Complete Nucleotide Sequence of the Cotton (*Gossypium barbadense* L.) Chloroplast Genome with a Comparative Analysis of Sequences among 9 Dicot Plants

Rashid Ismael Hag Ibrahim^{1,2*}, Jun-Ichi Azuma¹ and Masahiro Sakamoto¹

¹Graduate School of Agriculture, Kyoto University, Kyoto, 606-8502, Sakyu-ku, Kitashirakawa Oiwake-cho, Japan.

²Khartoum University, Faculty of Science, Botany Department, P. O. Box 321, P. C. 11115, Khartoum, Sudan.

(Received 13 May 2006, accepted 1 September 2006)

Recently, the complete chloroplast genome sequences of many important crop plants were determined, and this can be considered a major step forward toward exploiting the usefulness of chloroplast genetic engineering technology. Economically, cotton is one of the most important crop plants for many countries. To further our understanding of this important crop, we determined the complete nucleotide sequence of the chloroplast genome from cotton (*Gossypium barbadense* L.). The chloroplast genome of cotton is 160,317 base pairs (bp) in length, and is composed of a large single copy (LSC) of 88,841 bp, a small single copy (SSC) of 20,294 bp, and two identical inverted repeat (IR) regions of 25,591 bp each. The genome contains 114 unique genes, of which 17 genes are duplicated in the IRs. In addition, many open reading frames (ORFs) and hypothetical chloroplast reading frames (*ycfs*) with unknown functions were deduced. Compared to the chloroplast genomes from 8 other dicot plants, the cotton chloroplast genome showed a high degree of similarity of the overall structure, gene organization, and gene content. Furthermore, the sequences of the genes showed high degrees of identity at the DNA and amino acid levels. The cotton chloroplast genome was somewhat longer than the chloroplast genomes of most of the other dicot plants compared here. However, this elongation of the cotton chloroplast genome was found to be due mainly to expansions of the intergenic regions and introns (non-coding DNA). Moreover, these expansions occurred predominantly in the LSC and SSC regions.

Key words: Chloroplast DNA, Cotton, *Gossypium barbadense*

INTRODUCTION

The genus *Gossypium* L. comprises plants known as cotton, and includes about 50 species. The word cotton itself refers only to the four common cultivated species of the genus. *Gossypium arboreum* L. and *Gossypium herbaceum* L. are the two diploid cultivated species with the chromosome number $2n = 26$, and are known as Old World cotton (Afro-Asian). The other two cultivated species, *Gossypium hirsutum* L. (Upland cotton) and *Gossypium barbadense* L. (Sea Island cotton) are allotetraploids with the chromosome number $2n = 52$, and are known as New World cotton (American). The chromosome size, chromosome structure, chromosome pairing behavior, and relative fertility of inter-specific hybrids are useful

genetic typing tools and were used to group the genus *Gossypium* L. into eight diploid genome groups, designated A through G, in addition to K, and one allopolyploid genome group, which are widely distributed in the tropical areas of the world (Stewart, 1995).

Cytogenetically, the allotetraploid genome contains one genome similar to that of the Old World diploid A-genome and another genome similar to the one of the New World diploid D-genome (Endrizzi et al., 1985).

The genus *Gossypium* L., including both the diploid and allotetraploid cottons, has a chloroplast DNA (cpDNA) that is uni-parentally and especially maternally inherited. Furthermore, the allotetraploid cotton, AD-genome, has a chloroplast genome like that of the A-genome from the Old World diploid cotton (Wendel, 1989).

The complete sequences of the plastid genomes of many plants have been determined, and cover the major lin-

Edited by Toru Terachi

* Corresponding author. E-mail: rashid@kais.kyoto-u.ac.jp

eages, with the best representation from flowering plants, including monocot plants, dicot plants, gymnosperms, psilotophytes, bryophytes and algae. Also the genomic sequences of the apicoplast of some apicomplexans were determined (www.ncbi.nih.gov/genomes/organelles/plastids_tax.html). Comparative studies revealed that chloroplast genomes of higher plants are well conserved regarding gene content, gene order, and general structure (Palmer, 1991). The cpDNA was reported to be present in different topological forms (Oldenburg and Bendich, 2004). Structurally, it is generally believed to be a quadripartite double-stranded circle of DNA, which has an LSC region and an SSC region separated by two identical IR regions. The total length of the cpDNA ranges from 120 to 160 kb in higher plants (Sugiura, 1995; Gaut, 1998). Since they have lost most of the IR regions, conifers and some legumes are exceptions regarding this phenomenon (Tsudzuki et al., 1992).

The chloroplast genomes from many agricultural crop plants were sequenced, mainly from the cereal group; rice, corn, wheat, and sugar-cane (Hiratsuka et al., 1989; Maier et al., 1995; Ogiwara et al., 2002; Asano et al., 2004; Calsa et al., 2004). Cotton is the most important textile fiber in the world and it is the source of many other by-products, including cooking oil, and cellulose-derived products, and is used as animal fodder. Also cotton is grown in more than 90 countries and has a strong impact on their economies (Kumar et al., 2004). Thus the objective of this study was to sequence the chloroplast genome of cotton *Gossypium barbadense* L., as a dicot and a very important agricultural crop plant. We thereby aimed in the long run to facilitate future developments regarding cotton production and to encourage cotton improvement through chloroplast genetic engineering technology. The advantages of chloroplast genetic engineering include high-level transgene expression due to multiple chloroplast genomes per chloroplast and many chloroplasts per cell (DeCosa et al., 2001), transgene containment and prevention of gene flow via maternal inheritance (Daniell et al., 1998; Hagemann, 2004), and avoidance of gene silencing (Dhingra et al., 2004), undesirable foreign DNA (Daniell et al., 2004), position effect (Daniell, 2002), and pleiotropic effects (Lee et al., 2003) due to position-specific insertion of the transgene.

This manuscript had been finished when the complete sequence of cpDNA from *Gossypium hirsutum* L. was published (Lee et al., 2006). So a general comparison has been done, which showed very high identity and similarity between the two allotetraploid cotton species, *Gossypium hirsutum* L. and *Gossypium barbadense* L.

MATERIALS AND METHODS

Plant material Cotton plants (*Gossypium barbadense* L.) were grown under natural conditions in the experi-

mental farm of the Graduate School of Agriculture, Kyoto University, and Nippon Shinyaku Co., LTD, Kyoto, Japan.

DNA extraction Total genomic DNA was extracted from young and fully expanded leaves using the Plant Genomic DNA Extraction Miniprep System (Viogene, USA). The protocol of the manufacturer was followed and the extracted DNA was used as a template for PCR amplification (usually 0.5 to 1 µl).

Primers design The primer-walking strategy was adopted for this study. Primers were manually designed based on the tobacco cpDNA sequence as a reference (Shinozaki et al., 1986). Primers were designed to amplify cpDNA fragments ranging in size from 500 bp to 1800 bp.

PCR protocols Chloroplast DNA of cotton was amplified with the use of 1.25 units of the high-fidelity KOD Dash polymerase (TOYOBO, Japan) and suitable primers in final volumes of 25 µl in 0.2 ml tubes. A Bio-Rad iCycler Thermal Cycler (USA) was used to carry out the amplification reactions. Different PCR protocols were adopted, including:

Standard PCR 94°C for 2 minutes as a first denaturation step, followed by 35 cycles at 94°C for 30 seconds, 50–60°C (depending on the primer pair) for 2 seconds for annealing of primers and 74°C for 30–90 seconds (depending on the expected length of the PCR product) as an extension step. This was ended by a final extension at 74°C for 5 minutes.

Long PCR 94°C for 2 minutes as a first denaturation step, followed by 35 cycles at 94°C for 30 seconds, 50–60°C (depending on the primer pair) for 2 seconds, and 74°C for 120–180 seconds (depending on the expected length of the PCR product). The final extension was performed at 74°C for 5 minutes.

Touchdown PCR 94°C for 2 minutes as a first denaturation step, followed by 15 cycles at 94°C for 30 seconds, annealing of primers at 65–70°C (depending on the primer pair) for 2 seconds, and incubation at 74°C for 30–90 seconds (depending on the expected length of the PCR product) for extension. That was followed by 30 cycles at 94°C for 30 seconds, 50–60°C (depending on the primer pair) for 2 seconds for annealing of primers, and 74°C for 30–90 seconds (depending on the expected length of the PCR product). The final extension was performed at 74°C for 5 minutes.

Nested PCR Some of the long PCR products were used as templates to generate shorter PCR products. In these

cases the standard PCR protocol was followed.

Cloning and sequencing of PCR products The Wizard® SV Gel and PCR Clean – Up System (Promega, USA) was used to purify all PCR products. The purified PCR products were cloned using pGEM®-T Easy Vector System I (Promega, USA). DH5 α competent cells were used as the hosts for cloned DNA. Plasmid DNAs were extracted from colonies, and were confirmed to contain inserts using a plasmid DNA extraction kit “MagExtractor-Plasmid- (TOYOBO, Japan)”. DNA sequencing reactions were carried out by the modified dideoxy chain termination method using an ABI 373 DNA sequencer (Applied Biosystems, USA).

Data analysis The resultant sequences were analyzed using GENETYX software (GENETYX, Tokyo, Japan) and the Basic Local Alignment Search Tool (BLAST) at the National Center for Biotechnology Information website (Altschul et al. 1990).

RESULTS AND DISCUSSION

Overall Structure The overall structure, gene content, gene number and gene organization of the chloroplast genomes from different higher plant species are well conserved (Sugiura, 1995; Martin et al., 1998). However, micro- and macro-structural rearrangements exist in some chloroplast genomes, for example, small inversions (Hiratsuka et al., 1989), insertions and/or deletions (Ogihara et al., 1991; Kanno et al., 1993; Maier et al., 1995), base substitutions (Morton and Clegg, 1995), and translocations (Ogihara et al., 1988), as well as large inversions in the LSC regions in *Oenothera elata* (Hupfer et al., 2000) and *Lotus japonicus* (Kato et al., 2000).

The complete chloroplast genome of cotton is 160,317 bp in size and has the general quadripartite structure similar to the sequenced chloroplast genomes of the flowering plants group. It is composed of an LSC of 88,841 bp, an SSC of 20,294 bp, and a pair of identical IRs of 25,591 bp each, as shown in Fig. 1. At least 114 putative functional genes were annotated from the sequence, which is similar to the number of genes harbored by the cpDNA of *Nicotiana tabacum* (Shinozaki et al., 1986). In addition, many open reading frames (ORFs) and hypothetical chloroplast reading frames (*ycfs*) with unknown functions were deduced. The genes encoded by the cotton chloroplast genome are listed in Table 1.

Introns As shown in Table 2, the cotton chloroplast DNA possesses longer LSC and SSC regions than most of the other 8 dicot plants. This elongation can mainly be attributed to the expansions of intergenic regions and introns present in the LSC and SSC regions. Intron classification depends on the intron conserved-boundary

sequences, which play a crucial role in intron splicing, and the RNA folding patterns (Cech, 1990). The boundary sequences of the introns that were found in the cpDNA from cotton showed high identities when aligned with those of the plants under comparison. The introns in the chloroplast genomes belong predominantly to self-splicing group II, except in the case of the *trnL* (UAA) gene, which possesses a group I intron (Sugiura, 1992). In 17 annotated genes in cotton chloroplast DNA, the total number of introns was 20, which was similar to the number in most dicot plants investigated; only 3 genes, *ycf3*, *clpP*, and *rps12*, had 2 introns each. Fourteen introns are present in the LSC region, and these introns in cotton are longer than the introns in tobacco as a reference plant (Shinozaki et al., 1986). Six out of the 14 introns are longer in cotton than their counterparts in all the other dicot plants compared (Table 3). Furthermore, 4 of the 5 introns present in the IR regions in the *rpl2*, *ndhB*, *trnI* (GAU), and *trnA* (UGC) genes are longer in cotton than in tobacco, and 2 introns of the *rpl2* and *trnI* (GAU) genes are longer in cotton compared to their counterparts in the other 8 dicot plants. An exception is the intron of *3'rps12*, which is the same size as the one in tobacco. The only short intron in cotton cpDNA is the only intron in the SSC region, the intron of the *ndhA* gene, which means that the elongation of the SSC region is due only to elongations of the intergenic regions. These differences in introns and intergenic regions of the LSC and the SSC regions are consistent with the findings of previous studies, which showed that the LSC and the SSC regions have three times faster divergence than the IR regions (Maier et al., 1995; Sugiura, 1995).

Pseudo- and True Genes Some genes may exist as pseudo-genes in chloroplast genomes. For instance, *rpl23*, which encodes a protein component of the large ribosomal subunit, is present in *Gossypium barbadense* and many other plant species, while it is a pseudo-gene in *Spinacia oleracea* and has been substituted by a nuclear functional gene (Thomas et al., 1988; Bubuneko et al., 1994; Yamaguchi and Subramanian, 2000). The *infA* gene, which encodes an initiation factor protein, is present as a pseudo-gene in *Gossypium barbadense*, which is consistent with its presence in *Nicotiana tabacum* (Shinozaki et al., 1986) and *Atropa belladonna* (Schmitz-Linneweber et al., 2002). Millen and his colleagues (2001) demonstrated many parallel losses of the *infA* gene from the chloroplasts of many plants and its transfer to the nucleus: this gene is absent from the cpDNA of *Arabidopsis thaliana* (Sato et al., 1999), *Oenothera elata* (Hupfer et al., 2000) and *Lotus japonicus* (Kato et al., 2000). Other genes have been lost from the chloroplast genomes of some plants, for example, *sprA*, some ribosomal protein genes, *rpl22*, *rpl32*, *rps16*, some *ndh* genes, and *accD*. The lost genes might be trans-

Fig. 1. Gene organization of the chloroplast genome from cotton (*Gossypium barbadense* L.). Genes shown outside the circle are transcribed counterclockwise, while those located inside are transcribed clockwise. Intron-containing genes are indicated by asterisks (*). Genes for transfer RNAs are represented by the 1-letter code of amino acids with anticodons. When two genes overlap, the one that is located downstream or inside the other gene is displayed with a lower-height box.

Table 1. Genes annotated in the cotton (*G. barbadense*) chloroplast genome

Photosynthesis related genes	
RuBisCO large subunit:	<i>rbcL</i> .
Photosystem I genes:	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> , <i>psaI</i> , <i>psaJ</i> .
Assembly/stability of photosystem I:	<i>ycf3**</i> , <i>ycf4</i> .
Photosystem II genes:	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i> (<i>ycf9</i>).
Cytochrome <i>b/f</i> complex genes:	<i>petA</i> , <i>petB*</i> , <i>petD*</i> , <i>petG</i> , <i>petL</i> , <i>petN</i> .
c-type cytochrome:	<i>ccsA</i> (<i>ycf5</i>).
ATP synthase genes:	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF*</i> , <i>atpH</i> , <i>atpI</i> .
NADH dehydrogenase genes:	<i>ndhA*</i> , <i>ndhB**§</i> , <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i> .
Transcription and translation related genes	
RNA polymerase and related genes:	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1*</i> , <i>rpoC2</i> .
Ribosomal protein genes:	<i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7§</i> , <i>rps8</i> , <i>rps11</i> , <i>rps12**§</i> , <i>rps14</i> , <i>rps15</i> , <i>rps16*</i> , <i>rps18</i> , <i>rps19</i> , <i>rpl2*</i> , <i>rpl14</i> , <i>rpl16*</i> , <i>rpl20</i> , <i>rpl22</i> , <i>rpl23§</i> , <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i> .
RNA genes	
Ribosomal RNA genes:	<i>rrn23§</i> , <i>rrn16§</i> , <i>rrn5§</i> , <i>rrn4.5§</i> .
Transfer RNA genes:	<i>trnA</i> (UGC)*, <i>trnC</i> (GCA), <i>trnD</i> (GUC), <i>trnE</i> (UUC), <i>trnF</i> (GAA), <i>trnG</i> (GCC), <i>trnG</i> (UCC)*, <i>trnH</i> (GUG), <i>trnI</i> (CAU)§, <i>trnI</i> (GAU)*, <i>trnK</i> (UUU)*, <i>trnL</i> (CAA)§, <i>trnL</i> (UAA)*, <i>trnL</i> (UAG), <i>trnfM</i> (CAU), <i>trnM</i> (CAU), <i>trnN</i> (GUU)§, <i>trnP</i> (UGG), <i>trnQ</i> (UUG), <i>trnR</i> (ACG)§, <i>trnR</i> (UCU), <i>trnS</i> (GCU), <i>trnS</i> (GGA), <i>trnS</i> (UGA), <i>trnT</i> (GGU), <i>trnT</i> (UGU), <i>trnV</i> (GAC)§, <i>trnV</i> (UAC)*, <i>trnW</i> (CCA), <i>trnY</i> (GUA).
Others	
Maturase	<i>matK</i> .
Acetyl-CoA carboxylase subunit:	<i>accD</i> ,
ATP-dependent protease subunit:	<i>clpP**</i> .
Inorganic carbon uptake:	<i>cemA</i> (<i>ycf10</i>),
Open reading frames	
Conserved reading frames (<i>ycfs</i>):	<i>ycf1</i> , <i>ycf2§</i> , <i>ycf15§</i> , ORF40, ORF62, ORF77, ORF185, ORF230.
Non-conserved open reading frames (ORFs):	ORF55, ORF131, ORF49, ORF71, ORF61, ORF113, ORF86.

* Gene contains one intron.

** Gene contains two introns.

§ Gene present as a duplicate in the IR regions.

Table 2. Comparison of cotton (*Gossypium barbadense*) chloroplast genome with the chloroplast genomes from 8 dicot plants

Plant	cpDNA* (bp)	LSC (bp)	SSC (bp)	IR (bp)
<i>Gossypium barbadense</i>	160 317	88 841	20 294	25 591
<i>Nicotiana tabacum</i>	155 939	86 686	18 571	25 341
<i>Atropa belladonna</i>	156 688	68 868	18 008	25 906
<i>Arabidopsis thaliana</i>	154 478	84 170	17 780	26 264
<i>Cucumis sativus</i>	155 293	86 650	18 267	25 188
<i>Spinacia oleraceae</i>	150 725	82 719	17 860	25 073
<i>Panax schinseng</i>	156 318	86 106	18 070	26 071
<i>Lotus japonicus</i>	150 519	81 936	18 271	25 156
<i>Oenothera elata</i>	159 443	89 393	14 436	27 807

* (cpDNA) Chloroplast DNA.

Table 3. Lengths of introns detected in *Gossypium barbadense* (G.b.), *Atropa belladonna* (A.b.), *Arabidopsis thaliana* (A.t.), *Cucumis sativus* (C.s.), *Lotus japonicus* (L.j.), *Nicotiana tabacum* (N.t.), *Oenothera elata* (O.e.), *Panax schinseng* (P.s.) and *Spinacia oleracea* (S.o.) chloroplast DNA

Plant	G.b	A.b	A.t	C.s	L.j	N.t	O.e	P.s	S.o
Intron	Intron Length (bp) ^a								
<i>trnK</i>	2535	2519	2559	2497	2627	2526	2470	2524	2496
<i>rps16</i>	870	822	865	896	891	860	862	887	875
<i>trnG</i>	763	690	715	586	710	691	799	697	708
<i>atpF</i>	805	715	739	739	754	695	759	730	765
<i>rpoC1</i>	753	737	791	788	757	738	740	756	756
<i>ycf3-2</i>	789	763	787	749	695	783	777	758	759
<i>ycf3-1</i>	777	739	714	746	–	738	716	716	778
<i>trnL</i>	582	497	512	549	570	503	520	507	304
<i>trnV</i>	609	572	599	604	599	571	599	578	595
<i>clpP-2</i>	679	622	539	335	640	637	–	632	592
<i>clpP-1</i>	890	799	891	805	799	807	–	771	839
<i>petB</i>	821	759	804	785	824	753	771	783	776
<i>petD</i>	754	742	709	726	728	742	756	751	743
<i>rpl16</i>	1135	1019	1056	1129	1079	1020	1104	944	954
<i>rpl2</i>	688	664	682	665	687	666	662	660	–
<i>ndhB</i>	683	679	685	686	686	679	681	678	674
<i>3'rps12-2</i>	536	535	537	540	530	536	546	536	536
<i>trnI</i>	959	727	729	826	721	707	947	945	733
<i>trnA</i>	795	681	801	802	806	709	796	808	819
<i>ndhA</i>	1076	1150	1080	1138	1274	1148	1041	1023	1079

^a Longest introns are in bold.

(–) Absence of the intron.

ferred to the nucleus, and their protein products imported back via a chloroplast signal (Gantt et al., 1991). The *sprA* gene, which encodes a small plastid RNA of 218 bp and has been proposed to play a role in 16S rRNA maturation, seems to be absent in cotton, as was reported for some vascular plants, excluding tobacco, tomato and deadly nightshade plants (Vera and Sugiura, 1994; Sugita et al., 1997; Schmitz-Linneweber et al., 2002). Two ribosomal protein genes were reported to be lost from the chloroplast genomes of some plant species. The first gene is *rpl22*, which encodes a ribosomal protein component of the large subunit and is absent from legumes, including *Lotus japonicus*, (Gantt et al., 1991; Kato et al., 2000), while it is present in *Gossypium barbadense*, as in most other vascular plants. The second is the *rps16* gene, which encodes a ribosomal protein component of the small subunit and has been lost, as have all the NADH (*ndh*) dehydrogenase genes, from *Pinus thunbergii* (Wakasugi et al., 1994), *Marchantia polymorpha*, *Psilotum nudum* (Wakasugi et al., 1998) and *Physcomitrella patens* (Sugiura et al., 2003), while it was reported in *Gossypium barbadense*. The *accD* gene, which encodes the β subunit of prokaryotic-type acetyl-CoA carboxylase, was

not reported to be present in the cpDNA of cereals (Wakasugi et al., 2001), but was annotated in cotton. The *clpP* gene, which has two introns in tobacco but no intron in the monocots, nor in *Oenothera elata* or *Pinus thunbergii*, and which encodes the proteolytic subunit of ATP-dependent protease, was also annotated in cotton.

Hypothetical Chloroplast Reading Frames (*ycfs*)

Conserved open reading frames (*ycfs*) with unknown functions were reported in many chloroplast genomes of higher plants and algae (Stoebe et al., 1998). The chloroplast genome of cotton contains 7 *ycf* genes. Among them, there are 2 genes essential for plant survival (Martin et al., 1998; Drescher et al., 2000), *ycf1* and *ycf2*, which lack eubacterial orthologues. Four *ycfs* tentatively named depending on the number of codons they contain (ORF62, ORF77, ORF185, and ORF230) showed 92.5, 75.6, 90.6, and 90.1% identity at the nucleotide level, respectively, and exhibited 83.6, 33.3, 90.8, and 86% identity at the amino acid level, respectively, compared with their counterpart sequences, ORF62, ORF71, ORF185, and ORF230 from *Nicotiana tabacum* (Shinozaki et al., 1986; Stoebe et al., 1998). Another conserved open read-

Fig. 2. A comparison of *ycf15* among *Gossypium barbadense* (GB), *Cucumis sativus* (CS), *Oenothera elata* (OE), *Arabidopsis thaliana* (AT) and *Spinacia oleraceae* (SO). The inverted and complementary repeats are boxes 1 and 1'. The complementary and inverted incomplete sequences between cotton and cucumber are shown in box 2. The direct inserts in cotton are in boxes 3 and 4.

Table 4. Comparisons of *ycf15* conserved open reading frame from *Gossypium barbadense* with its counterparts from *Arabidopsis thaliana* (A.t.), *Cucumis sativus* (C.s.), *Oenothera elata* (O.e.), and *Spinacia oleraceae* (S.o.)

Plant	Nucleotide Homology (%)	<i>Gossypium barbadense</i>					
		Transitions		Transversions		Inserts (bp)	Deletions (bp)
		A↔G	T↔C	R→Y	Y→R		
A.t.	95.2	9	5	4	5	11+8+5	1
C.s.	93.1	8	13	8	4	11+5+5	—
O.e.	93.2	5	10	13	8	11+5+5	—
S.o.	95.1	3	2	2	4	11+51+5	4+1

R→Y from purine to pyrimidine.

Y→R from pyrimidine to purine.

ing frame named ORF40 showed 85.4% nucleotide identity and 67.7% amino acid identity with its counterpart ORF32 in *Spinacia oleraceae* (Schmitz-Linneweber et al., 2001). The unique-featured conserved open reading frame *ycf15* in cotton contains the short intron-like intervening sequence that was observed in many other plants, such as *Arabidopsis thaliana*, *Spinacia oleraceae*, *Oenothera elata*, and *Zea mays* (Schmitz-Linneweber et al., 2001). Inter-specific comparison of this intervening sequence revealed a high frequency of direct repeats. There are two direct repeat inserts of 11 bp (TATG-GATAATA) and 5 bp (TTCTA) in cotton. Also, many small inverted repeats were found in this *ycf15*. A well-conserved inverted and complementary repeat of 7 bp (AGAATT) length was found in all species examined. Near this inverted repeat there is an incomplete 21 bp (ATCCATACAT AGTGTTTTGA T) inversion between *Gossypium barbadense*, *Arabidopsis thaliana*, *Oenothera*

elata and *Spinacia oleraceae* on one side and *Cucumis sativus* on the other side (Fig. 2). This raises a question about the number and length of direct repeats and/or small inverted repeats, which are known to cause inserts or/and deletions through slipped-strand mis-pairing and/or illegitimate recombination, and their role in eliminating such sequences from other plants (Ogihara et al., 1988; Milligan et al., 1989; Nimzyk et al., 1993). This *ycf15* intervening sequence has been considered to be ancient (Schmitz-Linneweber et al., 2001) and has been eliminated by an unknown mechanism from many other plants such as *Nicotiana tabacum*, *Atropa belladonna*, *Cuscuta reflexa*, *Panax schinseng* and *Epifagus virginiana*. Except for some species-specific inserts or/and deletions, and nucleotide substitutions, this *ycf15* intron-like sequence shows high identity when compared among *Gossypium barbadense*, *Arabidopsis thaliana*, *Spinacia oleraceae*, *Oenothera elata*, and *Cucumis sativus*

either in pairs or as a group (Table 4).

Open Reading Frames (ORFs) Chloroplast genomes contain another kind of open reading frames of unknown functions. Their positions, lengths, and sequences are less conserved among different species of plants (Maier et al., 1995). In tobacco cpDNA, 11 such ORFs, which are at least 70 codons in size, were reported (Shinozaki et al., 1986). Only 6 of them were annotated and well con-

served in the related species *Atropa belladonna* (Schmitz-Linneweber et al., 2002). No homologues were found for the 11 ORFs in the other 7 species investigated (Schmitz-Linneweber et al., 2001, and this study). Compared to the tobacco ORFs, some inter-specific conservation was observed among the 11 ORFs, which is reflected by a high degree of sequence homology in different species, but they appear either reduced in size, fragmented, or both. As shown in Table 5, cotton is not an exception.

Table 5. Comparison of the 11 ORFs encoded by the tobacco, *Nicotiana tabacum* (N.t.) chloroplast DNA with those of other eight dicot plants including *Atropa belladonna* (A.b.), *Lotus japonicus* (L.j.), *Panax schinseng* (P.s.), *Spinacia oleracea* (S.o.), *Oenothera elata* (O.e.), *Arabidopsis thaliana* (A.t.), *Cucumis sativus* (C.s.) and cotton *Gossypium barbadense* (G.b.)

Intergenic Region	Plant Species								
	N.t.	A.b.	L.j.	P.s.	S.o.	O.e.	A.t.	C.s.	G.b.
<i>trnS(UGA)~ycf9</i>	ORF105 (37249) ^a	ORF58 (37129)	ORF74 (25351)	—	ORF43 (34438)	ORF21 (29525)	ORF57 (35594) ORF17 (35659)	—	—
<i>ycf3~trnS(GGA)</i>	ORF74 (46248)	ORF30 (46140) ORF35a (46014)	—	ORF90 (46480)	ORF27a (43271) ORF25 (43956)	ORF41 (20534)	—	—	—
<i>trnT(UGU)~trnL(UAA)</i>	ORF70A (48941)	ORF35b (48693)	ORF69 (14170)	—	ORF27b (45962)	ORF19 (17403)	ORF20 (46900)	—	ORF55 (49873) ORF131 (49897) ORF49 (67211) ORF71 (67405)
<i>petA~psbJ</i>	ORF99 (66176)	ORF99 (66428)	—	—	—	—	—	—	—
<i>psbE~petL</i>	ORF103 (67277)	ORF80 (67597)	—	ORF64 (67076)	—	—	—	—	—
<i>ycf15~trnL(CAA)</i>	ORF92 (96119) ORF115 (96060)	ORF92 (96385) ORF115 (96326)	— —	— —	ORF39a (91354) ORF39b (91354) ORF24 (91478)	ORF48 (99291) ORF71 (99430) ORF29 (99378) ORF26a (99560)	ORF39 (94034) ORF22 (94155) ORF27 (94213)	— —	ORF61 (98630)
<i>trnL(CAA)~ndhB</i>	ORF79 (96556)	ORF79 (96822)	—	—	ORF57 (91797) ORF38 (91967)	ORF22 (99693) ORF56 (99768)	—	ORF78 (145753)	—
<i>3' rps12~trnV(GAC)</i>	ORF70B (102102) ORF131 (101951)	ORF70B (102401) ORF131 (102250)	— —	ORF98 (101067) —	ORF47 (97056) ORF54a (97186) ORF54b (97503)	ORF48 (105379) ORF25 (105419) ORF26b (105419) ORF49 (100586)	— ORF36 (100191) ORF42 (100253)	—	ORF113 (104737) ORF86 (104818) —
<i>trnN(GUU)~ORF350</i>	ORF75 (110597)	ORF75 (110920)	—	—	ORF26 (106151)	ORF26c (114851)	—	—	—

^a The precise positions of the respective start codons for ORFs are given in parenthesis and were retrieved from the NCBI (National Center of Biotechnology Information) website.

Comparison with the cpDNA from *Gossypium hirsutum* L. The comparison of the cpDNA sequence from *Gossypium barbadense* L. with the recently published cpDNA sequence from *Gossypium hirsutum* L. (Lee et al., 2006) showed a very highly conserved gene content, gene order, similarity of sequences and total length (160,317 and 160,301 bp, respectively). However, a quick survey revealed some micro-structural differences, such as transitions, transversions and insertions/deletions (indel), and one macro-structural difference as an inversion of the SSC. It is noteworthy that two major insertion/deletions were found. One seems to be an indel of 51 bp as a direct repeat in the cpDNA sequence from *Gossypium hirsutum* L., which makes it longer in this part. This indel is located in the intergenic spacer of *petN* (*ycf6*) and *psbM*. The other indel detected has slightly complicated features, including many direct and inverted repeats with a short total loss in the cpDNA sequence from *Gossypium hirsutum* L. This indel is located in the intergenic spacer of *psbZ* (*ycf9*) and *trnG* (GCC). These indel results are consistent with results obtained by cpDNA PCR-RFLP of 4 cotton species, including *Gossypium barbadense* L. and *Gossypium hirsutum* L. (unpublished data). Now we are performing a detailed comparison of the cpDNA from the 2 allotetraploid cotton species, including re-sequencing of the different parts.

Cotton is known globally as the most economically important crop, and because of its strong impact on the economy of many nations, especially in developing countries, and also because of its unique feature as the only natural fiber-producing plant. To contribute to a better understanding of this important commercial crop we have presented here the complete chloroplast nucleotide sequence of cotton (*Gossypium barbadense* L.).

An additional aim of fundamental and developmental studies of sequenced plastid genomes is crop improvement. It is known that high-quality fiber of cotton comes from the allotetraploid group, especially *Gossypium barbadense*, which has the highest-quality of fiber. Furthermore, all allotetraploid cottons have the same chloroplast genome from the A-genome group of the diploid cottons (Wendel, 1989), which are *Gossypium arboreum* and *Gossypium herbaceum*. Therefore, we decided to sequence the cpDNA from *Gossypium barbadense*, as it is the source of the highest-quality cotton. On the other hand, *Gossypium barbadense* was considered to represent the cpDNA sequence from the whole group of cultivated cottons, which includes the other allotetraploid species, *Gossypium hirsutum*, and the 2 diploid species, *Gossypium arboreum* and *Gossypium herbaceum*, in addition to the 3 related wild species from the allotetraploid group, *G. mustelinum*, *G. darwinii*, and *G. tomentosum*. Since the cpDNA sequence from *G. hirsutum* has already been published, the cpDNA sequence from *G. barbadense* can be considered to be an additional

representative of the cultivated cotton species and to fulfill the need for cpDNA genome sequences from the allotetraploid cultivated cottons.

Finally, our hope and expectations is that the cotton cpDNA sequence will be valuable for the future of chloroplast biotechnology, transformation and genetic engineering, which in turn may have an impact on the quality and quantity of cotton production. This may, as a final aim, have some beneficial influence on the economy of many cotton-dependent communities around the world.

This work was supported by a Grant-in-Aid (No 020518) from the Ministry of Education, Science, Sports, and Culture of Japan. Cotton seeds were a gift from Nippon Shinyaku Co., LTD (Kyoto, Japan) to whom we express sincere gratitude. We would like to thank Prof. Hiroaki Shimada (Tokyo University of Science) for reading the manuscript. The complete sequence of the chloroplast DNA of cotton (*Gossypium barbadense* L.) has been deposited in the DNA Data Bank of Japan (DDBJ) and will appear in the DDBJ/EMBL/GenBank nucleotide sequence databases with the accession No AP009123.

REFERENCES

- Altschul, F. A., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Asano, T., Tsudzuki, T., Takahashi, S., Shimada, H., and Kadowaki, K. (2004) Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. *DNA Res.* **11**, 93–99.
- Bubunenko, M. G., Schmidt, J., and Subramanian, A. R. (1994) Protein substitution in chloroplast ribosome evolution: A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. *J. Mol. Biol.*, **240**, 28–41.
- Calsa, T. J., Carraro, M. D., Benatti, M. R., Barbosa, A. C., Kitajima, J. P., and Carrer, H. (2004) Structural features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast genome. *Curr. Genet.* **46**, 366–373.
- Cech, T. R. (1990) Self-splicing and enzymatic activity of an intervening sequence RNA from *Tetrahymena*. *Angew. Chem. Int. Ed. Engl.* **29**, 759–768.
- Daniell, H., Datta, R., Varma, S., Gray, S., and Lee, S. B. (1998) Containment of herbicide resistance through genetic engineering of the chloroplast genome. *Nat. Biotechnol.* **16**, 345–348.
- Daniell, H. (2002) Molecular strategies for gene containment in transgenic crops. *Nat. Biotechnol.* **20**, 581–586.
- Daniell, H., Cohill, P. R., Kumar, S., and Dfourmantel, N. (2004) Chloroplast genetic engineering. In: *Molecular Biology and Biotechnology of Plant Organelles* (eds.: H. Daniell and C. Chase), pp. 443–490. Springer Publishers, Dordrecht, The Netherlands.
- De Cosa, B., Moar, W., Lee, S. B., Miller, M., and Daniell, H. (2001) Overexpression of the *Bt* cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. *Nat. Biotechnol.* **19**, 71–74.
- Dhingra, A., Portis, A. R., and Daniell, H. (2004) Enhanced translation of a chloroplast-expressed *RbcS* gene restores small subunit levels and photosynthesis in nuclear *RbcS* antisense plants. *Proc. Natl. Acad. Sci. USA* **101**, 6315–

- 6320.
- Drescher, A., Ruf, S., Calsa, T. J., Carrer, H., and Bock, R. (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* **22**, 97–104.
- Endrizzi, J. E., Turcotte, E. L., and Kohel, R. J. (1985) Genetics, cytology, and evolution of *Gossypium*. *Adv. Genet.* **23**, 271–375.
- Gantt, J. S., Baldauf, S. L., Calie, P. J., Weeden, N. F., and Palmer, J. D. (1991) Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J.* **10**, 3073–3078.
- Gaut, B. S. (1998) Molecular clocks and nucleotide substitution rates in higher plants. In: *Evolutionary Biology* (eds.: Hecht, M. K.), vol. 30, pp.93–120. Plenum Press, New York.
- Hagemann, R. (2004) The sexual inheritance of plant organelles. In: *Molecular Biology and Biotechnology of Plant Organelles* (eds.: H. Daniell and C. Chase), pp. 93–113. Springer Publishers, Dordrecht, The Netherlands.
- Hiratsuka, J., Shimada, H., Whittier, R., et al. (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gen. Genet.* **217**, 185–194.
- Hupfer, H., Swiatek, M., Hornung, S., Herrman, R. G., Maier, R. M., Chiu, W. L., and Sears, B. (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.
- Kanno, A., Watanabe, N., Nakamura, I., and Hirai, A. (1993) Variation in chloroplast DNA from rice (*Oryza sativa*): Differences between deletions mediated by short direct-repeat sequences within a single species. *Theor. Appl. Genet.* **86**, 579–584.
- Kato, T., Kaneko, T., Sato, S., Nakamura, Y., and Tabata, S. (2000) Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res.* **7**, 323–330.
- Kumar, S., Dhingra, A., and Daniell, H. (2004) Stable transformation of the cotton plastid genome and maternal inheritance of transgenes. *Plant Mol. Biol.* **56**, 203–216.
- Lee, S. B., Kwon, H. B., Kwon, S. J., et al. (2003) Accumulation of trehalose within transgenic chloroplasts confers drought tolerance. *Mol. Breeding* **11**, 1–13.
- Lee, S. B., Kaittanis, C., Jansen, R. K., Hostetler, J. B., Tallon, L. J., Twon, C. D., and Daniell, H. (2006) The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. *BMC Genomics* **7**, 61 (doi: 10.1186/1471-2164-7-61).
- Maier, R. M., Neckermann, K., Igloi, G. L., and 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.
- Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M., and Kowallik, K. V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162–165.
- Millen, R. S., Olmstead, R. G., Adams, K. L., et al. (2001) Many parallel losses of *infA* from chloroplast DNA during Angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* **13**, 645–658.
- Milligan, B. G., Hampton, J. N., and Palmer, J. D. (1989) Dispersed repeats and structural reorganization in subcloroplast DNA. *Mol. Biol. Evol.* **6**, 355–368.
- Morton, B. R., Clegg, M. T. (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J. Mol. Evol.* **41**, 597–603.
- Nimzyk, R., Schöndorf, T., and Hachtel, W. (1993) In-frame length mutations associated with short tandem repeats are located in unassigned open reading frames of *Oenothera* chloroplast DNA. *Curr. Genet.* **23**, 265–270.
- Ogihara, Y., Terachi, T., and Sasakuma, T. (1988) Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proc. Natl. Acad. Sci. USA* **85**, 8573–8577.
- Ogihara, Y., Terachi, T., and 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., Isono, K., Kojima, T., et al. (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Mol. Genet. Genomics* **266**, 740–746.
- Oldenburg, D. J., and Bendich, A. J. (2004) Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *J. Mol. Biol.* **335**, 953–970.
- Palmer, J. D. (1991) Plastid chromosomes: Structure and evolution. In: *The Molecular Biology of Plastids* (eds.: Bogorad, L. and Vasil, I. K.), pp.5–53. Academic Press, San Diego.
- Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* **6**, 283–290.
- Schmitz-Linneweber, C., Maier, R. M., Alcaraz, J. P., Cottet, A., Herrmann, R. G., and Mache, R. (2001) The plastid chromosome of spinach (*Spinacia oleracea*): Complete nucleotide sequence and gene organization. *Plant Mol. Biol.* **45**, 307–315.
- Schmitz-Linneweber, C., Regel, R., Du, T. G., Hupfer, H., Herrmann, R. G., and Maier, R. M. (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: The role of RNA editing in generating divergence in the process of plant speciation. *Mol. Biol. Evol.* **19**, 1602–1612.
- Shinozaki, K., Ohme, M., Tanaka, M., et al. (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* **5**, 2043–2049.
- Stewart, J. McD. (1995) Potential for crop improvement with exotic germplasm and genetic engineering. In: *Challenging the Future: Proceedings of the World Cotton Research Conference-1* (eds.: G.A. Constable and N. W. Forrester), pp. 313–327. CSIRO, Melbourne, Australia.
- Stoebe, B., Martin, W. and Kowallik, K. V. (1998) Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. *Plant Mol. Biol. Rep.* **16**, 243–255.
- Sugita, M., Svab, Z., Maliga, P., and Sugiura, M. (1997) Targeted deletion of *sprA* from the tobacco plastid genome indicates that the encoded small RNA is not essential for pre-16S rRNA maturation in plastids. *Mol. Gen. Genet.* **257**, 23–27.
- Sugiura, M. (1992) The chloroplast genome. *Plant Mol. Biol.* **19**, 149–168.
- Sugiura, M. (1995) The chloroplast genome. *Essays Biochem.* **30**, 49–57.
- Sugiura, C., Kobayashi, Y., Aoki, S., Sugita, C., and Sugita, M. (2003) Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: Evidence for the loss and relocation of

- rpoA* from the chloroplast to the nucleus. *Nucleic Acids Res.* **31**, 5324–31.
- Thomas, F., Massenet, O., Dorne, A. M., Briat, J. F., and Mache, R. (1988) Expression of the *rpl23*, *rpl2* and *rps19* genes in spinach chloroplasts. *Nucleic Acids Res.* **16**, 2461–2472.
- Tsudzuki, J., Nakashima, K., Tsudzuki, T., et al. (1992) Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: Nucleotide sequence of *trnQ*, *trnK*, *psbA*, *trnI* and *trnH* and the absence of *rps16*. *Mol. Gen. Genet.* **232**, 206–214.
- Vera, A., and Sugiura, M. (1994) A novel RNA gene in the tobacco plastid genome: Its possible role in the maturation of 16S rRNA. *EMBO J.* **13**, 2211–2217.
- Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., and Sugiura, M. (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc. Natl. Acad. Sci. USA* **91**, 9794–9798.
- Wakasugi, T., Nishikawa, A., Yamada, K., et al. (1998) Complete nucleotide sequence of the plastid genome from a fern, *Psilotum nudum*. *Endocytobiosis Cell Res.* **13** (Suppl.), 147.
- Wakasugi, T., Tsudzuki, T., and Sugiura, M. (2001) The genomics of land plant chloroplasts: Gene content and alteration of genomic information by RNA editing. *Photosynthesis Res.* **70**, 107–118.
- Wendel, J. F. (1989) New World tetraploid cotton contains Old World cytoplasm. *Proc. Natl. Acad. Sci. USA* **86**, 4132–4136.
- Yamaguchi, K., and Subramanian, A. R. (2000) The plastid ribosomal proteins (2): Identification of all the proteins in the 50S subunit of an organelle ribosome (chloroplast). *J. Biol. Chem.*, **275**, 28466–28482.