Mol Genet Genomics (2002) 267: 713–720
DOI 10.1007/s00438-002-0706-1

ORIGINAL PAPER

Q. Yuan · J. Hill · J. Hsiao · K. Moffat · S. Ouyang
Z. Cheng · J. Jiang · C.R. Buell

Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion

Received: 18 February 2002 / Accepted: 27 May 2002 / Published online: 27 June 2002

Abstract In this study we describe a 239-kb region on the long arm of rice chromosome 10 that contains a high density (71%) of locally duplicated genes, including 24 copies of a glutathione S-transferase gene. Intriguingly, embedded within this cluster is a large insertion (~33 kb) of rice (*Oryza sativa*) chloroplast DNA that is derived from two separate regions of the chloroplast genome. We used DNA fiber-based fluorescence in situ hybridization (fiber-FISH) analyses of *O. sativa* spp. *japonica* nuclei to confirm that the insertion of organellar DNA was not a cloning artifact. The sequence of the chloroplast insertion is nearly identical (99.7% identity) to the corresponding regions in the published rice chloroplast genome sequence, suggesting that the transfer event occurred recently. PCR amplification and sequence analysis in two subspecies of rice, *O. sativa* spp. *japonica* and spp. *indica*, indicates that the transfer event predated the divergence of these two subspecies. The chloroplast insertion is flanked by a 2.1-kb perfect direct repeat that is unique to this location in the rice genome.

Keywords Glutathione S-transferase · Chloroplast · Gene families · Repetitive sequence · *Oryza sativa*

Introduction

While *Arabidopsis thaliana* represents a model for dicotyledonous plants, rice (*Oryza sativa*) provides a model for monocotyledonous plants, which include the agriculturally important cereal species maize, wheat, barley, oats, and sorghum (Goff 1999). As a consequence, the rice genome is the target of genome sequencing efforts. There are currently four separate projects focused on obtaining the rice genome sequence. Two corporate efforts were focused on sequencing *O. sativa* spp. *japonica* var. Nipponbare (Barry 2001; Goff et al. 2002) and the draft sequence data can be accessed through academic licensing agreements with each corporate entity. A public effort, also focused on sequencing the Nipponbare cultivar, is being carried out by the International Rice Genome Sequencing Project (Sasaki and Burr 2000). In this public effort, a BAC-by-BAC approach is being used to generate high-quality sequence that is available immediately through public databases such as Genbank. In addition to the public effort on the *japonica* cultivar Nipponbare, a second public effort focused the closely related subspecies *indica* has been completed with the recent release of a draft sequence of the *indica* genome (Yu et al. 2001, 2002; http://210.83.138.53/rice/). Collectively, these projects provide a resource that can be mined for genes and used to analyze genome structure in a second plant species.

Within a given genome, genes occur as single-copy sequences or as members of gene families. The organization of gene family members is varied; ranging from tandem duplication to random distribution within the genome. In *Arabidopsis*, in which an estimated 25,498 genes have been identified (The Arabidopsis Genome Initiative 2000), 65% of the genes are members of gene families, with 37.4% of the genes belonging to gene families that contain at least five members. Gene family members can be distributed throughout the genome, but a surprising 17% of all the *Arabidopsis* genes can be found in tandem repeats.

In addition to the nuclear genome, plants have mitochondrial and plastid genomes, all of which undergo intra- and inter-genomic recombination and rearrangement. In *Arabidopsis*, all three genomes have been sequenced: the nuclear, chloroplast and

Communicated by M.-A. Grandbastien

Q. Yuan · J. Hill · J. Hsiao · K. Moffat · S. Ouyang
C.R. Buell (✉)
The Institute for Genomic Research,
9712 Medical Center Drive, Rockville, MD 20850, USA
E-mail: rbuell@tigr.org
Tel.: +1-301-8383558
Fax: +1-301-8380208

Z. Cheng · J. Jiang
Department of Horticulture,
University of Wisconsin, Madison, WI 53706, USA

mitochondrial (Unseld et al. 1997; Sato et al. 1999; The Arabidopsis Genome Initiative 2000). With the completion of the *Arabidopisis* genome sequencing project, it was possible to identify transfer events from the organelles to the nucleus on a whole-genome scale. The most striking event is the insertion of more than 620 kb of the mitochondrial genome on chromosome 2 near the centromere (Lin et al. 1999; Stupar et al. 2001). However, other transfers, both from the mitochondrial and plastid genomes, have been detected reflecting additional transfer events (The Arabidopsis Genome Initiative 2000; Rujan and Martin 2001).

In this study, we analyzed the sequence of two rice bacterial artificial chromosome (BAC) clones (239 kb total) derived from the long arm of rice (*O. sativa* spp *japonica* cv. Nipponbare) chromosome 10 (10L) that contained a gene-rich region with a high frequency of locally duplicated genes. This region also contains a large segment of the rice chloroplast genome inserted between perfect direct repeats of 2.1 kb. The high frequency of local gene duplication and the chloroplast insertion provide a resource with which to examine evolutionary processes in a model plant species.

Materials and methods

DNA sequencing and molecular methods

BAC DNA was isolated separately from OSJNBb0005J14 and OSJNBa0034L04 using the Qiagen Maxi-prep method, following the directions supplied by the manufacturer (Qiagen, Valencia, Calif.). DNA was sheared using a nebulizer, treated with *Bal*31 exonuclease and T4 polymerase, and size fractionated on an agarose gel (Ausubel et al. 1994). Two shotgun libraries, containing inserts of 2–3 kb and 8–10 kb, respectively, were constructed in a modified pBR322 vector using standard molecular techniques (Ausubel et al. 1994; Tettelin et al. 2001). Templates were prepared using the Direct Bind kit (Eppendorf-5 Prime, Boulder, Colo.) and sequenced on ABI 3700 machines (Applied Biosystems, Foster City, Calif.) using standard sequencing methods. Shotgun clones were sequenced to generate 13-fold coverage for OSJNBb0005J14 and 8-fold for OSJNBa0034L04. The shotgun sequences for each BAC were assembled separately using TIGR Assembler (Sutton et al. 1995). A combination of primer walking, direct PCR sequencing and re-sequencing of individual shotgun clones was used to close the sequencing gaps in these two BAC clones (Lin et al. 1999; Tettelin et al. 2001).

To confirm that the sequence assembly was correct, BAC DNA was digested with restriction enzyme(s), and size fractionated on agarose gels (Ausubel et al. 1994). The resulting optical fingerprints were then compared to the electronic digests that were generated in silico. In addition, primers were designed based on the sequences that flank the junction of the chloroplast insertion and PCRs were conducted to amplify the targeted regions. PCR was performed using three BAC clones: OSJNBb0005J14 and two other clones (OSJNBb0002E24, OSJNBa0032B13) that overlapped OSJNBb0005J14 and contained the junction regions of the chloroplast insertion. Rice genomic DNA samples from *O. sativa* spp. *japonica* var. Nipponbare or *O. sativa* spp. *indica* var. IR64 were used for verification by PCR amplification. PCR products were either directly sequenced or cloned using the Invitrogen TOPO-TA cloning kit (Invitrogen, Carlsbad, Calif.) prior to sequencing.

Fiber-FISH

FISH procedures on genomic DNA fibers and BAC molecules were performed essentially as described in the published protocol (Jackson et al. 1998, 1999) with only minor modifications. For genomic fiber-FISH, a suspension (2 µl) of nuclei was deposited at one end of a poly-L-lysine slide (Sigma, St. Louis, Mo.) and allowed to dry for 5–10 min. STE lysis buffer [8 µl; STE: 0.5% (w/v) SDS, 5 mM EDTA, 100 mM TRIS-HCl (pH 7.0)] was pipetted on top of the nuclei on the slide and incubated at room temperature for 4 min. A clean coverslip was used to drag the solution slowly down the slide. For BAC molecule fiber-FISH, BAC DNA was isolated from overnight 5-ml cultures inoculated with single colonies. Approximately 20 ng of BAC DNA was diluted in 10 µl of distilled water and then pipetted onto a poly-L-lysine glass slide. A coverslip was slowly lowered onto the solution (drop coverslip extension method) thereby extending the DNA via fluid dispersion. Hybridization and post-hybridization washing stringencies were the same as those used for FISH on chromosomes (Jiang et al. 1995). The biotin-labeled probes were detected with avidin DN antibodies (Vector Laboratories, Burlingame, Calif.) followed by biotinylated anti-avidin DN (Vector Laboratories) and finally with fluorescein-avidin DN antibodies (Vector Laboratories). Digoxigenin-labeled probes were detected with mouse anti-digoxigenin (Boehringer Mannheim, Indianapolis, Ind.), followed by digoxigenin-conjugated sheep-anti-mouse IgG antibodies (Boehringer Mannheim) and finally with rhodamine anti-digoxigenin (Boehringer Mannheim).

Annotation of OSJNBb0005J14 and OSJNBa0034L04

To annotate the BAC clones, the assembled sequences were searched against a non-redundant amino acid sequence database, the TIGR plant gene index databases (http://www.tigr.org/tdb/tgi.shtml; Quackenbush et al. 2001), and the TIGR rice repeat sequence database (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml). The sequences were analyzed with several gene prediction programs, including FGENESH (http://www.softberry.com), Genemark.hmm (rice matrix; http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi), Genscan (maize matrix; http://genes.mit.edu/GENSCAN.html), Genscan+ (Arabidopsis matrix; http://genes.mit.edu/GENSCAN.html) and GlimmerM (rice matrix; http://www.tigr.org/tdb/glimmerm/glmr_form.html). Transfer RNAs were predicted by tRNAscan-SE (Lowe and Eddy 1997). Simple repeats were identified and annotated with RepeatMasker2 (A. F. A. Smith, and P. Green, unpublished; http://ftp.genome.washington.edu/RM/RepeatMasker.html). All working models were curated manually using both the database search evidence and the output from the *ab* initio gene finders. The predicted protein sequences from the edited gene models were then searched against Hidden Markov Models (HMMs) of protein domains from Pfam (Sonnhammer et al. 1998) and TIGRfam (Haft et al. 2001) using the HMMer program. For curated gene models that were identical to a previously characterized gene, the name of the gene was conserved in our annotation. For curated gene models with a blastp score greater than 100 to a known gene, the gene model was labeled as "putative XXX" or "XXX-like protein". For gene models that showed greater than 98% identity to a rice EST but no significant similarity to any known gene, the model was labeled as an "unknown protein". For gene models that displayed no similarity to an entry in the non-redundant amino acid database or dbEST database but were predicted by at least two *ab initio* gene finders, the model was labeled as "hypothetical protein".

Results

Sequence analysis

OSJNBb0005J14 (GenBank Accession No. AC074232) was anchored to rice chromosome 10L at 61.7 cM by the

genetic marker E10477S (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html). The BAC clone OSJNBa0034L04 (Accession No. AC091680) overlaps with OSJNBb0005J14 by 26,464 bp (Fig. 1). The correct assembly of these two BACs was confirmed by comparison of the optical and electronic digestion patterns (data not shown) and by comparison of sequence alignment in the 26-kb overlap region between these two BACs. We constructed a pseudomolecule of this region of rice 10L by ligating OSJNBa0005J14 with OSJNBa0034L04 to generate 239,079 bp of unique sequence.

Annotation of OSJNBb0005J14 and OSJNBa0034L04

Prior to annotation, sequence alignment of the 239-kb region with the rice chloroplast sequence (*O. sativa* spp. *japonica*; GenBank Accession No. NC_001320, gi:11466763; Hiratsuka et al. 1989) revealed the presence of a large chloroplast DNA insertion (32,974 bp). The chloroplast DNA insertion was derived from two separate fragments of the circular rice chloroplast genome. The larger chloroplast insertion fragment (22,013 bp) originated from the large single-copy (LSC) region, starting from the 3′-end of the chloroplast gene *rpoC2* and extending to the *tRNA-Val* gene (Hiratsuka et al. 1989; Fig. 1). The smaller chloroplast insertion fragment (10,961 bp) was from one of the two inverted repeat regions (IR_A) of the rice chloroplast genome, starting from rice chloroplast *tRNA-Asn* and stretching as far as *tRNA-Val*. Overall, the sequence of the chloroplast insertion was nearly identical (99.7% identity) to the corresponding regions of the rice chloroplast genome, with only 92 bp differing out of 32,974 bp (Fig. 2).

Annotation of this 239-kb region identified a total of 73 genes, including 24 copies of glutathione S-transferase (GST; 22 complete, 1 partial pseudogene, 1 pseudogene), 24 chloroplast genes, 17 hypothetical genes, and 8 putative genes. The putative gene models that were annotated included three cytochrome P450-like proteins, three putative polyproteins, a putative serine protease, and a putative disease resistance protein. Of the 17 hypothetical genes, five could be clustered into two separate gene families with two and three members. Thus, of the 49 non-chloroplast genes within this 239-kb region, 35 (71%) were members of locally duplicated gene families.

A survey of the publicly available rice genomic sequence data (a total of 256 Mb), revealed the presence of 63 GST genes in the rice genome (data not shown). Of these 63 genes, 24 are located on chromosome 10L.

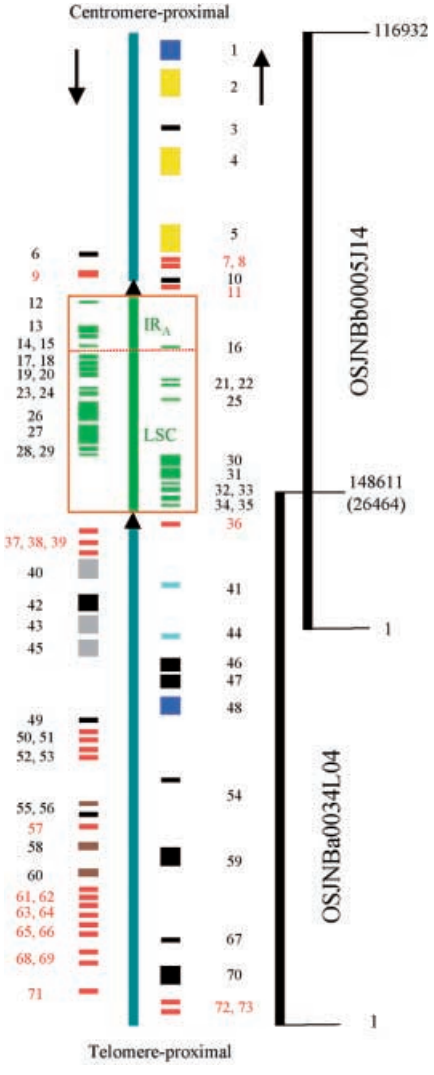

Fig. 1. Gene distribution within a 239-kb region surrounding the chloroplast insertion on rice chromosome 10L. The overall direction of transcription for each strand is indicated by a *black arrow*. The segment of the chloroplast genome inserted on the long arm of rice chromosome 10 is highlighted within the *box*. Gene models are depicted as *rectangles*, with the length of the rectangle reflecting the size of the ORF. The chloroplast-derived gene models are depicted in *green*. Putative GST genes are depicted in *red*. Cytochrome P-450-like proteins are depicted in *yellow*. Putative polyproteins are depicted in *gray*. The two conserved hypothetical gene familes are shown in *light blue* and *brown*. Other hypothetical gene models are in *black*. All other putative gene models are depicted in *blue*. The 2.1-kb direct-repeat sequence flanking the chloroplast insertion is indicated by *black triangles*. The 73 gene models were annotated as follows: 1: putative serine protease; 2, 4 and 5: cytochrome P450-like proteins; 3, 6, 10, 41, 42, 44, 46, 47, 49, 54–56, 58–60, 67 and 70: hypothetical proteins; 7–9, 37–39, 50–53, 57, 61–66, 68, 69, and 71–73: putative GSTs; 11: putative GST pseudogene; 36: putative GST pseudogene, 3′-partial; 12: tRNA-Val; 13: tRNA-Ile; 14: tRNA-Ala; 15: tRNA-Arg; 16: tRNA-Asn; 17: tRNA-Val; 18: NADH dehydrogenase ND3; 19: PSII G protein 5′-partial; 20: NADH-plastoquinone oxidoreductase subunit J; 21: tRNA-Phe; 22: tRNA-Leu; 23: tRNA-Thr; 24: ribosomal protein S4; 25: tRNA-Ser; 26: photosystem I P700 chlorophyll A apoprotein A1; 27: photosystem I P700 chlorophyll A apoprotein A2; 28: ribosomal protein S14; 29: tRNA-Arg; 30: ATPase alpha subunit; 31: ATPase I subunit; 32: ATPase III subunit; 33: ATPase A subunit; 34: ribosomal protein S2; 35: RNA polymerase beta′ subunit-2 5′-partial; 40, 43, and 45: putative polyproteins; 48: putative disease resistance protein

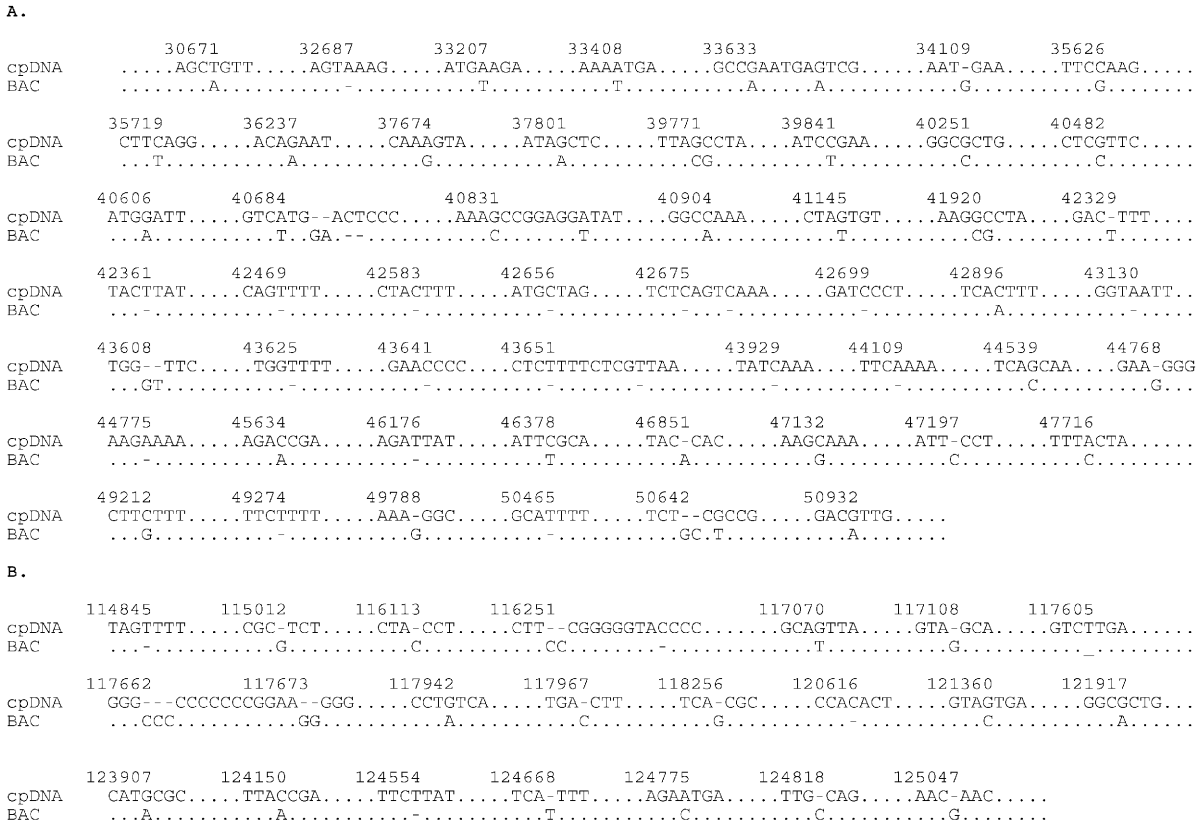

Fig. 2A, B. Sequence comparison between the chloroplast DNA insertion on rice 10L and the corresponding regions in the rice chloroplast genome. **A** The large single-copy region (LSC). **B** Inverted repeat region. The numbers indicate the positions in the rice chloroplast genome sequence (GenBank Accession: NC_001320, gi:11466763, labeled as cpDNA), which differ from their counterparts in OSJNBb0005J14 (labeled as BAC). The sequences in the BAC clone are only shown when they differ from the chloroplast genome sequence

Three other local clusters of GST genes representing an additional 16 GST genes were present in the rice genome; eight GSTs were clustered near the centromere on chromosome 1 at ~73.4 cM, four GSTs were clustered on chromosome 1L at ~170.4 cM, and four GST genes were clustered at ~9.3 cM on chromosome 3S. Thirteen of the 24 GST genes located on 10L are expressed, as revealed by high-identity (over 98%) alignments with rice ESTs, with some genes aligning with a single EST (models 11, 39, 50, and 73) suggesting weak expression and one gene (model 53) aligning with 18 ESTs suggesting a high level of expression in the rice transcriptome. Some of these GSTs were expressed in specific tissues, whereas other GSTs were expressed in multiple tissues. For example, the GSTs coded for in models 7 and 11 aligned with ESTs that were derived from root cDNA libraries, whereas models 39, 50, 57, 72, and 73 aligned with ESTs from shoot/leaf cDNA libraries and model 64 aligned with ESTs from a callus cDNA library. In contrast, GSTs coded for in models 52, 53, 61, 62, and 65 aligned with ESTs derived from multiple cDNA libraries, suggesting a broader range of expression in rice tissues for these GSTs.

A majority of the chloroplast genes were identical to their counterparts in the chloroplast genome. The chloroplast genes located within this insertion include ribosomal protein genes and tRNA genes, as well as photosystem and energy transport-related genes. Only a small fraction of the chloroplast-derived genes were identified using the gene prediction programs utilized in our annotation method. The ATPase A-subunit (gene model 33) was predicted by FGENESH and partially by Genscan+. Three other genes (models 26, 27 and 32) were partially predicted by FGENESH and/or Genscan+. In the 239-kb region described in this study, the average GC content for the chloroplast-derived genes is 41%, whereas the average GC content for nuclear-derived genes is 61%. The gene prediction programs utilized in our annotation pipeline are designed for nuclear genes and thus limited in their ability to identify genes with significantly different nucleotide composition, such as chloroplast-derived genes.

As described above, only 92 bp (out of a total of 32,974 bp) differed between the chloroplast-derived sequence in the insertion and the sequence found in the rice chloroplast (Fig. 2). These nucleotide alterations were distributed throughout the chloroplast insertion – irrespective of coding or non-coding regions – suggesting a lack of preferential accumulation of mutations within the insertion. Of the 92 altered nucleotides, 36% were in intergenic regions, 21% were within introns, and 43%

were in either an exon or an RNA gene. Twenty-four codons were affected by the nucleotide alterations, resulting in a change in the encoded amino acid in 15 cases. However, the position within the codon that was altered was uniform for all three codon positions with 32% of alterations occurring in position 1, 25% in position 2 and 43% in position 3.

The chloroplast insertion was flanked by a 2.1-kb perfect direct repeat (Figs. 1 and 3). A search of Genbank, which included 253 Mb of rice genomic DNA, with the 2.1-kb repeat did not reveal any sequence identity with other sequences, with the exception of the similarity at the end of the repeat to GST transcripts from rice and several other species as shown in Fig. 3. In fact, the chloroplast DNA was inserted into an intron of a GST pseudogene that carried a premature translational stop codon (UAG). As shown in the annotation of the 2.1-kb repeat sequence (Fig. 3), it is likely that the 2.1-kb sequence was duplicated at the time of the insertion, as the centromere-proximal 2.1-kb repeat sequence is flanked by the remaining exons and introns of the GST pseudogene.

Molecular verification of the insertion event

To verify that the insertion of the chloroplast sequence was not a cloning artifact that had occurred during BAC library construction, primers were designed to the junctions (centromere-proximal and telomere-proximal) between the chloroplast insertion and the nuclear DNA. These primers were used to amplify junction products not only from OSJNBb0005J14 but also from two other overlapping BAC clones (OSJNBb0002E24 and OSJNBa0032B13) and rice genomic DNA. Like OSJNBb0005J14, OSJNBb0002E24 was selected from an *Eco*RI-digested rice nuclear DNA library, while OSJNBa0032B13 was from a *Hind*III-digested rice nuclear DNA library (http://www.genome.clemson.edu/where/budiman/index.html). Using primers designed for both junctions, the products of the correct size were amplified from the BAC clones, confirming the correct assembly of OSJNBb0005J14 (data not shown).

We also amplified and sequenced the centromere-proximal and telomere-proximal junction products from rice genomic DNA (*O. sativa* spp. *japonica* var. Nipponbare). The sequences from both junction products verified the insertion of the chloroplast DNA in the rice genome. To ascertain the broader distribution of this insertion in rice species, we amplified and sequenced the centromere-proximal junction from *O. sativa* spp. *indica* var. IR64 DNA. The sequence of the *indica* centromere-proximal junction product (2.8 kb) was nearly identical to the sequence from OSJNBb0005J14, suggesting that the transfer event occurred before the divergence of these two subspecies.

Fiber-FISH analysis of the insertion

We used three clones from the shotgun library of OSJNBb0005J14 that span the chloroplast-derived sequences to confirm the insertion using fiber-FISH with BAC molecules and rice genomic DNA. These clones represent a total of 31,952 bp (96.9%) of the 32,974-bp chloroplast insertion. The DNA isolated from the three shotgun clones was mixed and labeled as a single probe. Fiber-FISH mapping on individual molecules of OSJNBb0005J14 confirmed a single location for this probe (Fig. 4A–C). The fiber-FISH signals derived from the BAC vector and from the shotgun clone probe were separated by two DNA fragments. Based on measurements on five molecules, the three shotgun clones cover 30.7 ± 2.4 kb; the two fragments flanking the shotgun clone probe are approximately 20.9 ± 2.0 kb and 65.3 ± 4.1 kb long, respectively, which is consistent with the sequence assembly data for OSJNBb0005J14.

Fiber-FISH analysis was also conducted using DNA fibers prepared from leaf nuclei of *O. sativa* spp. *japonica* var. Nipponbare. OSJNBb0005J14 was detected in the

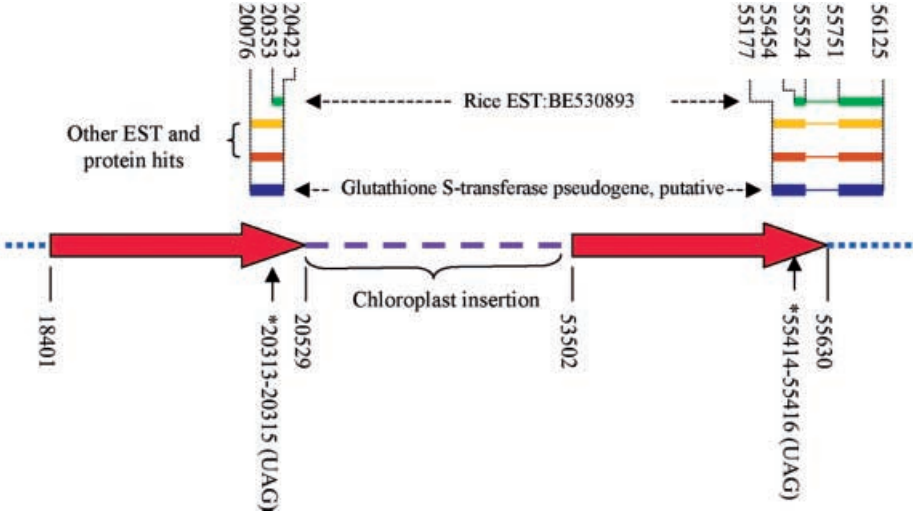

Fig. 3. Alignment of database matches within the 2.1-kb direct repeat. The two copies of the direct repeat are depicted as *red arrows*. The numbers are the coordinates in OSJNBb0005J14. The *solid blue boxes* represent the curated putative GST pseudogene. *Solid green boxes* represent rice EST BE530893, which was nearly identical to the corresponding coding region of the GST pseudogene. *Solid yellow and brown boxes* represent imperfect matches to additional GST ESTs and proteins

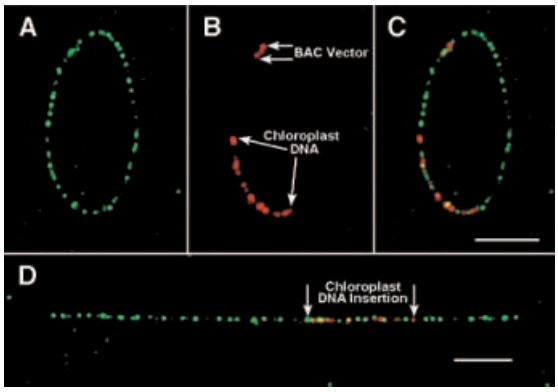

Fig. 4A–D. Fiber-FISH analysis of the chloroplast insertion in BAC OSJNBb0005J14 and in *O. sativa* spp. *japonica* var. Nipponbare. **A–C** Fiber-FISH signal derived from a single OSJNBb0005J14 molecule. **A** Green signals were derived from labeled OSJNBb0005J14 DNA. **B** Red signals were from a BAC vector (pBeloBAC11) probe and three shotgun clones (OTAWA47, OTAWA35, OTAWC79) that span the inserted chloroplast DNA. **C** Merged image of **A** and **B**. **D** A genomic fiber-FISH signal obtained from Nipponbare rice using OSJNBb0005J14 (*green*) and three shotgun clones (*red*) as probes. Bars = 10 μm

fluorescein isothiocyanate channel (FITC, green color) and the three shotgun clones that span the chloroplast insertion were detected in the rhodamine channel (red color) in genomic fiber-FISH experiments (Fig. 4D). We observed fiber-FISH signals with an identical pattern to those derived from individual OSJNBb0005J14 molecules. Based on five measurements, the red signal was approximately 29.3 ± 1.6 kb in length, and the two flanking green signals were 22.5 ± 1.3 kb and 65.1 ± 2.2 kb long, respectively (Fig. 4D)

Discussion

In this report we describe the annotation of 73 genes within a 239-kb region of rice chromosome 10L. This region was gene dense (one gene per 3.25 kb) and contained 24 chloroplast-derived genes and 49 nucleus-derived genes. Of the 49 nuclear genes, 35 (71%) were members of local gene families, including 24 which encode GST. As the chloroplast-derived genes could be due to a cloning artifact, we experimentally verified the insertion using both fiber-FISH and sequence analysis of PCR products from rice genomic DNA and confirmed the authenticity of the chloroplast DNA insertion of ~33 kb (24.5% of the chloroplast genome) into the rice nuclear genome. Based on the high degree of sequence conservation between the rice chloroplast genome and the rice genomic sequences within this region, the transfer of the 33 kb of chloroplast DNA to the nucleus must have occurred recently – prior to the divergence of *indica* and *japonica* about 2–3 million years ago (Huke and Huke 1990).

The mechanism by which the chloroplast DNA inserted into the long arm of rice chromosome 10 is unknown. One hypothesis is that the entire chloroplast genome was inserted into chromosome 10L and then underwent a deletion event, leaving the LSC and IR_A fragments. An insertion of the entire mitochondrial genome has been reported in *Arabidopsis* (Stupar et al. 2001), so transfer of the entire chloroplast genome to the nucleus would not be an unlikely event. Another possibility is that the chloroplast DNA insertion on rice 10L was the result of a random end-joining event, as co-ligated fragments from the LSC and IR_A of the circular chloroplast genome are present in the insertion. Rearranged chloroplast DNA molecules were discovered recently using fiber-FISH analysis (Lilly et al. 2001) and it is possible that the two chloroplast fragments within chromosome 10L were ligated together during the process of chloroplast DNA replication and prior to nuclear integration. Intriguingly, there is a perfect direct repeat (2.1 kb) flanking the chloroplast-derived sequence, which is reminiscent of the mechanism of transposable element insertion in which the target sequence is duplicated during the insertion event. However, no sequence similarity to any known transposable elements was detected within the 2.1-kb direct repeat. In addition, although this region on rice 10L appears to have a high density of locally duplicated genes, no similarity to other rice sequences was detected within the 2.1-kb repeat, with the exception of the similarity to GST genes.

In addition to the nuclear genome, plants have mitochondrial and plastid genomes. The sequences of complete mitochondrial genomes from three flowering plants are now available (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/euk_o.html). Although the rice mitochondrial genome has not been fully sequenced, the *Arabidopsis* mitochondrial genome (387 kb) has been sequenced and codes for 57 genes (Unseld et al. 1997). The chloroplast genome has been sequenced from several plant species (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/euk_o.html). The rice chloroplast genome is 134.5 kb long and codes for 76 proteins (Hiratsuka et al. 1989). According to the endosymbiotic theory, these organellar genomes are descendants of an ancient endosymbiote (for recent reviews see Martin and Herrmann 1998; Lang et al. 1999; Blanchard and Lynch 2000). The ancestor of present-day mitochondria is believed to be closely related to extant α-Proteobacteria. The smallest known genome of an α-proteobacterium is that of *Rickettsia prowazekii*, whose ~1.11-Mb genome codes for ~834 proteins (Andersson et al. 1998). The ancestor of present-day plastids is believed to be similar to the cyanobacterium *Synechocystis* PCC6803, which has a genome of 3.57 Mb that codes for 3168 proteins (Kaneko et al. 1996). With this large differential in genome size, it is believed that substantial portions of the mitochondrial and the plastid genomes have been lost or transferred to the nucleus over evolutionary time. Numerous transfer events involving fragments of mitochondrial and chloroplast DNA have been reported previously in plants (Baldauf and Palmer 1990; Ayliffe and Timmis 1992; Sun and Callis 1993; Blanchard and

Schmidt 1995; Martin et al. 1998; Millen et al. 2001) and an insertion of an extremely large mitochondrial DNA fragment (~620 kb) was recently discovered in the *Arabidopsis* nuclear genome after completion of the sequencing of chromosome 2 (Lin et al. 1999; Stupar et al. 2001). Although the insertion of 33 kb of the rice chloroplast genome into the nuclear genome is much smaller, it does represent a substantial and recent insertion of the chloroplast genome (~24.5%) that has been transferred to the rice nucleus.

Genome sequencing of *Arabidopsis* first revealed the presence of a large insertion of the mitochondrial genome on chromosome 2, and the report in this study of a large insertion of the chloroplast genome in the rice nucleus was also a direct consequence of a genome sequencing project. Not all genome sequencing projects will detect evolutionarily recent organellar insertion events as whole-genome shotgun sequencing projects actively remove organellar sequences from the assembly process to eliminate contaminating organellar sequences. This study suggests that the routine "filtering out" of chloroplast and mitochondrial sequences as contaminants in genome projects should be re-evaluated, as it is apparent that plant nuclear genomes are dynamic in nature and contain recent organellar DNA insertions that cannot be discriminated from true organellar DNA based solely on sequence identity. Continued genome sequencing efforts using a BAC-by-BAC approach will reveal the presence or absence of additional organellar DNA transfer events and provide additional data which should help to dissect the mechanism by which these transfer events occur.

Acknowledgements Funding for the work was provided by grants to C.R.B. from the US Department of Agriculture (99-35317-8275), the National Science Foundation (DBI998282), and the US Department of Energy (DE-FG02-99ER20357). The cytological work was supported by a grant to J.J. from the US Department of Energy (DE-FG02-01ER15266). The authors wish to thank Lowell Umayan, Jeremy Peterson, Hanif Khalak, Martin Shumway, Patee Gesuwan, Qi Yang, Brian Haas, Sam Angiouli, Owen White, M. Heaney, S. Lo, V. Sapiro, B. Lee, J. Shao, S. Gregory, C. Irwin, R. Kramchedu, J. Neubrech, M. Sengamalay and E. Arnold for their bioinformatic and IT support. The critical reading and comments by C. D. Town are greatly appreciated. The kind gift of *O. sativa* spp. *indica* cv. IR64 from Jan Leach is greatly appreciated.

References

Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson A-S, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–143

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (1994) Current protocols in molecular biology. Wiley, New York

Ayliffe MA, Timmis JN (1992) Plastid DNA sequence homologies in the tobacco nuclear genome. Mol Gen Genet 236:105–112

Baldauf SL, Palmer JD (1990) Evolutionary transfer of the chloroplast *tufA* gene to the nucleus. Nature 344:262–265

Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Blanchard JL, Lynch M (2000) Organellar genes: why do they end up in the nucleus? Trends Genet 16:315–320

Blanchard JL, Schmidt GW (1995) Pervasive migration of organellar DNA to the nucleus of plants. J Mol Evol 41:397–406

Goff SA (1999) Rice as a model for cereal genomics. Curr Opin Plant Biol 2:86–89

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

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29:41–43

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (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

Huke RE, Huke EH (1990) Rice: then and now. International Rice Research Institute, Metro Manila, The Philippines

Jackson SA, Wang ML, Goodman HM, Jiang J (1998) Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. Genome 41:566–572

Jackson SA, Dong F, Jiang J (1999) Digital mapping of bacterial artificial chromosomes by fluorescence in situ hybridization. Plant J 17:581–587

Jiang J, Gill BS, Wang GL, Ronald PC, Ward DC (1995) Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. Proc Natl Acad Sci USA 92:4487–4491

Kaneko T, et al (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109–136

Lang BF, Gray MW, Burger G (1999) Mitochondrial genome evolution and the origin of eukaryotes. Annu.Rev.Genet 33:351–397

Lilly JW, Havey MJ, Jackson SA, Jiang J (2001) Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants. Plant Cell 13:245–254

Lin X, et al (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761–768

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol 118:9–17

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermin LS, Wolfe KH (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645–658

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29:159–164

Rujan T, Martin W (2001) How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. Trends Genet 17:113–120

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

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283–290

Sonnhammer EL, Eddy SR, Birne E, Bateman A, Durbin R (1998) Pfam:Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26:320–322

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci USA 98:5099–5103

Sun CW, Callis J (1993) Recent stable insertion of mitochondrial DNA into an *Arabidopsis* polyubiquitin gene by nonhomologous recombination. Plant Cell 5:97–107

Sutton GG, White O, Adams MD and Kerlavage AR (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome 1:9–19

Tettelin H, et al (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293:498–506

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

Unseld M, Marienfeld JR, Brandt P, Brennicke A (1997) The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet 15:57–61

Yu J, et al (2001) A draft sequence of the rice *Oryza sativa* ssp. *indica*) genome. Chin Sci Bull 46:1937–19

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