial sequence lengths.

It was suggested that the rice genome could be independent of genome size decrease as concluded by phylogenetic analysis (Bennetzen and Kellogg, 1997). Recently, Bennetzen and his colleagues reported that the genome sizes of both *indica* and *japonica* rice have increased over 2 and 6%, respectively, mainly because of the amplification of LTR-retrotransponsons within 0.5 million years following their divergence from a common ancestor (Ma and Bennetzen, 2004). The observation of upward-directional changes of genome size was based on about 1 Mb near the terminal genomic region of the long arm of rice chromosome 4. The percentage of repetitive DNA in this region was similar to that of the duplicated segments of chromosomes 6 and 7 (Feltus et al., 2004), which were shown to increase 9.2 and 7.4%, respectively, in this study.

Patterns of genome size evolution within the same chromosome may become incongruent in a similar way. Our conclusion on chromosomal size changes was based on regions of chromosomes. An exact evolutionary scenario in the size of a particular chromosome needs more comprehensive analysis of its entire region, just as in the case of concluding the direction of size change for multi-chromosomal genome. Dramatically different percentages of repetitive DNA have been observed within some chromosomes of rice (Feltus et al., 2004), such as chromosomes 4, 9 and 10. For example, chromosome 4, was found to have a very high percentage (32.5%) of LRT-retrotransponsons in its heterochromatic region, and low percentage (10.9%) in other regions. Its heterochromatic region occupies as much as about half of the chromosome, including its short arm and parts of its long arm (Feng et al., 2002).

Chromosomes 2 and 3 were suggested to have shrunken dramatically during the last ~70 million years, after the genome duplication of rice. In addition, the results of whole genome duplication analysis have shown that the two chromosomes were completely covered by non-overlapping duplicated blocks from at least two other chromosomes (Figure S1) (Paterson et al., 2004; Zhang et al., 2005). Whether there could be any links between the two phenomena is still an interesting question to be solved. One of the evolutionary processes could be that after genome duplication event, some large duplicate chromosomes split into two or three pieces because of the expansion of the genomic sequence. Some pieces might have evolved into modern chromosomes. In modern rice, the ancestral duplicated syntenic chromosomes of modern chromosomes 2 and 3 might have split into several segments during a certain evolutionary period, which might have further evolved into modern chromosomes 4, 6, 7, 10, etc.

DNA loss and amplification of LTR-retroelement

Our data indicate that unbalanced presentation in non-repetitive DNA loss and the amplification of repetitive DNA led to the incongruent pattern of genome size evolution of chromosomes in rice. Of the two competitive forces, DNA loss had exerted the primary effect on size changes. The question of what mechanisms controlled or balanced the DNA loss in the two duplicate segments remains open, although some mechanisms have been put forward for the deletion of repetitive DNA (Bennetzen, 2002; Petrov, 2002). In addition, the evolutionary fates of duplicate genes have also been well documented (Lynch and Conery, 2000; Betran and Long, 2002). Paterson et al. (2004) reported a high level (21.4%) of rice genes retaining syntenic homologs found in duplicated blocks detected from version 1.0 (osa1) data of rice. In this study, we also identified a similar level in some duplicated regions based on current data (osa1, version 2.0) (Table S1). Will there be a bright future for the duplicated genes in rice genome? The answer could be no if the aspect of lost genes concerning initial sequence lengths or gene number of duplicate segments is taken into consideration. In this study, massive DNA loss was found after the genome duplication during the evolutionary process. In other words, the current duplicated blocks should be longer than they are. Thus, the percentage of retained genes should have decreased to the relatively low level (3.2-9.5%) at present, when considering the initial number of genes in the duplicated region (Table S1). More detailed research is needed to confirm this estimation.

Rice, as a model crop, has the smallest genome in the grass family. All members of the family originated from a common ancestor, which shared a whole genome duplication predating their divergence. Why and how the rice genome became the smallest one while others hold 2- or 4-fold larger genomes (such as maize) is unclear, which needs further investigation. Comparative analysis has shown that rice is more genomically stable than two other cereals, i.e., maize and sorghum (Ilic et al., 2003). Our results could provide some clues to understanding this disparity. The rice genome seemed to be subject to reduction or stability in its size after speciation, based on the five chromosomes surveyed in this study. Of the five chromosomes, only chromosome 10 expanded significantly whereas the others decreased or remained stable in size. Meanwhile, transposable elements (LTR-retrotransponsons, etc.) in rice do not display the same density as in maize (Bennetzen, 2002), except for some particular regions, such as part of chromosomes 10 and 4. This implies that transposable elements, the largest variable force relating to genome size in some plant genomes such as maize, wheat and barley seemed to show a relatively weak effect in rice.

Identification of new genes appearing after the whole genome duplication

Identifying new genes produced after the whole genome duplication was one of the most difficult tasks in this study. Evolutionary distance (amino acid substitution rate, d_A <0.2) was used to identify newborn genes after the whole genome duplication from others. The distribution peak of d_A values of duplicated gene pairs in all 9 duplicated blocks of the whole genome duplication was 0.35-0.40, and most d_A values of duplicated gene pairs from the large-scale duplication event were over 0.2 (Zhang et al., 2005). Therefore, the evolutionary distance is very balanced and conservative. Meanwhile, we set an identity for searching members belonging to a particular gene family. In this study, 50% identity was used as the lower threshold to identify members of a gene family. However, some factors often provide overestimation or underestimation of the number of new genes, such as the unfinished genome sequence of rice, which will narrow the size of the target dataset for searching new genes, although not by much (Yuan et al., 2003). It may result in the underestimation of the number of new genes.

ACKNOWLEDGMENTS

We thank Profs. Mingwei Gao and Mingsheng Chen for their critical reading of the manuscript. Research supported by grants from the National Natural Science Foundation of China (Nos. 30270810, 90208022 and 30471067) and IBM Shared University Research (Life Science).

REFERENCES

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

Bennett MD and Leitch IJ (1997). Nuclear DNA amount in angiosperms. *Philos*. *Trans*. *Royal Soc*. *Lond*. *B Biol*. *Sci*. 334: 309-345.

Bennetzen JL (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115: 29-36.

Bennetzen JL and Kellogg EA (1997). Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9: 1509-1514.

Betran E and Long M (2002). Expansion of genome coding regions by acquisition of new genes. *Genetica* 115: 65-80.

Delcher AL, Phillippy A, Carlton J and Salzberg SL (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. 30: 2478-2483.

Deutsch M and Long M (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*. 27: 3219-3228.

Devos KM, Brown JKM and Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res*. 12: 1075-1079.

Eckardt NA (2001). A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell* 13: 1699-1704.

Feltus FA, Wan J, Schulze SR, Estill JC, et al. (2004). An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res*. 14: 1812-1819.

Feng Q, Zhang Y, Hao P, Wang S, et al. (2002). Sequence and analysis of rice chromosome 4. *Nature* 420: 316-320.

Grover CE, Kim H, Wing RA, Paterson AH, et al. (2004). Incongruent patterns of local and global genome size evolution in cotton. *Genome Res*. 14: 1474-1482.

Guyot R and Keller B (2004). Ancestral genome duplication in rice. *Genome* 47: 610-614.

Ilic K, SanMiguel PJ and Bennetzen JL (2003). A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 100: 12265-12270.

Kent WJ, Baertsch R, Hinrichs A, Miller W, et al. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 100: 11484-11489.

Lynch M and Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155.

Ma J and Bennetzen JL (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 101: 12404-12410.

Ma J, Devos KM and Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res*. 14: 860-869.

Notsu Y, Masood S, Nishikawa T, Kubo N, et al. (2002). The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol*. *Genet*. *Genomics* 268: 434-445.

Ohno S (1970). Evolution by gene duplication. George Allen and Unwin, London, England.

Paterson AH, Bowers JE and Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 101: 9903-9908.

Petrov DA (2002). DNA loss and evolution of genome size in *Drosophila*. *Genetica* 115: 81-91.

Petrov DA, Lozovskaya ER and Hartl DL (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384: 346-349.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, et al. (2000). Evidence for DNA loss as a determinant of genome size. *Science* 287: 1060-1062.

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* 428: 475-476.

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

SanMiguel P and Bennetzen JL (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann*. *Bot*. 82: 37-44.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765-768.

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, et al. (2002). The genome sequence and structure of rice chromosome 1. *Nature* 420: 312-316.

Schwartz S, Kent WJ, Smit A, Zhang Z, et al. (2003). Human-mouse alignments with BLASTZ. *Genome Res*. 13: 103-107.

Shirasu K, Schulman AH, Lahaye T and Schulze-Lefert P (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res*. 10: 908-915.

Thomas Jr CA (1971). The genetic organization of chromosomes. *Annu*. *Rev*. *Genet*. 5: 237-256.

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, et al. (1999). Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell* 11: 1769-1784.

Vinogradov AE (1999). Intron-genome size relationship on a large evolutionary scale. *J*. *Mol*. *Evol*. 49: 376-384.

Vitte C and Panaud O (2003). Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol*. *Biol*. *Evol*. 20: 528-540.

Wang X, Shi X, Hao B, Ge S, et al. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol*. 165: 937-946.

Wendel JF (2000). Genome evolution in polyploids. *Plant Mol*. *Biol*. 42: 225-249.

Wendel JF, Cronn RC, Johnston JS and Price HJ (2002). Feast and famine in plant genomes. *Genetica* 115: 37-47.

Wicker T, Stein N, Albar L, Feuillet C, et al. (2001). Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J*. 26: 307-316.

Yang Z (1999). Phylogenetic analysis by maximum likelihood (PAML). University College, London, UK. http://abacus.gene.ucl.ac.uk/software/paml.html.

Yu J, Wang J, Lin W, Li S, et al. (2005). The genomes of *Oryza sativa*: a history of duplications. *PLoS. Biol*. 3: e38.

Yuan Q, Ouyang S, Liu J, Suh B, et al. (2003). The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res*. 31: 229-233.

Zhang Y, Xu G, Guo X and Fan L (2005). Two ancient rounds of polyploidy in rice genome. *J. Zhejiang Univ*. 6B: 87-90.