Funct Integr Genomics (2008) 8:101–108
DOI 10.1007/s10142-007-0067-2

ORIGINAL PAPER

Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved

Xingyi Guo · Songlin Ruan · Weiming Hu · Daguang Cai · Longjiang Fan

Received: 12 July 2007 / Revised: 15 October 2007 / Accepted: 15 October 2007 / Published online: 10 November 2007

Abstract Rice (*Oryza sativa*) is one of three predominant grain crops, and its nuclear and organelle genomes have been sequenced. Following genome analysis revealed many exchanges of DNA sequences between the nuclear and organelle genomes. In this study, a total of 45 chloroplast DNA insertions more than 2 kb in length were detected in rice nuclear genome. A homologous recombination mechanism is expected for those chloroplast insertions with high similarity between their flanking sequences. Only five chloroplast insertions with high sequence similarity between two flanking sequences from an insertion were found in the 45 insertions, suggesting that rice might follow the non-homologous end-joining (NHEJ) repair of double-stranded breaks mechanism, which is suggested to be common to all eukaryotes. Our studies indicate that the most chloroplast insertions occurred at a nuclear region characterized by a sharp change of repetitive sequence density. One potential explanation is that regions such as this might be susceptible target sites or "hotspots" of DNA damage. Our results also suggest that the insertion of retrotransposon elements or non-chloroplast DNA into chloroplast DNA insertions may contribute significantly to their fragmentation process. Moreover, based on chloroplast insertions in nuclear genomes of two subspecies (*indica* and *japonica*) of cultivated rice, our results strongly suggest that they diverged during 0.06–0.22 million years ago.

Keywords *Oryza sativa* · Chloroplast DNA insertion · Direct-repeat · Retrotransposon · *Indica* · *Japonica* divergence

Introduction

Rice (*Oryza sativa* L.) belongs to the grass family (Gramineae or Poaceae) and is one of three predominant grain crops—rice, wheat (*Triticum aestivum*), and maize (*Zea mays*)—that provide a major food source for humans. The grass family diverged from a common ancestor about 77 million years ago (mya), and the divergence between its subfamily Erhartoideae (rice) and Panicoideae (maize and sorghum) is estimated at 50 mya (Gaut 2002). The rice genus, *Oryza*, comprises approximately 23 species. Nine genome types of diploid ($2n=24$) and various combinations among them to allotetraploid ($2n=48$) have been recognized for 22 *Oryza* species (Ge et al. 1999; Vaughan et al. 2003). *O. rufigogon*, an Asian common wild rice that shows a range of variation from perennial to annual types, is the wild progenitor of cultivated rice (*O. sativa* L., AA genome). Cultivated rice has two main subspecies, *indica* and *japonica* subspecies (Khush 1997), which were estimated to separate about 0.05–0.44 mya (Ma and Bennetzen 2004; Tian et al. 2006).

Electronic supplementary material The online version of this article (doi:10.1007/s10142-007-0067-2) contains supplementary material, which is available to authorized users.

X. Guo · W. Hu · L. Fan (✉)
Institute of Crop Science, Zhejiang University,
Hangzhou 310029, China
e-mail: fanlj@zju.edu.cn

S. Ruan
Institute of Biotechnology,
Hangzhou Academy of Agricultural Science,
Hangzhou 310024, China

D. Cai
Plant Breeding Institute of Christian-Albrechts-University Kiel,
Olshausenstr. 49,
24118 Kiel, Germany

Organelle DNA transfer to the nucleus is ubiquitous and ongoing. Transfer of certain mitochondrial and plastid functional genes has occurred frequently in flowering plants in evolutionarily recent times, although the transfers of functional genes is now rare or has ceased in animals and many other eukaryotes (see review by Leister 2005). The high-frequency gene transfer from chloroplast genome to the nucleus has been experimentally demonstrated (Huang et al. 2003; Stegemann et al. 2003). Large, even whole plastid or mitochondrial genome, insertions have been detected in nuclear genomes (e.g., Yuan et al. 2002; Shahmuradov et al. 2003; Huang et al. 2005). Different genomic organization of organelle insertions, so-called continuous, rearranged, and mosaics, have been observed in nuclear genomes, and mechanisms or models of nuclear insertion of organelle DNA (such as simple end-joining and insertion and end-joining) have been proposed (Leister 2005). However, several outstanding questions still remain: (a) the nature of old insertions with short and/or diverged sequences; (b) the non-homologous end-joining repair (illegitimate repair, NHEJ, Leister 2005).

O. sativa ssp *japonica* cultivar Nipponbare has been sequenced by the integration of whole-genome shortgun (Goff et al. 2002) and extensive genetic and physical mapping efforts or "clone-by-clone" (Sasaki et al. 2002; Feng et al. 2002; The Rice Chromosome 10 Sequencing Consortium 2003; International Rice Genome Sequencing Project 2005). The draft genome sequence of an *indica* cultivar "93-11" was also generated by whole-genome shortgun (Yu et al. 2002, 2005). Besides nuclear genome, the chloroplast genome of rice was completed 15 years ago (Hiratsuka et al. 1989). The analysis of the available genome sequences revealed that many exchanges of DNA sequences among the nuclear, chloroplast, and mitochondrial genomes have occurred in rice (Notsu et al. 2002; Yuan et al. 2002; Shahmuradov et al. 2003; The Rice Chromosome 10 Sequencing Consortium 2003). The above genome sequences and large organelle DNA insertions, combining some new developed bioinformatic methods (such as GRIMM), provide an opportunity to further investigate the genes, sites, and ages involved chloroplast DNA insertions into the nuclear genome of rice.

Materials and methods

Sequence source and plant materials

Nuclear and chloroplast genome sequences of rice were downloaded from the Institute of Genome Research (TIGR) at http://www.tigr.org/tdb/e2k1/osa1/ (Release 5.0), the BGI-RIS of Beijing Genomics Institute at http://rise.genomics.org.cn/rice2/index.jsp, and GenBank (X15901) at http://www.ncbi.nlm.nih.gov, respectively. Sequence data from this article have been deposited to GenBank under accession numbers DQ973115–DQ973122.

Seven typical *indica* cultivars (Nanjing11, IR64, IR 36, Nantehao, Guangluai4, Teqing and 93-11), together with five typical *japonica* cultivars (Nipponbare, Akihikari, Youmangshajing, Balilla and Zhonghua11), were provided by China National Rice Research Institute.

Detection of chloroplast DNA insertions in nuclear genome

The sequences of 12 chromosomes of rice were aligned against its chloroplast genome using nucmer program of MUMmer package (Delcher et al. 2002) with 50-bp minimal length of exact match, respectively. Unusually, chloroplast DNA insertions were further inserted by short non-chloroplast sequences. In this case, the length of the non-chloroplast DNA sequence in an insertion was limited to a conservative standard (<400 bp; Huang et al. 2005). Otherwise, it was as two chloroplast insertions.

Similarity and annotation of flanking sequence

Similarities for two flanking nuclear sequences (5′ and 3′ end) of the chloroplast insertions were calculated based on sequence alignment using nucmer program of MUMmer package (Delcher et al. 2002) with 10-bp minimal length of exact match. Repeat elements were searched against repeat database (RepeatMasker.Lib) using RepeatMasker program. GO annotations were performed on the web server at TIGR (http://www.tigr.org/tdb/e2k1/osa1/batch_download.shtml). Based on GO hierachical class (http://www.geneontology.org/), molecular function and biological process were assigned to TIGR annotated genes at flanking nuclear sequences of chloroplast DNA insertions.

Rearrangement of chloroplast DNA insertions

Rearrangement scenario of chloroplast DNA insertion was constructed based on the Hannenhalli–Pevzner algorithms (Hannenhalli and Pevzner 1995) implemented in GRIMM web server (http://nbcr.sdsc.edu/GRIMM/grimm.cgi, Genome rearrangement algorithms).

Determination of common or subspecies-specific chloroplast insertion

Two methods were used to determine common or subspecies-specific insertions of chloroplast DNA for *indica* and *japonica* rice genome: (1) polymerase chain reaction (PCR) detection: Based on the sequences of nuclear–chloroplast junctions at two ends of those chloroplast insertions in *japonica* Nipponbare, seven representative *indica* cultivars

were detected by PCR amplification. Five *japonica* cultivars were used as controls. The reaction mixture (25 μl) consisted of 1× buffer, 100 μM dNTP, 0.4 μM of primers, 1 U of Taq plus enzyme from DingGuo company, and thermal cycled in an PTC-200 (MJ research) thermal cycler for 30 cycles for 1 min at 94°C, 1 min at 63°C, and 1 min at 72°C. The primer sequences and expected product length were listed in Table S1. As a quality control, all primers designed must represent positive PCR results for Nipponbare. Particularly, PCR products of three key insertions, insertion chr10-2, chr4-2, and chr2-2, in *indica* cultivar group were confirmed by sequencing: PCR products were recovered with glass milk kit (BioDev Company, China). The recovered fragments were ligated into PMD18-T vector and transformed into competent cells of *Escherichia coli* DH5α (TaKaRa Company). The positive clones were sequenced using ABI 3700 (ABI Company). Sequence data have been deposited to GenBank under accession numbers DQ973115–DQ973122. (2) Similarity search against *indica* genome: The sequences of chloroplast DNA insertions in *japonica* Nipponbare genome, including their two flanking nuclear sequences (2 kb in length, respectively), were searched against the nuclear genome of *indica* cultivar 93-11 using BLASTN.

Estimation of synonymous substitution rates

The average number of synonymous (d_S) per site of paralogous genes in a chloroplast DNA insertion and its corresponding chloroplast genomic region were calculated using yn00 program of PAML package with Yang's model (Yang 1999). The corresponding region (paralogous segment) of chloroplast genome (X15901) to an insertion was identified first using nucmer program, and then the paralogous coding sequences (the best hit with more than 95% length coverage to the corresponding cds in X15901) in the region were identified by BLASTN searching.

Results

Genes and repeat elements flanking chloroplast DNA insertions

Abundant chloroplast DNA insertions were found in currently assembled nuclear genome of *japonica* cultivar Nipponbare (Fig. 1). A total of 45 insertions more than 2 kb in length, of which five more than 40 kb, including a 131-kb insertion (almost equal to whole chloroplast genome) in chromosome 10, were detected. Some insertions clustered and were close to each other within 15 kb, suggesting that they might arise a similar insertion event and following nuclear DNA reinserting into the chloroplast DNA insertions (see next section). Of the 45 insertions, 24 do not have any neighbor chloroplast insertions within 15 kb.

To determine direct-repeat and other functional repeat elements flanking the chloroplast insertions, the above 24 chloroplast DNA insertions were used (for a detailed list of the 24 insertions, see Table S1). In Fig. 2, two flanking sequences from an insertion were ranked one by one in *X* and *Y* axes, respectively, and therefore, the dots in the boxes at diagonal line indicated the similarity between the two flanking sequences of the chloroplast insertion, while others represented the similarity between two flanking sequences

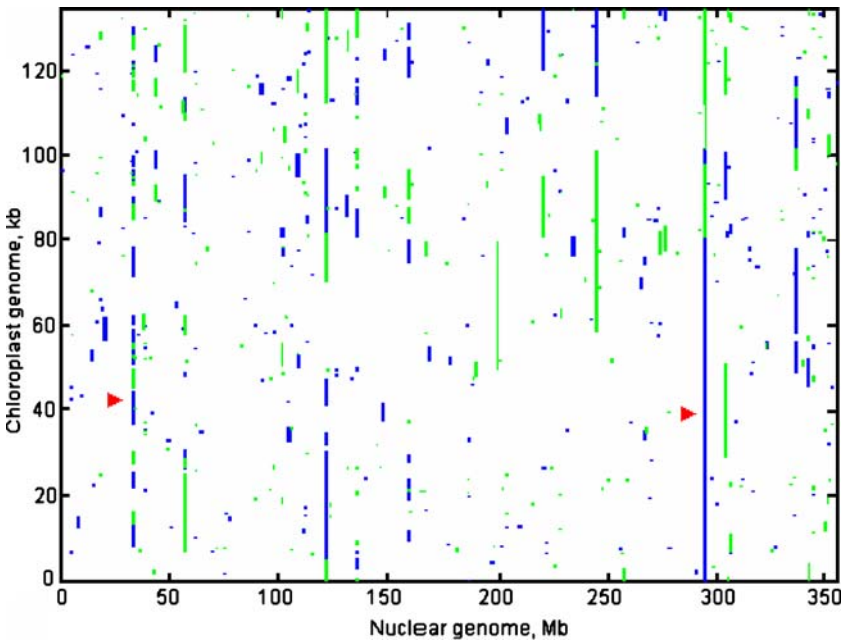

Fig. 1 Chloroplast DNA insertions in rice nuclear genome were detected by whole-genome alignment. Dot matrix analysis of chloroplast and nuclear genome performed using MUMmer. Nuclear chromosomes 1 ~ 12 were arranged from left to right one by one, and two whole plastid genome insertions at chromosome 1 and 10 are indicated by *arrow*. The *dark* (*blue*) and *light* (*green*) lines refer positive and reverse matches, respectively

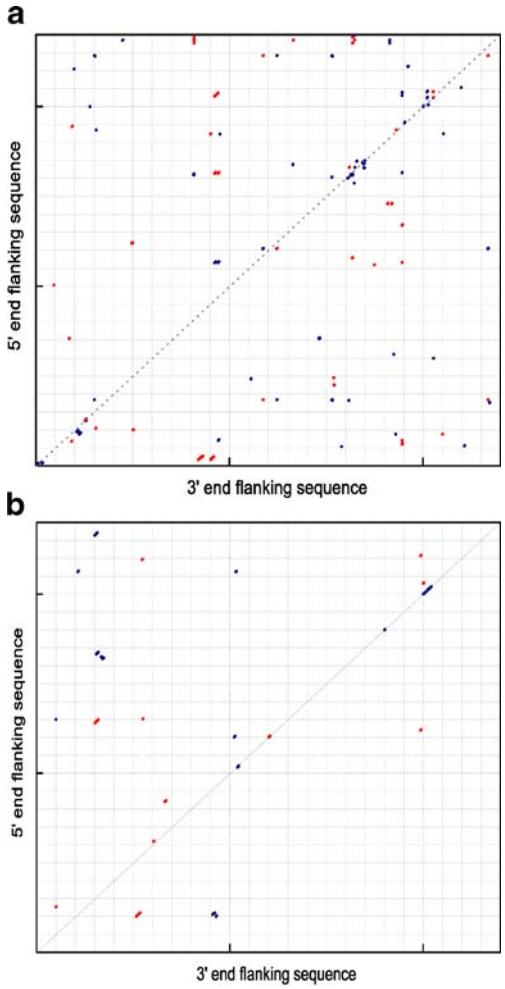

Fig. 2 Dot matrix of two flanking nuclear sequences of 24 chloroplast DNA insertions in rice. Each partition presents a 50 kb (**a**) or 5 kb (**b**) flanking sequence from an insertion. The *blue* and *red lines* refer positive and reverse matches, respectively

from two different insertions. The data in Fig. 2a suggest that seven insertions analyzed had direct- or reverse-repeats within their flanking sequences, as indicated by dots and lines on the diagonal in the comparison between 5′ and 3′ sequences. In a further investigation within 5 kb flanking sequences (Fig. 2b), only five insertions have repeats in both their flanking sequences, including an insertion in chromosome 10 with 2.1-kb direct-repeats at its two flanking sequences reported by Yuan et al. (2002). The results indicated that only a number (5/24) of chloroplast DNA insertions contain sequence similarity (potential direct-repeats) in both flanking nuclear sequences from a chloroplast insertion. Moreover, many dots or lines were observed in the cells that were not at the diagonal (Fig. 2); particularly, some were detected in several flanking sequences, suggesting that repeat elements might enrich in those flanking nuclear sequences.

Further investigations on repeat density in the flanking 50-kb regions suggested an interesting pattern of flanking sequences at both directions (Fig. 3). Apparent fluctuations of repeat (most are retrotranposons) densities across the flanking sequences at both ends were observed. Interestingly, repeat density within 20 kb from insertion point changed sharply from higher one at 5′ end to lower one at 3′ end regions. The results suggest that most chloroplast DNA insertions potentially occurred at those regions characterized by a sharp change of repetitive sequence density.

To find potential genes involved in chloroplast DNA insertions, TIGR annotated genes within 50-kb flanking nuclear sequences at both directions were collected. Functional category results showed that in a total of 340 annotated genes in flanking sequences of the 24 chloroplast insertions, genes involved in catalytic and transferase activity (18.4 and 15.2%, respectively; Fig. 4a), or biological processes of cellular process, response to abiotic stimulus and response to stress (10.3, 9.2 and 8.8%, respectively; Fig. 4b) were over-represented, which were different to the situations based on annotated genes from whole genome (International Rice Genome Sequencing Project 2005; Yu et al. 2005). The results provided some clues for identification of putative genes (e.g., the genes for catalytic and transferase activity) involved in plastid DNA insertions.

Rearrangement of chloroplast DNA insertions

A large insertion in chromosome 1 at 33.8–33.9 Mb has been suggested from a whole plastid genome insertion event, which later was subject to many potential duplication/deletion/inversion/translocation rearrangements (Shahmuradov et al. 2003). To date, it appears to be fractioned and disordered (Fig. 1). How to rearrange or what functional elements involved in mutational decay of a chloroplast DNA segment after it inserted into nuclear genome is always an interesting topic.

The chloroplast insertion could be defined as 28 pieces or blocks based on similarity with chloroplast genome

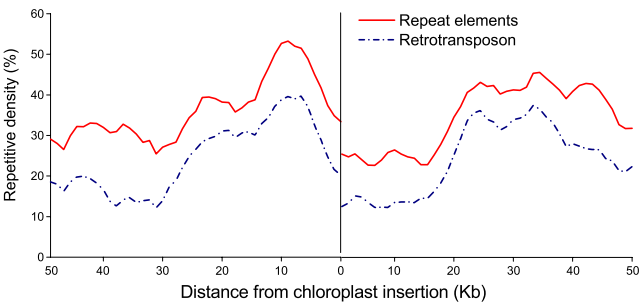

Fig. 3 Contents of repeat elements (*line*) and retrotransposons (*dash dot*) as function of positions of the nuclear DNA sequences flanking chloroplast insertions, relative to the inserting sites (position "0"). Left and right regions relative to the inserting position are 5′ and 3′ flanking sequences, respectively

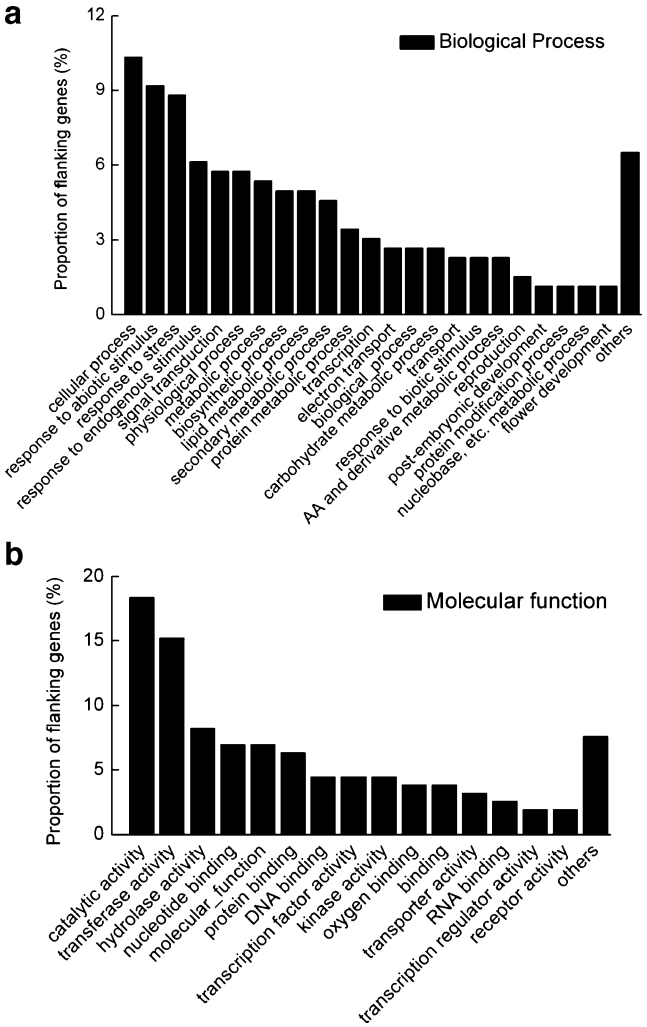

Fig. 4 GO functional classifications of 340 annotated genes in flanking nuclear sequences of 24 chloroplast DNA insertions in rice

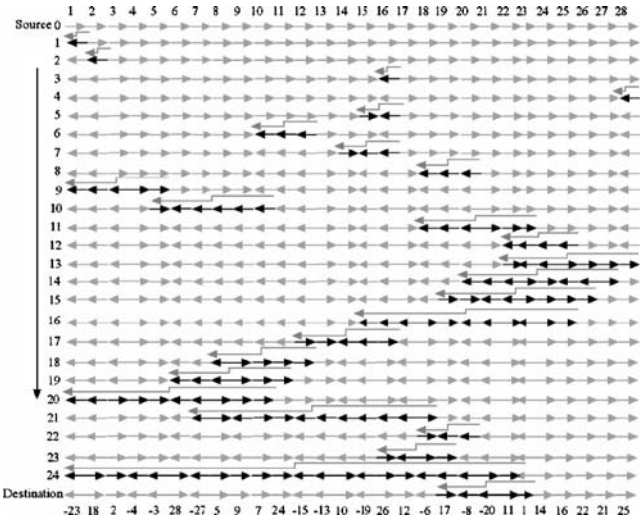

Fig. 5 A most parsimonious rearrangement scenario for a whole plastid genome insertion in rice chromosome 1 (see Fig. 1) (destination) and chloroplast genome (source). *Arrows* refer to 28 segments or blocks of the chloroplast DNA insertion, and the *Z arrow* indicates a reversion event of which block(s) in *dark arrow(s)* involved

using the Hannenhalli–Pevzner algorithms (Hannenhalli and Pevzner 1995). The method has been used to estimate genome rearrangements after divergence of species, such as human and mouse (Pevzner and Tesler 2003), but not in analysis of chloroplast insertions. We estimated a most parsimonious scenario of these 28 segments from the original order (i.e., plastid genome insertion) to the current insertion via inversion (Fig. 5). A total of 25 steps are needed to finish this dynamic process. Using coding sequences in this insertion, it was estimated to insert into the nuclear genome about 0.4 mya (for details, see next section). Moreover, several non-chloroplast DNA segments could be detected within the chloroplast DNA insertion, suggesting that these non-chloroplast DNAs reinserted the chloroplast insertion after it inserted into nuclear genome. We identified eight non-chloroplast DNA insertions from this chloroplast DNA insertion, and their average size is 4,009 bp (see supplementary Table S1). Then, we used RepeatMasker to further determine what kinds of sequences (e.g., retrotransposon or other elements) comprise those reinserting sequences. As expected, repeat elements were detected in seven of the eight insertions. Excluding the low-complexity repeats, Gypsy type LTR and RNA (LSU rRNA and tRNA) were only two other types of repeat elements in the seven insertions. For example, one insertion (no. 7) includes three LTR elements with a summing 2.8 kb in length, which comprise 90.4% of the non-chloroplast insertion (Table S1). The results suggest that retrotransposons may be the most active element involved in mutational decay of chloroplast DNA insertions in nuclear genome.

Timing the *indica–japonica* divergence based on chloroplast DNA insertions

Chloroplast DNA insertions in *indica* and *japonica* nuclear genomes provide an opportunity to time the *indica–japonica* divergence. The divergent time can be anchored through determining common and subspecies-specific insertions, suggesting that the insertions occurred before and after the *indica–japonica* divergence, respectively, and their insertion times. To test numerous confused time estimations (see "Discussion"), the *indica–japonica* divergence was estimated using this distinct method in this study.

Of the above 45 insertions, eight insertions of recent 1.0 (0.04–0.81) mya estimated by the average synonymous substitutions (d_S values) of coding sequences between the insertions and their paralogs from chloroplast genome, around the putative time (0.05–0.44 mya) of *indica–japonica* divergence suggested by previous studies, were selected for our estimation. Two methods are used to

determine whether the above eight *japonica* insertions can also be detected in the *indica* genome, which will determine the common in both subspecies, or *japonica*-specific insertions for the eight insertions. First, the eight insertions and their two flanking nuclear sequences were used as queries to be searched against the draft genome sequences of *indica* cultivar 93-11. The results indicated that four *indica* orthologs were returned for four *japonica* insertions including their franking sequences, suggesting that the four insertions were common for both subspecies, or the pre-divergence insertions. Another four insertions failed to be retrieved from the current *indica* genomic sequence database (Table 1), suggesting that the four insertions may be absent in the *indica* genome. But the absence needs to be further confirmed regarding to the current coverage of the *indica* draft sequence. Second, PCR amplifications were performed to confirm the above genome searching results. Fourteen insertion-specific primer combinations were designed based on the nuclear-chloroplast junction sequences of the ends of the eight chloroplast insertions, and seven representative *indica* cultivars, together with five *japonica* cultivars, as controls were analyzed. As expected, the 14 primer pairs could be amplified from the *japonica* genome (Table 1). However, only four insertions, chr4-6, chr5-3, chr7-3, and chr10-2, could be amplified from the *indica* genome, supporting the *indica* 93-11 database searching results, and insertions of chr7-2 and chr4-1 failed to be amplified from any *indica* cultivars, which was also consistent with the above searching results. Particularly, PCR products of several anchor insertions for estimation of *indica–japinica* divergence, chr10-2, chr4-2, and chr2-2, were confirmed by sequencing (GenBank accession nos.: DQ973115-DQ973122). In summary, in the eight *joponica* insertions, six insertions were found in the *indica* genome, suggesting that they were inserted before the *indica–japinica* divergence, while other two insertions (chr7-2 and chr4-1) were not found in *indica* genome and represented the post-divergence or *japonica*-specific insertion events.

d_S values of the eight insertions varied from 0.0004 to 0.0063 (Table 1). Insertion dates of the chloroplast insertion were estimated based on their d_S values with an assumption of molecular clock: $T = d_s/(r_{cp} + r_{nu})$ where chloroplast substitute rate $(r_{cp}) = 1.24 - 2.9 \times 10^{-9}$ (Muse 2000) and nuclear substitute rate $(r_{nu}) = 6.5 \times 10^{-9}$ (Gaut et al. 1996). Based on the total number of nucleotide substitutions, the eight insertions were dated 0.04 to 0.81 mya, assuming a nuclear DNA substitute rate for the chloroplast insertions in nuclear genome (Table 1). Together, the above detection results on common and *japonica*-specific insertions; the period of *indica–japonica* divergence can be determined based the eight chloroplast insertions. First, the above results indicate that chr10-2 insertion was the most recent common insertion for *indica* and *japonica* rice in the six pre-divergence insertions (chr10-2, chr4-2, chr2-2, chr4-6, chr5-3, and chr7-3). The insertion represents the most recent chloroplast DNA insertion events in the rice nuclear genome before the *indica–japonica* divergence in

Table 1 Timing *indica–japonica* divergence based on common and subspecies-specific chloroplast DNA insertions

Insertions / genome	Length (kb)	cds number[a]	Substitution rate d_S	Date (mya)[b]	*Indica* search[c]	PCR detection[d]										
						japonica Cultivars					*indica* Cultivars					
						1	2	3	4	5	1	2	3	4	5	6	7
Chr7-2	14.4	10	0.0004	0.04–0.05	ND	+	+	−	+	−	−	−	−	−	−	−	−
Chr4-1	14.4	7	0.0006	0.06–0.08	ND	+	+	+	−	−	−	−	−	−	−	−	−
Chr10-2	33.0	16	0.0014	0.15–0.18	I	+	+	+	−	+	−	+	−	−	−	−	+
Chr4-2	61.9	31	0.0017	0.18–0.22	ND	+	+	+	+	+	+	−	+	+	−	+	−
Chr2-2	49.7	25	0.0018	0.19–0.23	ND	+	+	+	+	+	+	+	−	+	−	+	+
Chr4-6	7.1	5	0.0026	0.28–0.34	I	+	+	+	+	+	+	+	+	+	−	−	+
Chr5-3	11.1	5	0.0051	0.54–0.66	I	+	+	+	+	+	+	+	+	−	+	+	+
Chr7-3	4.6	5	0.0063	0.67–0.81	I	+	+	−	−	+	+	−	−	−	+	−	+

[a] The number of predicted coding sequences (cds)

[b] Insertion dates of the chloroplast insertion were estimated based on their d_S values with an assumption of molecular clock: $T = d_s/(r_{cp} + r_{nt})$ where chloroplast substitute rate $(r_{cp}) = 1.24 - 2.9 \times 10^{-9}$ (Muse 2000) and nuclear substitute rate $(r_{nt}) = 6.5 \times 10^{-9}$ (Gaut et al. 1996);

[c] When both nuclear junction sequences of a chloroplast DNA insertion were perfectly matched by a intact *indica* genomic sequence based on the alignment results against the nuclear genome of *indica* cultivar 93-11, two situations maybe existed at the corresponding insertion region of the *indica* sequence region: chloroplast sequence were found (I) and no genomic sequence data of 93-11 could be found (ND);

[c] + refer to positive (at least one end) or negative PCR results that were obtained for two junction ends of an insertion, respectively. *japonica* cultivars 1–5: Nipponbare, Akihikari, Balilla, Youmangshajing, Zhonghua11; *indica* cultivars 1–7: Nanjing11, IR64, IR36, Nantehao, Guangluai4, Teqing and 93-11

[d] Average synonymous substitutions per site (d_S) or insertion times, PCR amplification results in genomes of *indica* cultivar group and genome-scale search results against *indica* genome (93-11) of the selected insertions were listed. The insertions were sorted by their d_S values.

current data set. d_S values of insertion chr10-2 is 0.0014, corresponding to the time of 0.15–0.18 mya, suggesting that *indica* and *japonica* rice did not separate until 0.15–0.18 mya. The insertion chr4-2, a 0.18–0.22 mya insertions found in at least four *indica* cultivars, might provided a more solid estimation in terms of potential *japonica* introgressive hybridizations in *indica* 93-11 and IR64. In other words, their divergence must have occurred within 0.22 mya. Second, the insertion of chr4-1 was the most ancient insertion in the *japonica*-specific insertions. Its d_S value indicated that it should have inserted into the *japonica* genome 0.06–0.08 mya. The results implied that *indica* and *japonica* subspecies have separated at least 0.06 mya. Taking together the above two threshold time estimates, our results strongly suggested that *indica* and *japonica* subspecies should have separated between 0.06 and 0.22 mya (Fig. S1).

Discussion

One mechanism that organelle DNA can integrate into the nuclear genome by NHEJ repair (illegitimate repair) of double-stranded breaks (DSBs) has been found in yeast (Yu and Gabriel 1999; Ricchetti et al. 1999). Generally, DSBs can be repaired via two different ways, either via homologous recombination or via NHEJ (also known as illegitimate recombination). Whereas in the former way the sequences are linked in regions that are identical to each other, in the latter, the sequence information does not play a major role in the rejoining of the two double strands. Repair of DSBs by NHEJ requires little or no sequence homology between the termini, enabling organelle DNA to be "pasted" to one another. Several studies have indicated that the NHEJ might be a phenomenon common to all eukaryotes (Leister 2005). Only five chloroplast insertions with high sequence similarity between two flanking sequences from an insertion were found in our study, suggesting that rice maybe also follow this mechanism. Moreover, according to the model, any increase in the frequency of formation of DSBs should influence the rate of organelle DNA insertions generation. Our studies indicate that the most chloroplast insertions occurred at a region characterized by a sharp change of repetitive sequence density (Fig. 3). One potential explanation is that such regions might be susceptible target sites or "hotspots" of DNA damage in both strands. The observation in this study that genes involved in catalytic and transferase activity are over-represented in the nuclear DNA flanking chloroplast insertions needs further investigation, as they may relate to the repair mechanisms that operate as part of the insertion process.

We have detected many non-chloroplast DNA segments (most are LTR retrotransposons) reinserting into chloroplast DNA insertions. The phenomenon is also found in human genome that repetitive sequences, in particular to mobile elements, re-localize mitochondrial insertions (Mishmar et al. 2004). It suggests that the insertion of non-chloroplast DNA, including transposable elements, into chloroplast DNA insertions may contribute significantly to fragmentation process (i.e., rearrangement) of chloroplast DNA insertions.

Our results provided evidence for recent divergence (0.06–0.22 mya) of the two subspecies (*indica* and *japonica*) of rice, which is similar to the estimated time of 0.086–0.200 mya based on chloroplast genome comparisons (Tang et al. 2004) 0.045–0.250 mya based on mitochondrial genome comparisons (Tian et al. 2006), and compatible with the estimation of 0.44 mya based on nuclear gene data by Ma and Bennetzen (2004) and 0.4 mya based on intronic sequences of nuclear genes by Zhu and Ge (2005). Another study has suggested that the two genomes diverged from one another at least 0.2 mya based on LTR retrotransposons (Vitte et al. 2004). Huang et al. (2005) documented a 131-kb chloroplast fragment which was believed to be transferred about 0.148 (0.074–0.296) mya to the nuclear genome of *japonica* after it diverged from *indica*. In our study, a clear and narrowed divergent period was shown for the two subspecies based on an independent method (chloroplast insertion). Because of several advantages of chloroplast DNA, such as maternal inheritance, rare or no recombination, etc., chloroplast DNA had been successfully used in the estimation of divergent dates of plants (Wolfe et al. 1989; Gaut 2002; Sall et al. 2003). The total number or average of nucleotide substitutions in an insertion and a molecular clock were used for time estimation in our study. The method has been used other estimations for *indica–japonica* divergence (Ma and Bennetzen 2004; Vitte et al. 2004; Zhu and Ge 2005). Although there are many limitations to the use of clocks based on sequence data, including substitution rate heterogeneity among lineages, uncertainties of clock calibration and other potential sources for estimation errors (Gaut 2002), our estimations do provide an approximate timeframe. It is noted that the analysis of synonymous divergence for all genes in an insertion should minimize the inaccuracy. The estimation based on any single gene might have an unusual rate of synonymous site change. A key step in our study is to determine the common or *japonica*-specific insertions for the two species. The key insertion, chr10-2, was detected as a common insertion for the two species and an upper threshold time of *indica–japonica* divergence in this study. This result is the same as an early study by Yuan et al. (2002) who first sequenced the large chloroplast insertion in rice and found that the insertion was an insertion of pre-divergence of *indica* and *japonica* rice.

Acknowledgments We thank Dr. Jianzhi Zhang (University of Michigan) for his critical reading of the manuscript. This work was supported by National Natural Science Foundation of China (30270810) and National High Technology Research and Development Program of China (2006AA10A102).

References

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

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

Gaut BS (2002) Evolutionary dynamics of grass genomes. New Phytol 154:15–28

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*. Proc Natl Acad Sci USA 93:10274–10279

Ge S, Sang T, Lu B, Hong D (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci USA 96:14400–14405

Goff SA et al (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp *japonica*). Science 296:92–114

Hannenhalli S, Pevzner PA (1995) Transforming men into mice (polynomial algorithm for genomic distance problem). Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science. IEEE, Milwaukee, Wisconsin, pp 581–592

Hiratsuka J et al (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185–194

Huang C et al (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72–76

Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol 138:1723–1733

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35:25–34

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet 21:655–663

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

Mishmar D et al (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat 23:125–133

Muse SV (2000) Examining rates and patterns of nucleotide substitution in plants. Plant Mol Biol 42:25–43

Notsu Y, Masood S, Nishikawa T, Hubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K (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 Gen Genomics 268:434–445

Pevzner P, Tesler G (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res 13:37–45

Ricchetti M et al (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature 402:96–100

Sall T, Jakobsson M, Lind-hallden C, Hallden C (2003) Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*. J Evol Biol 16:1019–1029

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

Shahmuradov IA, Akbarova YY, Solovyev VV, Aliye JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis.* Plant Mol Biol 52:923–934

Stegemann S et al (2003) High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci USA 100:8828–8833

Tang J et al (2004) A comparison of rice chloroplast genomes. Plant Physiol 135:412–420

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

Tian X, Zheng J, Hu S, Yu J (2006) The rice mitochondrial genomes and their variations. Plant Physiol 140:401–410

Vaughan DA, Morishima H, Kadowaki K (2003) Diversity in the Oryza genus. Curr Opin Plant Biol 6:139–146

Vitte C, Ishii T, Lamy F, Brar D, Panaud O (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol Gen Genomics 272:504–511

Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of the monocot–dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Yang Z (1999) Phylogenetic Analysis by Maximum Likelihood (PAML) (University College, London, UK), http://abacusgeneuclacuk/software/pamlhtml

Yu X, Gabriel A (1999) Patching broken chromosomes with extranuclear cellular DNA. Mol Cell 4:873–881

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

Yu J et al (2005) The genome of *Oryza sativa*: a history of duplication. PloS Biol 3(2):e38

Yuan Q, Hill J, Hsiao J, Moffat K, Ouyang S, Cheng Z, Jiang J, Buell CR (2002) Genome sequencing of a 239-kb region of rice chromosome 10 L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. Mol Gen Genomics 267:713–720

Zhu Q, Ge S (2005) Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytol 167:249–265