Mol Genet Genomics (2002) 266: 740–746
DOI 10.1007/s00438-001-0606-9

ORIGINAL PAPER

**Y. Ogihara · K. Isono · T. Kojima · A. Endo
M. Hanaoka · T. Shiina · T. Terachi · S. Utsugi
M. Murata · N. Mori · S. Takumi · K. Ikeo
T. Gojobori · R. Murai · K. Murai · Y. Matsuoka
Y. Ohnishi · H. Tajiri · K. Tsunewaki**

Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA

Received: 17 July 2001 / Accepted: 15 October 2001 / Published online: 22 November 2001

Abstract Structural features of the wheat plastome were clarified by comparison of the complete sequence of wheat chloroplast DNA with those of rice and maize chloroplast genomes. The wheat plastome consists of a 134,545-bp circular molecule with 20,703-bp inverted repeats and the same gene content as the rice and maize plastomes. However, some structural divergence was found even in the coding regions of genes. These alterations are due to illegitimate recombination between two short direct repeats and/or replication slippage. Overall comparison of chloroplast DNAs among the three cereals indicated the presence of some hot-spot regions for length mutations. Whereas the region with clustered tRNA genes and that downstream of *rbcL* showed divergence in a species-specific manner, the deletion patterns of ORFs in the inverted-repeat regions and the borders between the inverted repeats and the small single-copy region support the notion that wheat and rice are related more closely to each other than to maize.

Keywords Chinese Spring wheat · Chloroplast DNA · Complete sequence · Hypervariable region · Structure analysis

Communicated by R. Hagemann

Y. Ogihara (✉) · T. Kojima · A. Endo
Kihara Institute for Biological Research
and Graduate School of Integrated Science,
Yokohama City University, Yokohama 244-0813, Japan
E-mail: ogihara@yokohama-cu.ac.jp
Tel.: +81-45-820-1903
Fax: +81-45-820-1901

K. Isono
PE Biosystems Japan Ltd, Chiba 279-0011, Japan

M. Hanaoka · T. Shiina
Graduate School of Human and Environmental Studies,
Kyoto University, Kyoto 606-8501, Japan

T. Terachi
Department of Biotechnology, Faculty of Engineering,
Kyoto Sangyo University, Kyoto 603-8047, Japan

S. Utsugi · M. Murata
Research Institute for Bioresources,
Okayama University, Kurashiki 710-0046, Japan

N. Mori · S. Takumi
Laboratory of Plant Genetics,
Department of Biological and Environmental Science,
Faculty of Agriculture, Kobe University,
Kobe 657-8501, Japan

K. Ikeo · T. Gojobori
Center for Information Biology,
National Institute of Genetics, Mishima 411-8540, Japan

R. Murai · K. Murai · Y. Matsuoka
Y. Ohnishi · H. Tajiri · K. Tsunewaki
Department of Bioscience,
Fukui Prefectural University, Fukui 910-1195, Japan

Introduction

A characteristic feature of plastomes is the occurrence of long inverted repeats, ranging in size from 10 to 85 kb (Palmer 1991), that separate the rest of the molecule into large and small single-copy regions. It is well known that the structure and gene content of plastomes are conserved among diverse plants. The complete sequencing of chloroplast DNAs from various plants, such as tobacco (Shinozaki et al. 1986), rice (Hiratsuka et al. 1989), maize (Maier et al. 1995), Arabidopsis (Sato et al. 1999), Oenothera (Hupfer et al. 2000), spinach (Schmitz-Linneweber et al. 2001), black pine (Wakasugi et al. 1994), liverwort (Ohyama et al. 1986), and the algae Chlorella (Wakasugi et al. 1997), and Euglena (Hallick et al. 1993), has confirmed the conservative nature of plastome evolution. Although the overall structure of the plastome is thought to be conserved by the stabilizing action of the long inverted repeats, structural alterations in plastomes, such as inversions (Howe et al. 1988; Hiratsuka et al. 1989), translocations (Ogihara et al.

1988), and deletions (Palmer 1991), have been found among angiosperms. Furthermore, sequence analysis of chloroplast DNAs in related plants has allowed the identification of hot spots for length mutations (Ogihara et al. 1988), and this comparative approach is expected to further our understanding of the mechanism(s) of, and better define the traits affected by, plastome evolution.

Wheat is one of the major crops in the Gramineae, and is the most important member of the Triticeae. The genome constitution of each species belonging to the *Triticum-Aegilops* complex and the phylogenetic relationships among them are well defined (Kihara 1954). In addition to the nuclear genome analysis, nucleus-cytoplasm (NC) hybrids have been established by combining tester nuclei from common wheat with the cytoplasms from all other *Triticum-Aegilops* species (Tsunewaki 1993). By cultivating these NC hybrids, the biological effects of alien plasmons can be analyzed precisely. Accordingly, the molecular basis of nucleus-cytoplasm interaction in wheat species is now open for clarification.

In order to analyze the overall structure of the wheat plastome, we have recently completed the sequencing of the wheat chloroplast DNA, which represents the third such plastome to be fully sequenced among the grasses (Ogihara et al. 2000). In that report, we listed the plasmid clones which cover the entire wheat plastome, and the gene content of each. By comparison of the entire sequence of the plastomes of three cereals, namely, wheat, rice and maize, we have now clarified the structural features of the hypervariable regions in the plastomes of these grass plants and the alterations in gene structure between them. The results of this analysis are presented here.

Materials and methods

The entire sequence of wheat chloroplast DNA was determined as previously reported (Ogihara et al. 2000). Nucleotide sequences were retrieved from the DDBJ, EMBL and NCBI nucleotide data banks. Assembly and manipulation of sequences were performed with the GENETYX program (Software Development Co., Tokyo,). Identity searches were carried out with the FASTA (Pearson and Lipman 1988) and BLAST (Altschul et al. 1990) programs. The entire sequences of chloroplast genomes from wheat, rice, maize and tobacco were compared with each other by using the harr plot program (Sonnhammer and Durbin 1995).

Results and discussion

Size and genetic organization of the wheat chloroplast genome

The circular wheat chloroplast DNA is 134,545 bp long, and each of the inverted repeats (IRs) is 20,703 bp in length. The IR regions divide the rest of the sequence into segments of 80,349 bp [the large single-copy (LSC) region] and 12,790 bp [the small single-copy (SSC) region], as shown in Fig. 1. Although the molecular size of wheat chloroplast DNA was previously reported to be 134,540 bp (Ogihara et al. 2000), further examination of sequence traces have led to the revised size of 134,545 bp. The arrangement and locations of the plastome genes are also given in Fig. 1. The number and content of functional wheat chloroplast genes, so far identified, are identical to those of rice and maize (Hiratsuka et al. 1989; Maier et al. 1995).

However, some structural alterations in the chloroplast genes have been recognized even among grass plastomes. These alterations are the result of two processes: replication slippage or intramolecular recombination mediated by simple direct repeats, and simple nucleotide loss during replication (Ogihara et al. 1991). As regards the former, an 81-bp deletion is found in the middle of the wheat *rpoC2* gene, but not in that of rice or maize. Since a duplicated 5-bp oligonucleotide (CTTTT) is located at each end of the deleted portion in the chloroplast DNAs of rice and maize, this 81-bp deletion is assumed to have been caused by recombination via the short direct repeats (Ohnishi et al. 1999). A second example is an 18-bp deletion in wheat *infA*. A tandem duplication of an 18-bp unit is located near the 3′-terminus of the *infA* genes from rice and maize. One of these tandem repeats was lost in the wheat gene. As for the latter process, a 19-bp deletion at the 3′-end of the wheat *rpl22* gene, a 6-bp deletion near the 5′-region of rice *ndhI*, a 3-bp deletion near the 5′-region of wheat *ndhK*, a 2-bp deletion at the 3′-end of wheat *atpA*, and a 1-bp deletion at the 3′-end of rice *atpF* and wheat *ycf5* were detected by comparison of the entire sequences of the chloroplast DNAs of the three grasses. These nucleotide eliminations resulted in deletions of one amino acid (*rpl22* and *ndhK* of wheat), two amino acids (wheat *atpA* and rice *ndhI*), or three amino acids (rice *atpF*), and the addition of one amino acid in one case (wheat *ycf5*). These structural alterations are not located in conserved regions of these genes, and therefore, these changes probably do not affect protein function.

Sequence comparison of the whole plastome among grass plants

The wheat chloroplast genome represents the third to be completely sequenced in grass plants, and also in monocots. Because rice (which provided the first complete sequence of a plastome from monocots; Hiratsuka et al. 1989), maize (second candidate; Maier et al. 1995) and wheat have diverged almost equally among grasses (e.g. Chase et al. 1993), the availability of the complete plastome sequences of these three grass plants should help to clarify the variability of the plastomes during the evolution of grass plants.

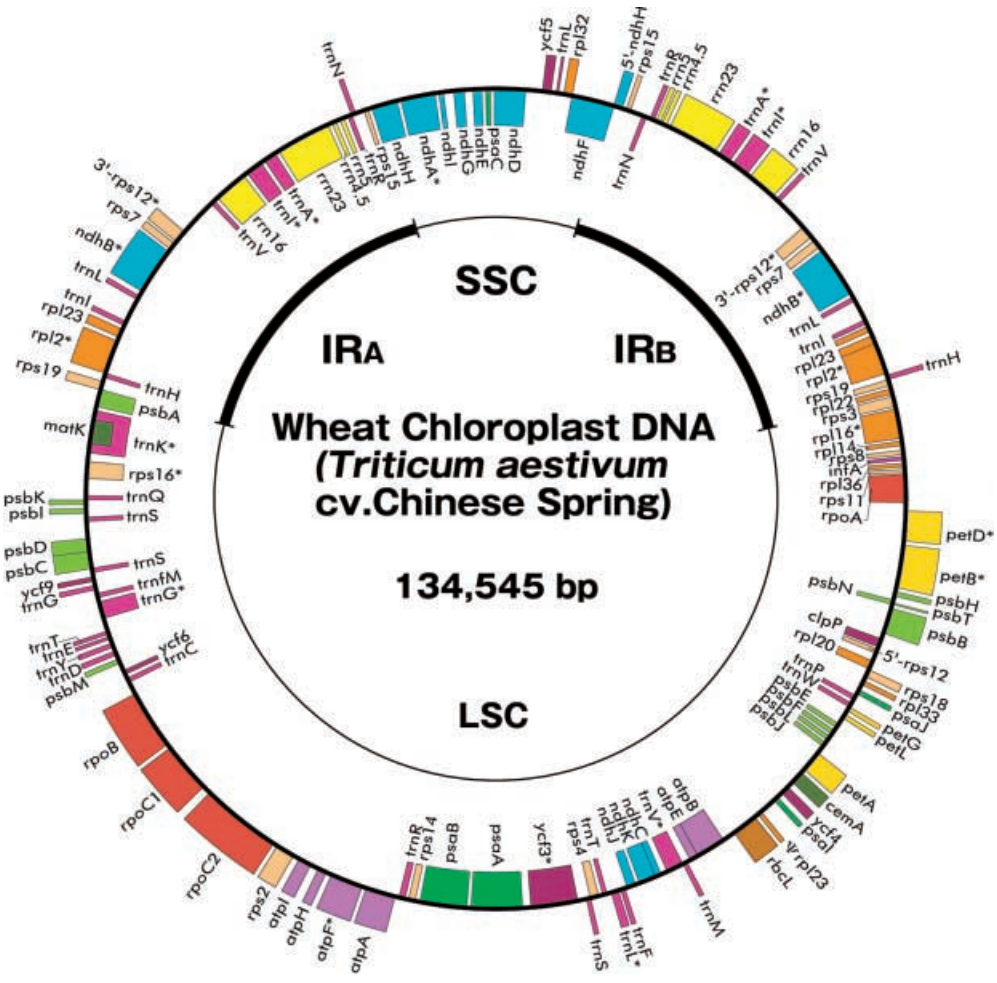

Fig. 1 Organization of the chloroplast genome of common wheat (*Triticum aestivum*) cv. Chinese Spring. Genes depicted on the *outer* rim of the map are transcribed counterclockwise. The *asterisks* indicate genes that harbor introns in their coding sequences

The results of a harr plot analysis comparing the entire sequences of the three grass plastomes are shown in Fig. 2. As pointed out earlier, the wheat plastome does not show extensive genome rearrangements except in the large inverted repeats (Fig. 2A). When the entire sequence of the wheat plastome was compared with that of rice, two remarkably variable regions were discovered (Fig. 2B). One is located in the region approximately 16 kb from the start of the LSC, containing the genes *trnSer(UGA)* to *trnCys(GCA)*, where a number of pseudogenes have accumulated. The other lies downstream of *rbcL*, at ~55 kb, where a hot-spot for length mutations was reported in the wheat complex (Ogihara et al. 1991) and the grass family (Morton and Clegg 1993). Furthermore, comparison of the entire sequence of the wheat plastome to that of maize (Fig. 2C) revealed another structural alteration of this grass plastome in the IR region between *trnIle(CAC)* and *trnLeu(CAA)*. From comparative studies of the grass plastome relative to that of tobacco, three major inversions were found, as shown in Fig. 2D (Howe et al. 1988; Hiratsuka et al. 1989). Thus, with respect to basic genome structure, the plastomes of grass plants are highly conservative, but some hot-spots for plastome structural mutation were observed. In the case of three of these hot-spot regions, precise sequence comparisons among the three grass plastomes from wheat, rice and maize were carried out.

The hot-spot region for length mutations that lies about 16 kb from the start of the LSC, and contains *trnSer(UGA)* to *trnCys(GCA)* is located close to the breakpoint of the first of the three large inversions found in grass plastomes (Howe et al. 1988; Hiratsuka et al. 1989). The first inversion took place during the course of differentiation of the Restionaceae from the Centrolepidaceae (Katayama and Ogihara 1996). These inversions mainly resulted from homologous recombinations mediated by tRNA genes (Hiratsuka et al. 1989). Actually, a number of tRNA genes have accumulated in this hot-spot region (Fig. 3). Although the gene content and arrangement involved in the hot-spot region have been conserved among three grass plastomes, the intergenic regions are hypervariable, i.e., a number of insertions/deletions have accumulated in the region. In addition, the positions of these insertions/deletions varied from species to species (Fig. 3). Moreover, some pseudogenes translocated from the other regions were found in these intergenic regions. These lines of evidence suggest that this hot-spot region harbors unstable sequences and that the clustered tRNA genes might

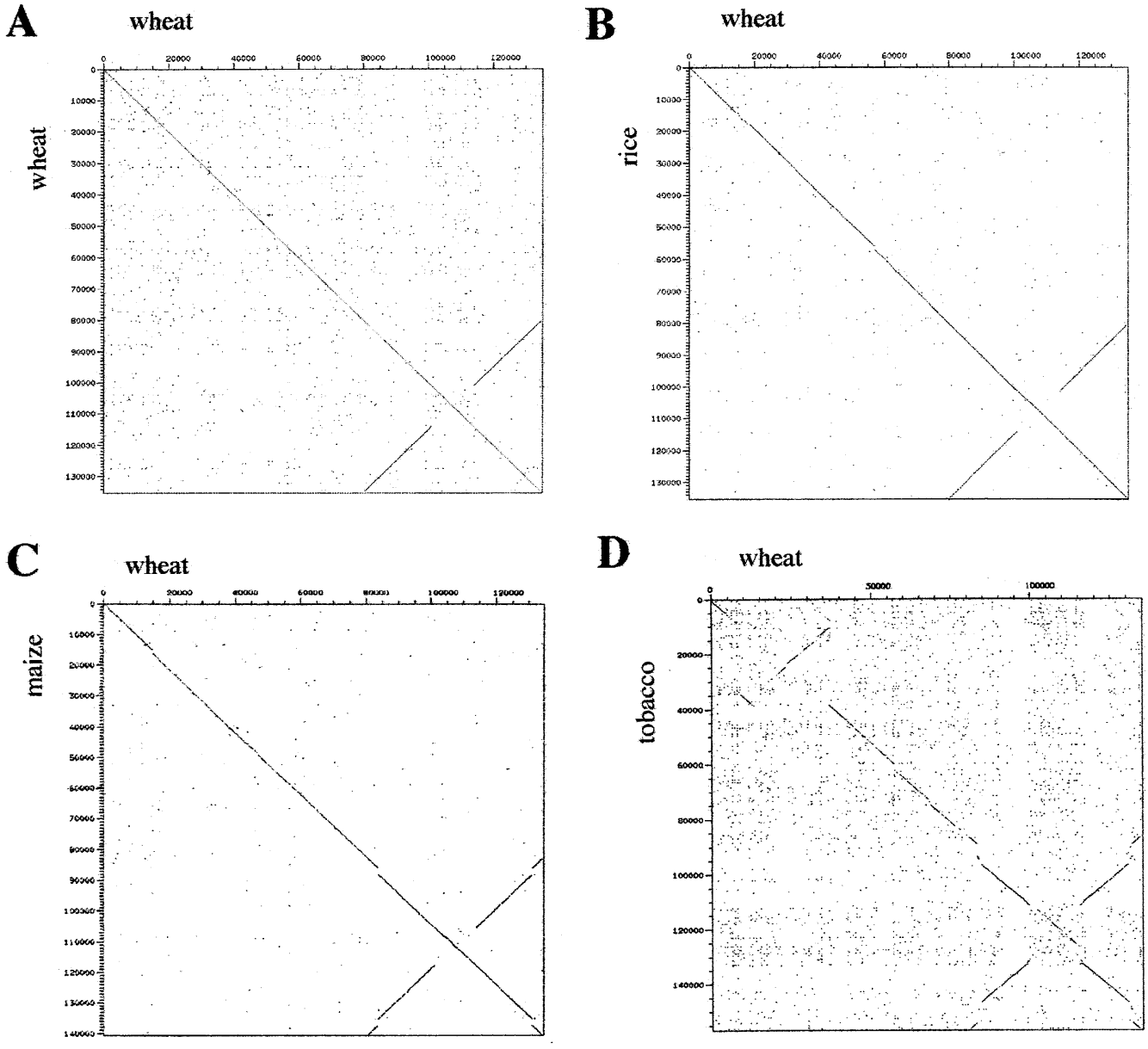

Fig. 2A–D Harr plot analysis of the entire plastomes of wheat, rice, maize and tobacco. The entire sequences of the plastomes were compared by the harr plot method. The comparison was carried out for the plastomes of wheat vs. wheat (**A**), wheat vs. rice (**B**), wheat vs. maize (**C**) and wheat vs. tobacco (**D**)

become targets for illegitimate recombination (Ogihara et al. 1992).

A second hot-spot for length mutation is located in the large single-copy region between the genes *rbcL* and *cemA*. A comparison of the nucleotide sequences with the corresponding region from dicots (Shinozaki et al. 1986; Hupfer et al. 2000; Schmitz-Linneweber et al. 2001) revealed two rearrangements in the cereal plastomes, i.e., deletion of the *accD* gene and non-reciprocal translocation of *rpl23* (Ogihara et al. 1988). When the *accD* gene was traced in the plastomes of monocots, no detectable Southern hybridization was found in the plastome of the progenitor of Poales and Cyperales (Katayama and Ogihara 1996). However, because a remnant of the *accD* gene was found in the rice plastome (Hiratsuka et al. 1989), deletion of the *accD* gene appears to have occurred several times during the course of the evolution of Poaceae. Non-reciprocal translocation of the *rpl23* gene took place recently during the course of differentiation of Poaceae from its direct ancestor, Joinvilleaceae (Katayama and Ogihara 1996). After non-reciprocal translocation of the *rpl23* gene, deletions involving the homologous portion of the *accD* gene have occurred recurrently during the differentiation of the wheat and maize lineages. Another deletion involving a part of the translocated *rpl23* gene occurred in the rice plastome. Furthermore, a number of illegitimate recombinations took place around the translocated *rpl23* genes in the wheat species, so as to make this region

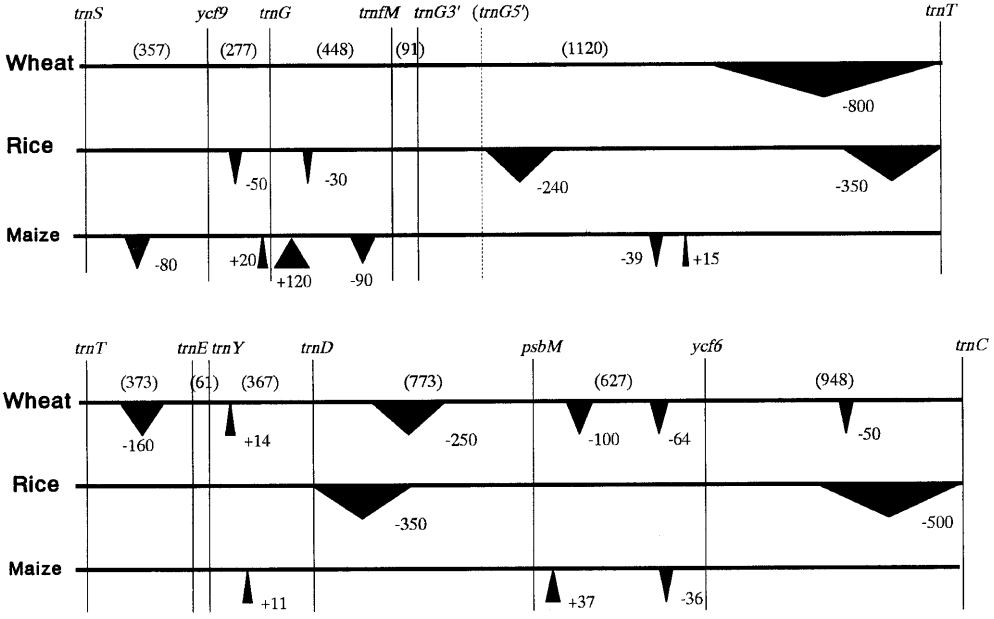

Fig. 3 Sequence comparison of a hot-spot region for length variation between *trnSer(UGA)* and *trnCys(GCA)* in the LSC region. The corresponding segments of the plastomes of wheat, rice and maize are depicted schematically in the alignment. *Upwardly pointing triangles* represent insertions relative to the other sequences; the *downwardly pointing arowheads* stand for deletions. The numbers of nucleotides *between* the genes are given in *parentheses*, and the numbers of altered nucleotides are indicated by + (insertion) or – (deletion)

hypervariable with respect to length mutations (Ogihara et al. 1991). In contrast, little sequence divergence has occurred here in maize, as there were no sequence deviations in the corresponding region (Doebley et al. 1987; Morton and Clegg 1993).

A third hot-spot for length mutations was found in the *ycf2* gene, which is conserved in dicots (Shinozaki et al. 1986; Hupfer et al. 2000; Schmitz-Linneweber et al. 2001), liverwort (Ohyama et al. 1986) and black pine (Wakasugi et al. 1994). Rearrangements of the *ycf2* gene were initiated at the time of the differentiation of the progenitor of Centrolepidaceae, Restionaceae, Joinvelleaceae and Poaceae from Anarthriaceae, because no Southern hybridization signals were detected when these plastomes were probed with tobacco *ycf2* (Katayama and Ogihara 1996). However, comparison of the nucleotide sequences of the corresponding region among higher plants makes it possible to conclude that plastomes of grass plants harbor parts of the *ycf2* gene remaining after its truncation. Gene arrangements corresponding to the *ycf2* gene region of tobacco in the three grass plastomes are shown in Fig. 4. When the nucleotide sequences of the tobacco *ycf2* region and those of grass plants were aligned, at least two deletions were noted: one deletion occurred within *ycf2*, so that two ORFs (ORF46 and ORF34) were generated. Also, the DNA segment including ORF92 was deleted from the grass plastomes, and subsequently a novel ORF (ORF173 for maize and ORF249 for rice and wheat) was produced. After base substitutions and/or minor rearrangements of the sequences, several smaller ORFs were generated in the maize plastomes in place of the large ORF (*ycf2*). The remaining ORFs (ORF46, ORF34, ORF 241, ORF139 and ORF38 of maize) show homology to their counterparts in tobacco. When a sequence comparison of the *ycf2* region within the grass plastomes

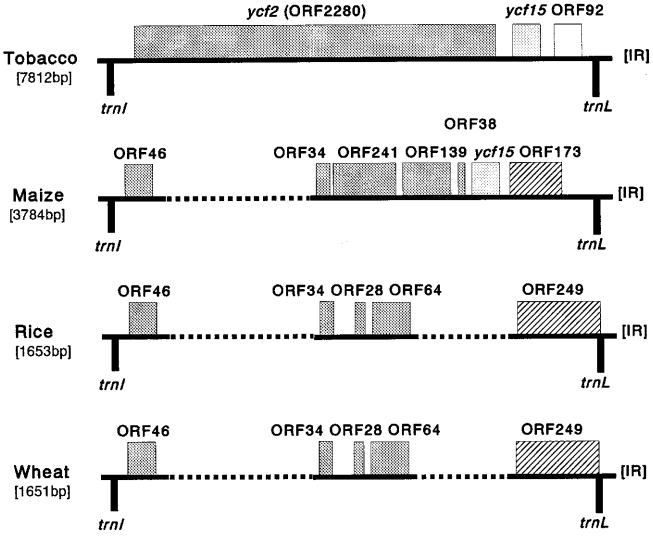

Fig. 4 Comparison of plastome structure around the *ycf2* gene in the IR between grass species. Homologous genes or ORFs are marked by the same shading pattern. ORFs *above* the base line are transcribed from left to right. The *dotted lines* indicate deleted portions in each plastome

was conducted, the region containing ORF139, ORF38 and *ycf15* was found to have been eliminated from the plastomes of both wheat and rice (Fig. 4). Since the remaining region containing ORFs (ORF28 and ORF64) in wheat and rice showed high homology with ORF241 of maize, it is plausible to assume that the rearrangement was a simple deletion event. The portions deleted from the wheat and rice plastomes were almost identical, suggesting that wheat and rice plastomes shared a common progenitor after the deletion involving this region had occurred. It is still unclear whether or not these ORFs are functional in grass plastomes.

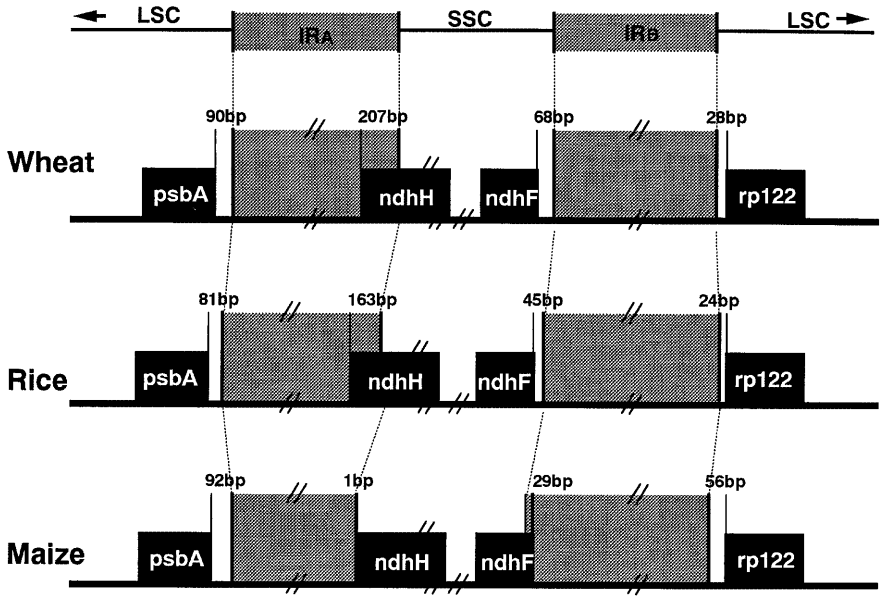

Fig. 5 Comparison of the positions of the junctions between IR regions and single copy (SC) regions in the plastomes of wheat, rice and maize. The numbers of base-pairs between the border of the IR and the adjacent genes are given

Junctions between IR and SC regions in the three cereal plastomes

The borders between the two inverted repeats (IR_A and IR_B) and the two single-copy regions (LSC and SSC) vary greatly among plant species, although the nucleotide sequences in the IR regions are conserved (Palmer 1991; Maier et al. 1995). Accordingly, ebb and flow of inverted repeats have brought about expansions and reductions in plastome genome sizes (Goulding et al. 1996). In Fig. 5, the exact positions of the IR borders in each of the three cereal plastomes are compared. Whereas the gene contents and arrangements in these regions of the three cereal plastomes correspond, the border positions are not identical. In all cases, the borders between the IR and LSC were located within the intergenic region between *rps19* (inside the IR) and *psbA* or *rpl22* located in the LSC. On the other hand, the border positions between IR_A and SSC lie in the middle of the coding region of the *ndhH* gene in wheat and rice. However, the corresponding border in the maize plastome has shifted to the initiation codon of the *ndhH* gene, so that almost all of the coding region of *ndhH* in maize is located in the SSC. Similarly, the *ndhF* gene of wheat and rice is located entirely within the SSC region, whereas the border between SSC and IR_B in maize has shifted so as to fall within the *ndhF* gene. The characteristic features of the maize plastome with regard to the borders between IR and SSC are inferred to have arisen due to intramolecular recombination (Maier et al. 1995). These structural features of cereal plastomes demonstrate that wheat and rice are more closely related to one another than to maize.

In conclusion, structural features were compared among three cereal plastomes. Two hypervariable regions, i.e., a region with clustered tRNA genes and another region downstream *of rbcL*, represented typical variation types among the three plastomes. However, the comparison of the deletion pattern in the *ycf2* gene and of the junctions between IR and SSC clearly showed that wheat and rice are more related to one another than to maize. This phylogenetic relationship is supported by previously published sequence data from some limited regions (Ogihara et al. 1991; Chase et al. 1993; Ohnishi et al. 1999). Phylogenetic comparisons based on the entire sequence of chloroplast DNA from wheat, rice and maize are now being conducted.

Acknowledgements This work was supported by a Grant-in-Aid (No. 10309008) from the Ministry of Education, Science, Sports, and Culture of Japan. The complete sequence of the chloroplast DNA of Chinese Spring wheat has been deposited in the DNA Data Bank of Japan (DDBJ) under the Accession No. AB042240.

References

Altschul FA, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Chase MW, et al (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann Miss Bot Gard 80:528–550

Doebely J, Renfroe W, Blanton A (1987) Restriction site variation in the *Zea* chloroplast genome. Genetics 117:139–147

Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Curr Genet 252:195–206

Hallick RB, Hong L, Drager RG, Favreau MR. Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21:3537–3544

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

Howe CJ, Barker RF, Bowman CM, Dyer TA (1988) Common features of three inversions in wheat chloroplast DNA. Curr Genet 13:343–349

Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, Sears BB (2000) Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable Euoenothera plastomes. Mol Gen Genet 263:581–585

Katayama H, Ogihara Y (1996) Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. Curr Genet 29:572–581

Kihara H (1954) Consideration on the evolution and distribution of *Aegilops* species based on the analyzer-method. Cytologia 19:336–357

Maier RM, Neckermann K, Igloi GL, Kössel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251:614–628

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

Ogihara Y, Terachi T, Sasakuma T (1988) Intramolecular recombination of chloroplast genome mediated by a short direct-repeat sequence in wheat species. Proc Natl Acad Sci USA 85:8573–8577

Ogihara Y, Terachi T, Sasakuma T (1991) Molecular analysis of the hot spot region related to length mutations in wheat chloroplast DNAs. I. Nucleotide divergence of genes and intergenic spacer regions located in the hot spot region. Genetics 129:873–884

Ogihara Y, Terachi T, Sasakuma T (1992) Structural analysis of length mutations in a hot-spot region of wheat chloroplast DNA. Curr Genet 22:251–258

Ogihara Y, et al (2000) Chinese Spring wheat (*Triticum aestivum* L.) chloroplast genome: complete sequence and contig clones. Plant Mol Biol Rep 18:243–253

Ohnishi Y, Tajiri H, Matsuoka Y, Tsunewaki K (1999) Molecular analysis of a 21.1-kbp fragment of wheat chloroplast DNA bearing RNA polymerase subunit (*rpo*) genes. Genome 42:1042–1049

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

Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK (eds) The molecular biology of plastids. Academic Press, San Diego, pp 5–53

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

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

Schmitz-Linneweber C, Maier RM, Alcaraz JP Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol Biol 45:307–315

Shinozaki K, et al (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–10

Tsunewaki K (1993) Genome-plasmon interaction in wheat. Jpn J Genet 68:1–34

Wakasugi T, Tsudzuki J, Ito S, NakashimaK, Tsudzuki T, Sugiura M (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thumbergii*. Proc Natl Acad Sci USA 91:9794–9798

Wakasuge T, Nagai T, Kapoor M, Sugita M, Ito M, Itso S, Tsudzuki J, Nakashima K, Tsudzuki T, Suzuki Y, Hamada A, Ohta T, Inamura A, Yoshinaga K, Sugiura M (1997) Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulagaris*: The existence of genes possibly involved in chloroplast division. Proc Natl Acad Sci USA 94:5967–5972