A Comparison of Rice Chloroplast Genomes[1][w]

Jiabin Tang[2], Hong'ai Xia[2], Mengliang Cao, Xiuqing Zhang, Wanyong Zeng, Songnian Hu, Wei Tong, Jun Wang, Jian Wang, Jun Yu, Huanming Yang, and Lihuang Zhu*

Institute of Genetics and Developmental Biology (J.T., H.X., X.Z., W.Z., J.Y, H.Y., L.Z.) and Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China (J.T., H.X., X.Z., S.H., W.T., J.W., J.W., J.Y., H.Y, L.Z.); National Hybrid Rice Research and Development Center, Changsha 410125, China (M.C.); and Hangzhou Genomics Institute, Hangzhou 310007, China (S.H., J.W.)

Using high quality sequence reads extracted from our whole genome shotgun repository, we assembled two chloroplast genome sequences from two rice (*Oryza sativa*) varieties, one from *93-11* (a typical *indica* variety) and the other from *PA64S* (an *indica*-like variety with maternal origin of *japonica*), which are both parental varieties of the super-hybrid rice, *LYP9*. Based on the patterns of high sequence coverage, we partitioned chloroplast sequence variations into two classes, intravarietal and intersubspecific polymorphisms. Intravarietal polymorphisms refer to variations within *93-11* or *PA64S*. Intersubspecific polymorphisms were identified by comparing the major genotypes of the two subspecies represented by *93-11* and *PA64S*, respectively. Some of the minor genotypes occurring as intravarietal polymorphisms in one variety existed as major genotypes in the other subspecific variety, thus giving rise to intersubspecific polymorphisms. In our study, we found that the intersubspecific variations of 93-11 (*indica*) and PA64S (*japonica*) chloroplast genomes consisted of 72 single nucleotide polymorphisms and 27 insertions or deletions. The intersubspecific polymorphism rates between *93-11* and *PA64S* were 0.05% for single nucleotide polymorphisms and 0.02% for insertions or deletions, nearly 8 and 10 times lower than their respective nuclear genomes. Based on the total number of nucleotide substitutions between the two chloroplast genomes, we dated the divergence of *indica* and *japonica* chloroplast genomes as occurring approximately 86,000 to 200,000 years ago.

The intracellular organelle chloroplast has its own genome that encodes a number of chloroplast-specific components (for review, see Palmer, 1985; Sugiura, 1989). In nature, the size of this circular genome varies from 39.4 to 200.8 kb among photosynthetic plant species (Kohler et al., 1997; Turmel et al., 1999). Since the sequencing of the first two complete chloroplast genomes from tobacco (*Nicotiana tabacum*; Shinozaki et al., 1986) and liverwort (*Marchantia polymorpha*; Ohyama et al., 1986) in 1986, more than 30 other chloroplast genomes from both eukaryotic algae and land plants have been reported (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html). The rice (*Oryza sativa*) chloroplast genome sequence from Nipponbare was published in 1989 (Hiratsuka et al., 1989) and was reported to have a length of 134,525 bp. Comparative studies of the genome architecture showed that the gene order and essential gene content are highly conserved for most of the chloroplast genomes (Kato et al., 2000; De Las Rivas et al., 2002). Nevertheless, variations among the different and closely related genomes do exist over evolutionary time scales. Comparative analyses reported thus far remain incomplete for most of the extant plant taxa (Shimada and Sugiura, 1991; Morton and Clegg, 1993). It is noted that all of these chloroplast genomes were sequenced on the basis of cloned chloroplast DNA from purified organelles, which is a classical but still viable strategy for organelle genome sequencing.

Recently, the whole genome shotgun approach has been successfully applied to sequencing nuclear genomes for large eukaryotes, such as *Drosophila* (Adams et al., 2000) and rice (Yu et al., 2001; Goff et al., 2002; Wang et al., 2002; Yu et al., 2002). The original raw data from the rice whole genome shotgun sequencing project contain a large amount of reads from chloroplast and mitochondrial genome sequences that are often removed before the assembly of nuclear genomes (Yu et al., 2001; Goff et al., 2002; Wang et al., 2002; Yu et al., 2002). However, organelle genomes can also be reassembled from the organelle sequence reads with extremely high coverage if complications of gene transfer between the nuclear and organelle genomes are carefully handled. For example, if the total number of sequencing reads sampled yields nuclear genome coverage only a few folds in magnitude, it is reasonable to assume that the polymorphic sites exceeding this genome coverage are mainly legitimate chloroplast sequences. In addition, homologous sequences between nuclear and chloroplast genomes range in identity from 97% to 99%, and the sequence reads are assembled into separate contigs. Only a small fraction of the assembled contigs

[1] This work was supported by project grants from the Chinese Academy of Sciences to J.Y. and H.Y. and by grants from the National Natural Science Foundation of China (90208001) and the Chinese Academy of Sciences (KSCX2–SW–306) to L.Z.

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail lhzhu@genetics.ac.cn; fax 86–10–64873428.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.103.031245.

have reads of both nuclear and chloroplast origin with a perfect (100%) match, but all of these are less than 300 bp in length. Goff and colleagues (2002) assembled a *Nipponbare* chloroplast genome (referred to herein as *Nipponbare-S*) using their whole genome shotgun sequences, and identified some variations compared to the other published *Nipponbare* sequence (referred to as *Nipponbare-H*; Hiratsuka et al., 1989). The reported differences included 141 insertions or deletions (InDels) and 78 single nucleotide polymorphisms (SNPs), but the details of the sequence assembly and validation were not published. Therefore, we reassembled the *Nipponbare* chloroplast genome (referred to as *Nipponbare-G*) from the same shotgun sequencing data of the Syngenta Company (San Diego).

One of the goals in the Super-Hybrid Rice Genome Project carried out at the Beijing Genomics Institute has been to sequence parental genomes of a super-hybrid rice cultivar, *Liang-You-Pei-Jiu*, or *LYP-9* (Yi and Xiao, 2000). The maternal and paternal varieties of *LYP-9* are *PA64S* and *93-11*, respectively. Paternal *93-11* is a typical *indica* variety despite having gone through a rather complex breeding process for agriculturally favorable traits. Maternal *PA64S* is a composite of all three major rice subspecies, *indica*, *japonica*, and *javanica*; it is *indica*-like (estimated as 50% *indica* based on breeding history). From the recorded breeding processes, the maternally inherited chloroplast of *PA64S* is expected to be of the *japonica* variety in origin (Longping Yuan, personal communication). Previously, using high quality chloroplast sequence reads extracted from our whole genome shotgun repository, we assembled both *PA64S* and *93-11* chloroplast genomes. Comparing the *PA64S* chloroplast genome to that of *Nipponbare-G*, we found that they are identical, confirming that *PA64S* is indeed *japonica* in origin. In this paper, we compare the *93-11* assembly of an *indica* genome to the *japonica* genomes, *PA64S* and *Nipponbare-G*, and analyze the variations to estimate the date at which these genomes diverged in evolutionary history.

RESULTS

Sequence Assemblies and Validation

We reassembled the *Nipponbare* chloroplast genome using 28,000 reads (about 110 times the length of the chloroplast genome) released from Syngenta Company (referred to as *Nipponbare-G*). *Nipponbare-G* (accession no. AY522330 in GenBank) is 134,551 bp long and 26 bp longer than previously reported for *Nipponbare-H* (Hiratsuka et al., 1989). A total of 189 variations, including 79 SNPs and 110 InDels (data not shown), were found between the two independently assembled *Nipponbare* chloroplast sequences. We are unable to carry out further comparisons for *Nipponbare-G* and *Nipponbare-S* since the raw sequence data with critical information about the *Nipponbare-S* assembly are not publicly available. To date, only a general comparative result between these sequences has been published as supplemental data (http://www.tmri.org) in a recent report (Goff et al., 2002).

For each assembly project, we used the sequencing data for *PA64S* and *93-11* to screen out 52,000 chloroplast sequencing reads (with an average length greater than 540 bp at quality Q20), which are about 250 times of the genome length. The reads were assembled into one major contig in each project based on the quality of our selected raw data. Some smaller contigs were excluded because of either low coverage (i.e. the nuclear equivalent coverage was less than 4 for the entire data set) or less than perfect identity (i.e. <100% over 500 bp and <99% over the entire length) to the main contig. Each genome was finally assembled into one contig, 134,551 bp for *PA64S* (AY522331) and 134,496 bp for *93-11* (AY522329). The length of *PA64S* chloroplast genome is the same as that of *Nipponbare-G*, which is 55 bp longer than that of *93-11*. These data are also publicly available at our institutional website (http://www.genomics.org.cn/bgi/rice/main.htm).

To validate our three assemblies, we simulated their restriction maps with all 6-bp cutting restriction enzymes. The maps are identical except for those of *Nipponbare-H*, which showed some length polymorphisms with CCCGGG as the cutting site, including Cfr9I, *Sma*I, and *Xma*I. The differences can be attributed to GG and CC deletions at the sequence locations of 98,867 and 116,251 bp, respectively, in the *Nipponbare-H* assembly compared with *93-11*, *PA64S*, and *Nipponbare-G* assemblies. We noted that one fragment of 20.8 kb in *Nipponbare-H* corresponds to three fragments of 17.2 kb, 1.8 kb, and 1.8 kb in the other three assemblies. Furthermore, when we digested chloroplast DNA samples isolated from *93-11*, *PA64S*, and *Nipponbare* with *Sma*I, the three predicted restriction fragments observed each differed from the 20.8-kb predicted fragment of the *Nipponbare-H* assembly (Supplemental Fig. 1, available at www.plantphysiol.org). We concluded that the three chloroplast sequences from *93-11*, *PA64S*, and *Nipponbare-G* are very similar, as predicted. The *Nipponbare-H* assembly may be somewhat anomalous as the templates used for sequences may be variants of the *Nipponbare* chloroplast. Thus, the differences found between our assemblies and the *Nipponbare* sequence may be attributed to different methods used for the respective sequencing.

Sequence Polymorphisms

The chloroplast genome is believed to be clonal and has its own replication and DNA repair systems. A given cell, such as that of the plant leaf, often contains 400 to 1,600 copies of chloroplast genome (Pyke, 1999). These multiple-copied clones of chloroplast genome can be regarded as a population if genetic heterogeneity exists among them. Therefore, when hundreds of high quality chloroplast sequence reads are aligned,

minor polymorphic sites at a small frequency can be detected. The process resembles the sampling of a population of genomes. When we compare two subspecific varieties, the major genotypes are used, and thus the resulting variations detected are intersubspecific. The minor genotypes identified within a variety, termed intravarietal variations, are also useful because the origin of the polymorphism is attributable to a low mutation rate of the chloroplast genome. There is a possibility of homologous sequences of chloroplast DNA in nuclear and mitochondrial genomes with an identity exceeding 98%, which may also be assembled into a contig as minor genotypes. In such instances, high quality and redundancy of the sequences allow us to distinguish the real chloroplast sequences from the contaminants. We have not yet encountered such an ambiguity in our assemblies. The frequencies of major and minor genotypes at each polymorphic site can be computed from the numbers of high-quality sequence reads overlapping the site by visual and manual inspection. The results can also be verified experimentally to resolve any discrepancy that appears problematic.

In this study, we did not detect any intervarietal polymorphisms between the *PA64S* and *Nipponbare-G* sequences. The *PA64S* chloroplast genome is a typical *japonica* variety as predicted from its breeding genealogy and does not appear to have diverged from its common *japonica* ancestor. The alignments between *93-11* and *PA64S/Nipponbare-G* indicated that *93-11* is a typical *indica* variety in agreement with its recorded breeding history. Therefore, the polymorphisms we identified can be regarded as examples of intersubspecific (*indica* and *japonica*) polymorphisms.

Table I. *Minor genotype frequency of each SNP type among three chloroplast genomes*

The intersubspecific SNP types and their respective minor genotype frequencies among three varieties, *93-11, PA64S,* and *Nipponbare-G* (*Nip-G*), are listed. The oblique line in the first column (/) separates the major genotype in *93-11* and *PA64S* /*Nipponbare-G*. Minor genotype frequencies (MF) were indicated as percentage of the minor genotype in each SNP type. The statistics was based on the total sites of each SNP type in a given variety. The number in the parenthesis shows the total number of the surveyed sequence traces covering the SNP loci of a given type. MF in *93-11* is about twice as high as that of *PA64S* or *Nipponbare-G* chloroplast genome.

SNP type in 93-11/ PA64S & Nip-G	MF in 93-11	MF in PA64S	MF in Nip-G
		%	
G/A	4.3 (1,583)	2.4 (1,384)	2.5 (2,961)
C/T	5.1 (1,014)	3.9 (934)	2.7 (1,875)
A/G	4.5 (2,207)	1.5 (882)	2.9 (1,978)
T/C	4.5 (1,591)	1.8 (529)	2.6 (1,375)
A/T	3.9 (2,224)	2.4 (660)	0.7 (848)
T/A	2.3 (911)	2.4 (1,667)	0.9 (2,109)
A/C	6.6 (1,219)	3.0 (568)	2.9 (1,415)
C/A	4.2 (1,154)	1.4 (709)	1.9 (1,442)
G/C	0	0	0
C/G	9.8 (384)	0	1.2 (482)
G/T	6.0 (990)	3.7 (859)	2.5 (1,760)
T/G	5.2 (1,485)	0.5 (610)	2.2 (1,051)

Single Nucleotide Polymorphisms

A total of 72 SNPs, including 30 transitions and 42 transversions, were identified between the *93-11* and *PA64S* chloroplast genomes. The frequencies of major and minor genotypes at each polymorphic site were categorized (Table I and Supplemental Table I). In general, SNPs in the chloroplast genome occurred at a rate of 5 in 10,000 bases, which is about 8 times lower than that in its nuclear genome (estimated as 0.43%;Yu et al., 2002). Only 6 SNPs were detected in the inverted repeat (IR) regions (about 40 kb). The SNP rate in these regions was 4 times lower than that in the single copy regions. This result is consistent with a previous report that the synonymous substitution rate of IR regions was roughly 5 times lower than that of the single copy regions when the chloroplast genome sequences were compared among different species (Muse, 2000).

In all the SNPs between *93-11* and *PA64S* chloroplast genomes, only 15 SNPs (about 21%) did not change the GC content, of which 2 SNPs were transversions between G and C, and the other 57 SNPs were related to GC content changes. There were 7 variable sites that involved simple repetitive sequences. Four of them were reverse-complemented sequences, from AGACCAAG, CGTT, TTT, and AAA to CTTGGTCT, AACG, AAA, and TTT, respectively.

The number of SNPs in intergenic regions (55 SNPs) was approximately twice that of gene-coding regions (17 SNPs), and several in the gene-coding regions have produced amino acid changes (Table II). Only one hotspot region (GCTT/AAGC) was detected in ORF321 between the *93-11* and *PA64S*, resulting in one amino acid change, from Leu to Ser. Therefore, it is expected that SNPs between intersubspecific chloroplast sequences may not give rise to significant functional changes among different rice varieties.

Since chloroplast genomes of *PA64S* and *Nipponbare-G* are identical with regard to intervarietal polymorphisms, we carefully inspected the intravarietal changes. Almost all of the minor genotypes in *PA64S* and *Nipponbare-G* were also found to be the major genotypes in *93-11* assembly or vice versa (Table I and Supplemental Table I). In almost all the cases, we found more intravarietal than intersubspecific SNPs, indicating that only some of intravarietal mutations were fixed gradually and inherited stably among different rice subspecies (Fig. 1). One exceptional site was noted, located at 51,349 bp (positioned in the *PA64S* sequence). In *PA64S*, among 142 sequencing reads covering this locus, the major genotype was T (82%) and the minor was C (18%), but this minor genotype (C) was not found in the *Nipponbare-G* assembly among the 234 sequences we carefully surveyed. One potentially important observation is that the minor genotype frequency at each SNP site in

Table II. *SNPs in the coding region between 93-11 and PA64s chloroplast genome*

The locations of polymorphic sites between chloroplast genomes of *93-11* and *PA64S* were documented according to the nucleotide order of the *93-11* chloroplast DNA sequence. The oblique line (/) separates the corresponding variations of codons and amino acids in the involved genes between *93-11* and *PA64S*. Genes were annotated according to the published chloroplast genome (Hiratsuka, et al., 1989). The gene symbols in this table are as follows: ORF, open reading frame; rsp16, ribosomal protein S16; psbk, PSII K protein; rpoC2, RNA polymerase β'-subunit-2; atpA, ATPase α-subunit; rbcL, ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit; rp120, ribosomal protein L20; psbB, PSII 47-kD protein; rp116, ribosomal protein L16; rps3, ribosomal protein S3; and ndhF, NADH dehydrogenase ND5.

Location in 93-11	Gene	*Codon variation* in 91-11/PA64S	Amino acid variation in 91-11/ PA64S
4,547	rsp16	ACA/ACC	
7,141	psbk	AAC/AAT	
14,166	ORF91	AGC/AAC	Ser/Asn
27,979	rpoC2	TTG/TGG	Leu/Trp
29,073	rpoC2	GAC/AAC	Asp/Asn
35,342	atpA	GCA/GCG	
54,869	rbcL	GGA/GGG	
56,818	ORF106	GGT/GGC	
66,354	rpl20	CCA /TCA	Pro/Ser
69,301	psbB	GTG/GCG	Val/Ala
77,750	rp116	GTA/GCA	Val/Ala
79,426	rps3	GAA/GAG	
102,887	ndhf	GCA/GCG	
105,740	ORF321	TTG/TTA	
105,741; 105,742; 105,743	ORF321	CTT/AGC	Leu/Ser

93-11 is nearly twice as high as that in *PA64S* and *Nipponbare-G* (Table I). We are not able to generalize these findings to other *indica* varieties at the present time but speculate that the *indica* chloroplast population may be more polymorphic than those of *japonica*.

Insertions and Deletions

InDels among rice chloroplast genomes are quite limited in number compared to SNPs. Only 27 InDels in the *93-11*-to-*PA64S* comparisons were detected. Frequencies of the major and minor genotypes in different varieties are summarized in Table III. The cumulative length differential attributable to the InDels is 55 bp, which is consistent with the total difference in length between the *93-11* and *PA64S* assemblies. In general, the InDel rate calculated from the *93-11*-to-*PA64S* comparison is about 0.02%, nearly one-half of the SNP rate of chloroplast genomes (0.05%) and about 10 times lower than that of its nuclear counterpart (a genome average of 0.23%; Yu et al., 2002). Most of the InDels were found in repetitive sequences and located in the noncoding region of chloroplast genomes, such as the InDels of I-32 and D-69 at the *93-11* chloroplast genome positions 17,741 bp and 8,554 bp, respectively (Table III). We have further noted that all single nucleotide InDels are A or T variations, and no single G or C InDel was detected. There were no InDels detected in the IR regions where the GC content was estimated to be 57.2%, much higher than the chloroplast genome average. Some of InDels were linked as haplotypes, such as D-9 (AAAAAAAAG) and D-2 (TT) deletions in the *93-11* assembly at positions 57,009 bp and 57,011 bp, respectively. These results indicate that the replication and DNA repair systems in the chloroplast may be biased, resulting in more frequent sequence variations in the low GC content regions and repeat sequences.

Experimental Confirmation of Sequence Polymorphisms between *Indica* and *Japonica*

Although sequence redundancies (about 250 times the coverage of the total chloroplast genome length) guaranteed high-quality assemblies, we nevertheless conducted experiments to confirm several of the observed variations. Primers were designed (see "Materials and Methods") to amplify the two InDels, D-69 and I-32 (Supplemental Fig. 2), and one hotspot region (located at 62,474 bp of the *93-11* assembly) that harbors the nucleotide changes from AGACCAAG in *93-11* to CTTGGTCT in *PA64S*. The amplified DNA

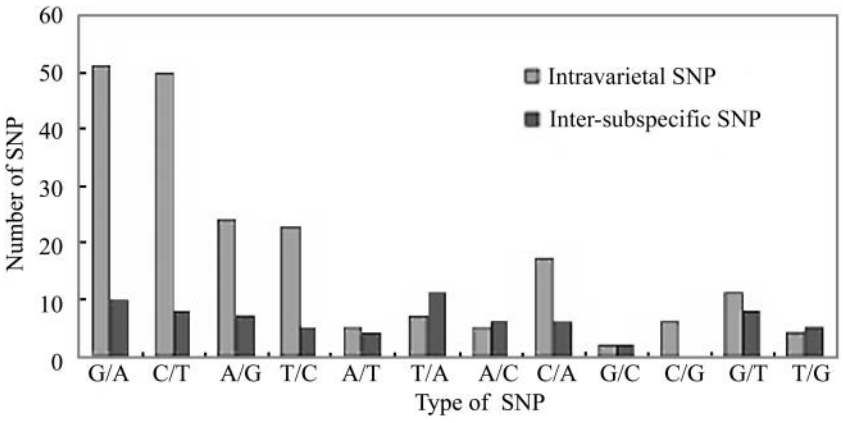

Figure 1. Comparisons of intravarietal and intersubspecific SNPs in rice chloroplast genomes. Intersubspecific SNPs are the ones found between *japonica* (*Nipponbare-G*) and *indica* (*93-11*). Intravarietal SNPs are those found within *Nipponbare-G*. The analysis was based on SNP types, including four transitions (G/A, C/T, T/C, and A/G) and eight transversions (A/T, T/A, A/C, C/A, G/C, C/G, G/T, and T/G). The oblique line (/) separates the major and minor genotypes in intravarietal SNPs and the major genotype in intersubspecific SNPs, respectively.

Table III. *Genotype frequencies of InDels among rice chloroplast genomes*

All the InDels between *93-11* and *PA64S*/*Nipponbare-G* are listed according to the nucleotide order in the *93-11* chloroplast sequence. I or D depict insertion or deletion of a major genotype at a given polymorphic site. The numbers following D or I, joined with hyphens, indicate the InDel length in basepairs. At a given intersubspecific InDel site, the major genotype in *93-11* could be the minor genotype in *PA64S* or *Nipponbare-G* or vise versa. Frequency (F) and Minor genotype frequency (MF) are percentages of the major and minor genotype at a given polymorphic locus. The number in the parentheses is the total number of the surveyed sequence traces.

InDels	Locus in 93-11	Sequence	F in *93-11*	MF in *93-11*	MF in *PA64S*	MF in *Nipponbare-G*
				%		
I-7	5,016	CTTTATC	94 (205)	6 (205)	0 (103)	0 (234)
D-1	6,252	T	92 (193)	8 (193)	1 (108)	3 (232)
D-69	8,554	GAATCCTATTTTTGTTCTTATA CCCATGCAATAGAGAGCGAG TGGGAAAAGGGAGGTTACTT TTTTTCA	100 (69)	0 (69)	0 (118)	0 (188)
D-4	12,610	AGGG	96 (200)	4 (200)	5 (137)	2 (352)
D-2	13,946	AC	95 (228)	5 (228)	0 (150)	3 (314)
I-1	16,613	A	98 (215)	2 (215)	0 (114)	0 (230)
D-6	17,324	TAGAAA	99 (306)	1 (306)	0 (134)	0 (336)
I-32	17,741	TAACAAATTCTTAGAGTATTTC TGGTAGAATT	97 (261)	3 (261)	1 (184)	2 (375)
I-1	43,861	A	98 (206)	2 (206)	0 (114)	0 (177)
D-5	46,050	TATAT	93 (209)	7 (209)	1 (95)	11 (232)
D-1	46,136	T	100 (129)	0 (129)	0 (66)	0 (140)
D-6	46,494	AGAAAA	96 (227)	4 (227)	0 (128)	0 (222)
I-1	47,166	T	96 (137)	4 (137)	6 (50)	3 (241)
D-1	56,998	T	90 (186)	10 (186)	0 (86)	3 (144)
D-9	57,009	AAAAAAAAG	90 (187)	10 (187)	0 (95)	0 (119)
D-2	57,011	TT	90 (187)	10 (187)	0 (95)	0 (119)
I-5	57,591	AAAGT	96 (227)	4 (227)	0 (148)	0 (236)
I-5	60,813	TGTAT	98 (159)	2 (159)	2 (113)	5 (195)
D-2	65,568	TT	91 (202)	9 (202)	0 (125)	2 (234)
D-1	75,933	T	95 (152)	5 (152)	3 (89)	3 (161)
D-1	76,184	A	99 (213)	1 (213)	5 (116)	1 (166)
D-1	76,524	T	100 (216)	0 (216)	4 (129)	0 (106)
I-3	77,677	TGG	96 (197)	4 (197)	0 (142)	0 (206)
D-1	78,383	T	96 (142)	4 (142)	0 (130)	0 (100)
I-2	80,564	TT	100 (242)	0 (242)	0 (196)	1 (173)
I-4	104,486	CAAA	99 (287)	1 (287)	3 (61)	2 (400)
D-4	134,497	AATA	100 (130)	0 (130)	1 (150)	0 (158)

fragments were subsequently sequenced, and all were verified at the sequence level (Fig. 2). The results demonstrated that two InDels are duplications or deletions of low complexity sequences or simple repeats. The mutation hotspot was also verified as the reverted variation. Among the polymorphisms between the two subspecific chloroplast genomes identified by us, only D-69 was reported previously (Kanno and Hirai, 1993).

We attempted to validate some of the InDels found between *Nipponbare-H* and other *japonica* varieties. A pair of primers was designed to test the InDel I-15 that represented a 15-bp-deletion (CGAATTCCTATAGTA) located at position 53,857 bp in the *Nipponbare-H* sequence. This deletion, as well as three other predicted variations, was not found in the *indica* or *japonica* varieties used for our experiment, including cultivars of *Nipponbare* (Fig. 2c and Fig. 3). Exhaustive searches for such a 15-bp-deletion over 550 times the redundant sequences among all raw data traces available did not yield a single relevant sequence too (data not shown).

Segregation analysis of the InDels D-69 and I-32 in the F_2 populations from the cross combination of *PA64S* and *93-11* showed that all individual plants in the F_2 generation having *PA64S* as the maternal parent had the same PCR product band as *PA64S*, further validating our sequence assemblies (Supplemental Fig. 3). In order to study the distribution of the two large InDels, D-69 and I-32, 27 different cultivars from *indica* and *japonica* subspecies were surveyed. The result showed that these InDels were common polymorphisms between *indica* and *japonica* subspecies, with only one exception occurring in an *indica* variety (Supplemental Fig. 4). These polymorphisms were absent in only 1 out of 35 *japonica* varieties (Supplemental Fig. 5) and 7 out of 27 *indica* varieties (Supplemental Fig. 6). This result is in accordance with previous reports (Dally and Second, 1990; Sun et al., 2002). In addition, our findings indicated that

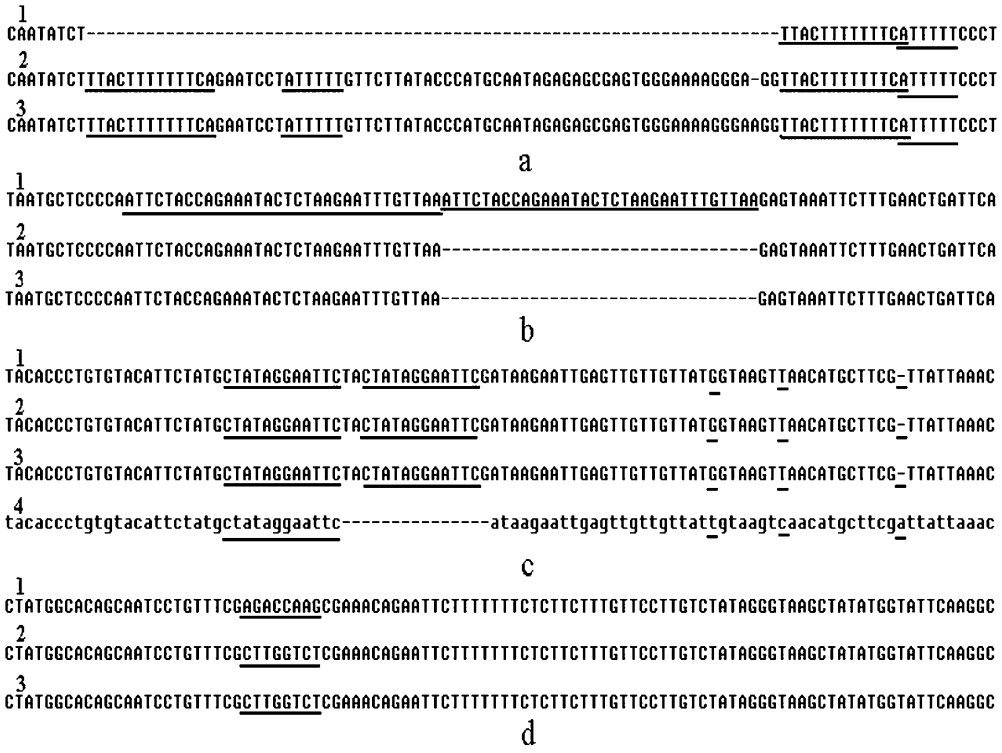

Figure 2. Multiple alignments of the intersubspecific chloroplast variations. Sequences of the intersubspecific variations of D-69 InDel (a), I-32 InDel (b), I-15 InDel (c), and one hotspot (d) were aligned. The corresponding chloroplast sequences amplified from *93-11, PA64S,* and *Nipponbare-G* are indicated in numbers (1, 2, and 3). The published sequence (4) of *Nipponbare-H* from the National Center for Biotechnology Information databases is shown in section c. The repeated sequences and SNPs are underlined. The dashed lines indicate deleted bases.

D-69 and I-32 are genetically linked as one haplotype (Supplemental Figs. 5 and 6).

DISCUSSION

While it is well known that chloroplast genomes in plant leaf cells are not absolutely homogenous, rigorous confirmation of the nature of their differences requires cloning and sequencing. We have defined a way to study such intravarietal polymorphisms, and we have demonstrated that minor genotypes are detectable at frequencies ranging from a few percents to a few tens of percents. It is noteworthy that a few polymorphic sites were found to have more than one minor genotype. For example, at the polymorphic site located at 51,292 bp in the *93-11* chloroplast genome, among 198 reads surveyed the major genotype is A (85%), and two other minor genotypes were, respectively, T (5%) and G (10%). In addition, each major genotype detected in the intersubspecific chloroplast genome comparisons, such as between *93-11* and *PA64S,* could also be identified as either major or minor genotypes at the corresponding polymorphic site among the intravarietal variations. Often, the intravarietal minor genotype in one subspecies would be observed as a major genotype in other subspecies,

Figure 3. Sequence alignment of the I-15 InDel and three SNPs among different *japonica* varieties. The sequences were aligned around I-15 InDel from nine *japonica* varieties: *MiYang46* (1), *ShenNong1033* (2), *Taibei309* (3), *DV10* (4), *DV85* (5), *E32* (6), *DiGu* (7), *CBB7* (8), and *C418* (9). The 10th row is the published chloroplast sequence of *Nipponbare-H* from GenBank. SNPs are underlined and dashed lines indicate deletions.

or vice versa. This finding suggests that the minor genotypes are chloroplast in origin rather than results of nuclear or mitochondrial DNA contaminations. For example, one chloroplast genome polymorphic site with A in *93-11* and T in *PA64S* is found in *93-11* with A as a major genotype and T as a minor genotype. For *PA64S*, T is a major genotype and A is a minor genotype. Therefore, the frequency at which a minor genotype is detected at a given polymorphic site provides a useful statistical basis for comparing sequence variations among the chloroplast genomes.

The sequence polymorphisms among multiple copies of the chloroplast genomes are inherited maternally as an intravarietal population, but the inheritability of the resultant chloroplast mutations are different from those of endoreduplication that frequently occurs in certain somatic cells of rapidly growing plant tissues (Joubes and Chevalier, 2000). Only when the mutations resulted from endoreduplication that occurred in germ cells could they become inheritable. Nevertheless, we did not find intravarietal sequence polymorphisms in the nuclear genome sequences in our assemblies of the two rice varieties, *93-11* and *PA64S* (data not shown).

To evaluate the role of inter-genomic gene transfer between the organelle and the nuclear genomes, we surveyed all publicly available rice genomic sequences for the presence of any major and minor genotype sequences found in this study. Of the 99 polymorphisms (including 72 SNPs and 27 InDels) discovered as intersubspecific chloroplast DNA variations, eleven SNPs and seven InDels did not have matching variable sites in the homologous nuclear counterparts. This result suggests that most of the inter-subspecific SNPs and some of the InDels may be quite old, or, alternatively, that recent transfer events between the chloroplast and nuclear genomes have occurred. The chloroplast homologous sequences are nevertheless easily identifiable from the surrounding sequences and higher variation rate coexisting in the nuclear genome. In a previous study of the chloroplast homologous sequences in rice mitochondrial genome, it was reported that a total of sixteen chloroplast sequences (about 22 kb), ranging from 32 bases to 6.8 kb in length were dispersed throughout the mitochondrial genome (490.570 kb; Nakazono and Hirai, 1993). In our survey, only 3 out of 99 chloroplast sequences containing polymorphic sites were found to have mitochondrial homologous sequences. The transfer of DNA from chloroplast to mitochondria appears to be a very rare event and might require a nuclear intermediate.

Most of the InDels in chloroplast genomes between two rice subspecies exist as short and simple repetitive sequences in noncoding regions, which, therefore, may not have functional consequences. Furthermore, only a few SNPs in coding sequences cause amino acid changes in the chloroplast encoded proteins. The rate of transversion versus transition in intersubspecific SNPs is different from that in intravarietal SNPs. The rate of transversion versus transition in the intersubspecific SNPs between *93-11* and *PA64S* is 1.4, and a similar rate is found between chloroplast genomes *93-11* and *Nipponbare-G*. On the other hand, the rate of transversion versus transition in the 205 intravarietal SNPs of the *Nipponbare-G* is only 0.4, clearly biased toward transitions and consistent with previously reported results (Alain et al., 2002). Furthermore, the sites of G/A and C/T transition detected in the intravarietal SNPs of *Nipponbare-G* are about twice those of the A/G and T/C transitions (data not shown). These results are again attributable to a GC content bias of the DNA polymerase specific to the chloroplast, as the rate of transition versus transversion in rice nuclear genome between *93-11* and *Nipponbare* is close to 1.0 (unpublished data available from authors upon request).

Chloroplast genomes diverge at a much different rate than their nuclear genomes. The overall sequence difference between rice subspecific varieties in the nuclear genomes is about 130 times higher than that of the chloroplast (0.12%; Yu et al., 2002). When two diverged nuclear genomic sequences are compared, a large amount of the transposable elements often appear to interrupt the comparison. A valid alternative approach is, thus, to compare SNPs and InDels in the context of repetitive and unique sequences. Since most of the chloroplast genome is coding sequence, the comparison of mutation rates between nuclear and organelle genomes should be conducted with the unique sequences. When the rates of SNP and InDel sites between the two rice varieties are compared in this strict sense, we can conclude that these rates for chloroplasts (0.05% and 0.02%) are approximately 8 and 10 times lower than those of nuclear genomes (0.43% and 0.23%), respectively. These results are slightly different from a previous study in which the chloroplast DNA was reportedly evolving 4 times more slowly than its nuclear counterpart (Wolfe et al., 1987). According to the total number of nucleotide substitutions ($ds = 0.0005$) between *93-11* and *PA64S* calculated by using ClustalX software (ftp://ftp-igbmc.u-strasbg.fr/pub/clustalX), we roughly deduced that the divergent time ($T = ds/2r$, $r = 1.242.94 \times 10^{-9}$; Muse, 2000) of *indica* and *japonica* chloroplast genomes was about 86,000 to 200,000 years ago. In contrast, based on the nuclear genome comparison, the estimated divergent time for *indica* and *japonica* nuclear genomes was about 1,000,000 years ago. It is not inconceivable that the nuclear genome and the organelle genome diverged at different time in evolutionary history since the organelle genome is mainly maternally inherited.

MATERIALS AND METHODS

Sequence Assembly and Analysis

High quality sequencing reads were extracted from our whole genome shotgun sequence repository (continual nucleotide length more than 50 bp at Phred value Q20; http://www.genomics.org.cn; Yu et al., 2002) according to

their identities to known rice chloroplast genome sequences with the expectation value of 1e–100. The sequences were assembled into contigs by using the software package Phred-Phrap-Consed (Ewing and Green, 1998; Gordon et al., 1998, 2001). The DNA sequence data used for *Nipponbare-G* assembly were from Syngenta Company (http://www.tmri.org; Goff et al., 2002). Polymorphisms, including SNPs and InDels, among the chloroplast genome sequences were identified with a compiled software tool, SNP_CROSS.PL. Results were confirmed by careful visual inspections. The frequencies of major and minor genotypes were manually tabulated.

Rice Materials

The *93-11* is a typical *Oryza sativa* cv *indica* that was bred in Jiangsu Academy of Agricultural Sciences, China (Dai et al., 1997). *PA64S* is a photoperiod- and temperature-sensitive male sterile cultivar and was bred in National Hybrid Rice Research and Development Center (NHRRDC), China (Yi and Xiao, 2000). Its maternal parent is *Nong-Ken-58*, a *japonica* cultivar. Both *93-11* and *PA64S* were provided by Longping Yuan of NHRRDC. F_2 population was come from the cross combination of *PA64S* and *93-11*.

The japonica varieties used for the experiments are Taibei309, LiJiangXin-TuanHeiGu, Lemont, ShenNong1033, DV10, CBB7, DV85, ZhongHua8, C418, JiangNanXiangNuo, E32, DiGu, JingXi17, MiYang46, 02428, Yongjing36, Qiuguang, Wanhui31, Yongjing27, Nongken58, Jiejieqing, Qihongqing, *Jin1244-2, Ji86-11,* Jia64, Chujing23xuan, Chunjiangnuo3, Huangjinqing, Heizhong, and Manhonggu. The indica varieties are: TeQing, MingHui63, NanJing6, XiaoQingZhan, MoLiZhan, ZhaiYeQing8, Gui630, IR24, IRBB5. *B6532-MR-2-5-1, B6582F-MR-14-1-2-3,* 1nongzhu, Pin9501, Paozhugu4, Qingzhen8, Hao'an2, Luweidao2, Guisi, Yigenmiao, Dalidao, Youzhidao2, Luweidao, Erkuaigu3, Haonuo2, and Wangdao1. Three javanica varieties are: SR3, C bao, and Dular.

Restriction Digestion Analysis

Extraction of chloroplast DNA and restriction digestion were carried out according to published protocols (McCouth et al., 1988; Triboush et al., 1998). The restriction-digested DNA was separated in 1.2% agarose with either constant field electrophoresis or pulse-field-electrophoresis. For pulse-field-electrophoresis, the published parameters of Carle et al. (1986) were used: 6 V/cm (voltage), 1 s (initial switch time), 5 s (final switch time), and 120° (including-angle). The gel was stained with ethidium bromide and destained with water before photography.

PCR Analysis

The sequences of the three of primer pairs for the two larger InDels (D-69 and I-32) and one mutation hotspot region (or highly variable region; Ogihara et al., 2002) found in the *93-11-to-PA64S* comparison are 5′-GAAAGAAAGA-TAAAGTAAG-3′ and 5′-TTCTTCTTTAGTAACTCTAC-3′, 5′-CCCCAT-ACTACTAGTGAAAGAG-3′ and 5′-GCACCCCGCATAAGTTG-3′, and 5′-GCACCCCGCATAAGTTG-3′, and 5′-GCAATGAATAGGGAAGGTA-TAG-3′ and 5′-GGAAAGGTTGGATTCGAA-3′, respectively. The sequences of the primer pair for testing a 15-bp-deletion in *Nipponbare-H* are 5′-GATT-TGGGGTTGCGCTAT-3′ and 5′-CTGAGGAGTTACTCGGAATGCT-3′. PCR reactions were performed in a final volume of 25 μL, containing 0.7 units of Ampli-Tag polymerase, dNTPs (200 μM), 1× GeneAmp PCR buffer (ABI, Sunnyvale, CA), and 10 pmol primers.

Sequence data from this article have been deposited with the GenBank data library under accession numbers AY522330, AY522331, and AY522329.

ACKNOWLEDGMENTS

We are grateful to Dr. Qian Qian (Chinese National Center for Rice Improvement) and Syngenta Company (http://www.tmri.org) for kindly supplying the rice materials and the sequencing reads of *Nipponbare*, respectively. We thank Drs. Gwendolyn Zahner and Lin Wu for critical reading of the manuscript.

Received August 1, 2003; returned for revision January 28, 2004; accepted February 10, 2004.

LITERATURE CITED

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al (2000) The genome sequence of Drosophila melanogaster. Science **287:** 2185–2195

Alain V, Denis M, Magali SC, Andrè E (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol **34:** 275–305

Carle GF, Frank M, Olson MV (1986) Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science **232:** 65–68

Dai ZY, Zhao BH, Liu XJ (1997) A new medium *indica* variety with fine quality, high yield and muti-disease resistance. Jiangsu Agricultural Sciences **1:** 13–14

Dally AM, Second G (1990) Chloroplast DNA diversity in wild and cultivated species of rice (genus O*ryza,* Section O*ryza).*Cladistic-mutation and genetic-distance analysis. Theor Appl Genet **80:** 209–222

De Las Rivas J, Lozano JJ, Ortiz AR (2002) Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res **12:** 567–583

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:** 186–194

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296:** 92–100

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res **8:** 195–202

Gordon D, Desmarais C, Green P (2001) Automated finishing with autofinish. Genome Res **11:** 614–625

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY (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

Joubes J, Chevalier C (2000) Endoreduplication in higher plants. Plant Mol Biol **43:** 735–745

Kanno A, Hirai A (1993) A transcription map of the chloroplast genome from rice (Oryza sativa). Curr Genet **23:** 166–174

Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S (2000) Complete structure of the chloroplast genome of a legume, Lotus japonicus. DNA Res **7:** 323–330

Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS (1997) A plastid of probable green algal origin in Apicomplexan parasites. Science **275:** 1485–1489

McCouth SR, Kochert G, Yu ZH (1988) Molecular mapping of rice chromosomes. Theor Appl Genet **76:** 815–829

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (Poaceae). Curr Genet **24:** 357–365

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

Nakazono M, Hirai A (1993) Identification of the entire set of transferred chloroplast DNA sequences in the mitochondrial genome of rice. Mol Gen Genet **236:** 341–346

Ogihara Y, Isono K, Kojima T, Endo A, Hanaoka M, Shiina T, Terachi T, Utsugi S, Murata M, Mori N, et al (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol Genet Genomics **266:** 740–746

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, et al (1986) Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature **322:** 572–574

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet **19:** 325–354

Pyke KA (1999) Plastid division and development. Plant Cell **11:** 549–556

Shimada H, Sugiura M (1991) Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res **19:** 983–995

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al (1986) The complete nucleotide sequence of tobacco

chloroplast genome: its gene organization and expression. EMBO J **5:** 2043–2049

Sugiura M (1989) The chloroplast chromosomes in land plants. Annu Rev Cell Biol **5:** 51–70

Sun CQ, Wang K, Yoshimura A, Doi K (2002) Genetic differentiation for nuclear, mitochondrial and chloroplast genomes in common wild rice (Oryza rufipogon Griff.) and cultivated rice (Oryza sativa L.). Theor Appl Genet **104:** 1335–1345

Triboush SO, Danilenko NG, Davydenko OG (1998) A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. Plant Mol Biol Rep **16:** 183–189

Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci USA **96:** 10248–10253

Wang J, Wong GK, Ni P, Han Y, Huang X, Zhang J, Ye C, Zhang Y, Hu J, Zhang K, et al (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. Genome Res **12:** 824–831

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA **84:** 9054–9058

Yi JZ, Xiao WZ (2000) The production technology of the Liang-You-Pei-Jiu (LYPJ). Hybrid Rice **1:** 76–77

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2001) A draft sequence of the rice (Oryza sativa ssp. indica) genome. Chin Sci Bull **46:** 1937–1942

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296:** 79–92