The Complete Chloroplast Genome Sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): Evolutionary Comparison of *Cephalotaxus* Chloroplast DNAs and Insights into the Loss of Inverted Repeat Copies in Gymnosperms

Xuan Yi[1,†], Lei Gao[1,†], Bo Wang[1], Ying-Juan Su[2,3,*], and Ting Wang[1,*]

[1]CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, China

[2]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China

[3]Institute for Technology Research and Innovation of Sun Yat-sen University, Zhuhai, Guangdong, China

*Corresponding authors: E-mail: tingwang@wbgcas.cn, tingwangtingwang@gmail.com; suyj@mail.sysu.edu.cn.

[†]These authors contributed equally to this work.

Accepted: March 12, 2013

Data deposition: The complete cp genome of *Cephalotaxus oliveri* has been deposited at GenBank under the accession KC136217.

Abstract

We have determined the complete chloroplast (cp) genome sequence of *Cephalotaxus oliveri*. The genome is 134,337 bp in length, encodes 113 genes, and lacks inverted repeat (IR) regions. Genome-wide mutational dynamics have been investigated through comparative analysis of the cp genomes of *C. oliveri* and *C. wilsoniana*. Gene order transformation analyses indicate that when distinct isomers are considered as alternative structures for the ancestral cp genome of cupressophyte and Pinaceae lineages, it is not possible to distinguish between hypotheses favoring retention of the same IR region in cupressophyte and Pinaceae cp genomes from a hypothesis proposing independent loss of IR_A and IR_B. Furthermore, in cupressophyte cp genomes, the highly reduced IRs are replaced by short repeats that have the potential to mediate homologous recombination, analogous to the situation in Pinaceae. The importance of repeats in the mutational dynamics of cupressophyte cp genomes is also illustrated by the accD reading frame, which has undergone extreme length expansion in cupressophytes. This has been caused by a large insertion comprising multiple repeat sequences. Overall, we find that the distribution of repeats, indels, and substitutions is significantly correlated in *Cephalotaxus* cp genomes, consistent with a hypothesis that repeats play a role in inducing substitutions and indels in conifer cp genomes.

Key words: chloroplast genome sequence, *Cephalotaxus oliveri*, chloroplast DNA isomer, inverted repeats, *accD*, cupressophyte.

Introduction

Cephalotaxus oliveri Masters (Cephalotaxaceae) is a vulnerable conifer endemic to China (Fu et al. 1999). It is a dioecious woody shrub or small tree up to 4 m in height, with leaves densely arranged on leafy shoots (Fu et al. 1999). Mainly based on the distinctive characters of leaf morphology and anatomy as well as the alkaloid composition, Fu (1984) has suggested separation of *C. oliveri* from the other *Cephalotaxus* species, establishing a new section, Sect. *Pectinatae*. This treatment has been supported by subsequent studies such as the anatomy of the secondary phloem of the stem (Hu and Shao 1986), embryonic development (Li et al.1986), karyomorphology (Gu et al. 1998), pollen morphology and exine ultrastructure (Xi 1993), and molecular phylogenetic analyses (Wang et al. 2002). Recently, the chloroplast (cp) genome of *C. wilsoniana* has been sequenced, which provides

fresh insights into the loss of inverted repeat (IR) copies from conifer cp genomes and the influence of high heterotachous cp genes on the reconstruction of gymnosperm phylogeny (Wu, Wang, et al. 2011). Considering the unique status of *C. oliveri* in the genus *Cephalotaxus*, we have further sequenced the entire cpDNA of *C. oliveri* to perform a fine comparative analysis of *C. wilsoniana* and *C. oliveri* cp genomes.

The existence of large IRs may confer on the circular cp genome a number of physical properties—formation of head-to-head dimers, copy correction between the IR segments, resistance to intramolecular recombinational loss, reversal of the relative orientation of the single copy sequences, and stabilization of the cp genome against rearrangement (Kolodner and Tewari 1979; Gillham et al. 1985; Palmer 1983). Nevertheless, these potential functional implications do not prevent the variation of IRs. On the contrary, IRs represent hotspots for structural rearrangements within the cp genome (Wicke et al. 2011), because they are frequently subject to expansion, contraction, or even complete loss as mentioned earlier. Through this process known as "the ebb and flow," IRs can easily gain or loose genes from the neighboring single copy regions (Goulding et al. 1996). Goulding et al. (1996) has proposed that short IR expansions may be the result of gene conversion, whereas large IR expansions need to involve a double-strand DNA break. As for IR contractions, they can simply be explained by a deletion of DNA from one copy of the IR, either within the IR or across one of the IR/single-copy boundaries. For example, Wu, Wang, et al. (2011) suggest that deletions of genes including *ycf2* and *psbA* have led to the IR loss from the cpDNAs of *Cedrus deodara* and *C. wilsoniana* and that each has lost a different IR copy (either IR_A or IR_B). More recently, Lin et al. (2012) has further suggested that the deletion of one copy of *ycf2* has caused the IR contraction of *Ginkgo biloba*. However, the evolutionary mechanism underlying the dynamic changes of IRs remains to be elucidated. In particular, Yi et al. (2012) stress that the IR is under stronger evolutionary constraint than the large single copy (LSC) or small single copy (SSC) as shown by the distribution of repeated elements and indels and the base substitution patterns.

Currently, the information available on cp genomes remains insufficient for elucidating the general evolutionary mechanisms or patterns of the genome (Yi et al. 2012). In this study, we have 1) determined the entire cp genome sequence of *C. oliveri*; 2) compared the overall gene content and genomic structure of *C. oliveri* cp genome with those of other gymnosperms, detailing differences between *C. oliveri* and *C. wilsoniana*, which represent the two sections in the same genus *Cephalotaxus*; 3) explored the roles of cpDNA isomers that are generated from the IR-facilitated flip-flop recombination in the dynamic changes of gymnospermous IRs; and 4) investigated whether a genome-wide association of repeats, indels, and substitutions occurs as has been previously reported in angiosperms (Ahmed et al. 2012).

Materials and Methods

Plant Materials and DNA Extraction

Young fresh leaves of *C. oliveri* were collected from a single individual growing in Wuhan Botanical Garden, Chinese Academy of Sciences. Two grams of the leaves was washed and weighed, and total DNA was extracted following the protocol of Su et al. (1998). The purity of the extracted DNA was determined by calculating the optical density 260/280 ratio. The DNA sample with a 260/280 ratio greater than 1.8 was collected and stored at −20 °C for further experiments.

cp DNA Sequencing and Genome Assembly

The cpDNA fragments of *C. oliveri* were amplified by use of polymerase chain reaction (PCR). Because the *C. wilsoniana* cp genome had not been published when this experiment was being conducted, degenerate PCR primers were designed from alignments of coding sequences (CDSs) in 13 gymnosperms (*Cryptomeria japonica*, NC_010548; *Picea sitchensis*, NC_011152; *Ephedra equisetina*, NC_011954; *Welwitschia mirabilis*, NC_010654; *Gnetum parvifolium*, NC_011942; *Pinus thunbergii*, NC_001631; *Keteleeria davidiana*, NC_011930; *Pseudotsuga sinensis* var. *wilsoniana*, NC_016064; *Cathaya argyrophylla*, NC_014589; *Larix decidua*, NC_016058; *P. morrisonicola*, NC_016069; *Ced. deodara*, NC_014575; *Cycas taitungensis*, NC_009618), two ferns (*Alsophila spinulosa*, NC_012818; *Adiantum capillus-veneris*, NC_004766), and one angiosperm (*Amborella trichopoda*, NC_005086).

PCR reactions were carried out in a total volume of 50 μl, containing 2.5 ng of DNA template, 5 μl 10×LA PCR Buffer II (Mg^{2+} Plus), 8 μl deoxyribonucleoside triphosphates mixture (each 2.5 mM), 2.5 U of LA Taq (TaKaRa Bio Inc, Dalian, China), and 2 μl each of forward and reverse primers (10 μM). After an initial denaturation step at 97 °C for 3 min, PCRs were performed in a S1000 Thermal Cycler (Bio-Rad, Hercules, CA) programmed as follows: 35 cycles of denaturation at 94 °C for 30 s and combined annealing and elongation at 62 °C for 3 min. The samples underwent a final extension step at 72 °C for 20 min. Amplification products were then evaluated by running on 1% agarose gel. PCR products, ranging from 500 bp to 5 kb, were cloned into PCR2.1 vector using a TA Cloning Kit (Invitrogen, Carlsbad, CA). Overlapping regions of adjacent PCR products were set to 150–300 bp. At least six clones of each product were selected randomly and sequenced using ABI 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA). For sequencing fragments with size more than 1.4 kb, walking primers were designed based on the sequences obtained to determine the remaining region. Gap regions (caused by unsuccessful PCR

amplifications or incomplete primer-walking sequencing) were filled by sequencing PCR products amplified using primers designed from the regions flanking the gaps. After vector, primer, and low-quality sequences were removed, the remaining sequences were assembled using the software Bioedit (Hall 1999). Contigs were then manually assembled into the complete circular genome sequence. To check the assembly result, nine partially overlapped fragments covering the entire *C. oliveri* cp genome were amplified and sequenced. The fragments were 14–17 kb in length, with an overlapping region of 1 kb (supplementary fig. S1, Supplementary Material online). Each fragment was sequenced for three times by using different primers to generate amplicons. We have generated a total of 917,104 bp sequences, representing approximately 6.8-fold coverage of the *C. oliveri* cp genome.

Annotation

Initial annotation of the *C. oliveri* cp genome was performed using Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004). Genes and open reading frames (ORFs) that may not be annotated by DOGMA, such as *matK*, *ndhF*, *ndhG*, *rps16*, *ycf1*, and *ycf2*, were identified with the aid of Blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). From this initial annotation, putative starts, stops, and intron positions were determined based on comparisons to homologous genes in other cp genomes and by taking into account the possibility of RNA editing, which can modify the start and stop positions (Kugita et al. 2003). In addition, all tRNA genes were further verified by using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/). GC content was calculated with BioEdit (Hall 1999). The circular *C. oliveri* cp genome map was drawn using the program GenomeVx (Conant and Wolfe 2008) followed by manual modification.

RNA Extraction and Reverse Transcription-PCR

Total RNA was extracted following the protocol of RNAiso for Polysaccharide-rich Plant Tissue (Takara Bio Inc, Dalian, China). cDNA templates were synthesized using Reverse Transcriptase M-MLV Kit (Takara Bio Inc, Dalian, China).To examine whether the *accD* gene was transcribed, a reverse transcription (RT)-PCR was carried out using a pair of primers designed based on the full-length CDS of the gene in *C. oliveri*.

Comparative Analysis of cp Genome Rearrangements

As shown in figures 1 and 2, the complete cp genomes of *Cyc. taitungensis*, *C. oliveri*, and *Ced. deodara* were used to perform the analysis. We have specifically conducted two comparisons: *Cycas* versus *Cephalotaxus* and *Cycas* versus *Cedrus*. In figure 3, the IRs in *C. oliveri*, *C. wilsoniana*, *Taiwania cryptomerioides*, and *Cry. japonica* cpDNAs were identified through alignments of the complete genome sequences against themselves via Basic Local Alignment Search Tool 2 sequences at the National Center for Biotechnology Information (http://blast.ncbi.nlm.nih.gov/). Conserved gene blocks were identified by comparing the gene orders of the cp genomes examined. The genome comparison tool GRIMM (Tesler 2002) was used to estimate the minimum number of inversions required for the transformations of the gene orders.

Sequence Analyses and Computational Methods

All sequences of protein-coding genes, tRNA genes, rRNA genes, introns, and intergenic spacers (IGSs) were extracted from the *C. oliveri* and *C. wilsoniana* cp genomes based on their annotation. The Perl script MISA (MIcroSAtellite; http://pgrc.ipk-gatersleben.de/misa/) was applied to identify simple sequence repeats (SSRs) in the cp genome. The criteria for SSRs search were set as follows. The size of motifs was one to six nucleotides, and the minimum repeat unit was defined as 10 for mononucleotides, five for dinucleotides, and four for tri-, tetra-, penta-, and hexa-nucleotides. The Tandem repeats finder software (Benson 1999) was used to identify tandem repeats.

The *C. oliveri* and *C. wilsoniana* cp genome sequences were aligned by the MAFFT online software (Katoh et al. 2005). Setting *C. wilsoniana* cp genome as a reference, indels and substitutions were counted in the comparison of 545 non-overlapping bins each with a size of 250 bp. A total of 10,899 overlapped repeats (6,089 forward and 4,810 reverse repeats) were identified by using REPuter (Kurtz et al. 2001) with a minimum size of 14 bp and a maximum of one nucleotide mismatch between the two repeat copies.

Results and Discussion

General Features of the *C. oliveri* cp Genome and Comparison with That of *C. wilsoniana*

The complete cp genome of *C. oliveri* (KC136217) is 134,337 bp in length (supplementary fig. S2, Supplementary Material online), slightly shorter than that of *C. wilsoniana* (136,196 bp) (Wu, Wang, et al. 2011) but longer than those of *T. cryptomerioides* (132,588 bp) (Wu, Wang, et al. 2011) and *Cry. japonica* (131,810 bp) (Hirao et al. 2008), which all belong to the cupressophytes. Compared with other gymnosperms (noncupressophytes), *Cyc. taitungensis* has the largest cp genome size (163,403 bp) to date (Wu et al. 2007), and *G. biloba* (156,945 bp) ranks second (Lin et al. 2012). Excluding these two species, all other gymnosperms whose whole cp genome has been sequenced have smaller cp genomes than that of *C. oliveri* and *C. wilsoniana*. Similar to *C. wilsoniana* and some other seed plants, *C. oliveri* has no typical IR regions in its cp genome, resulting in the LSC and SSC regions unable to be defined.

The *C. oliveri* cp genome encodes a total of 113 genes (supplementary table S1, Supplementary Material online),

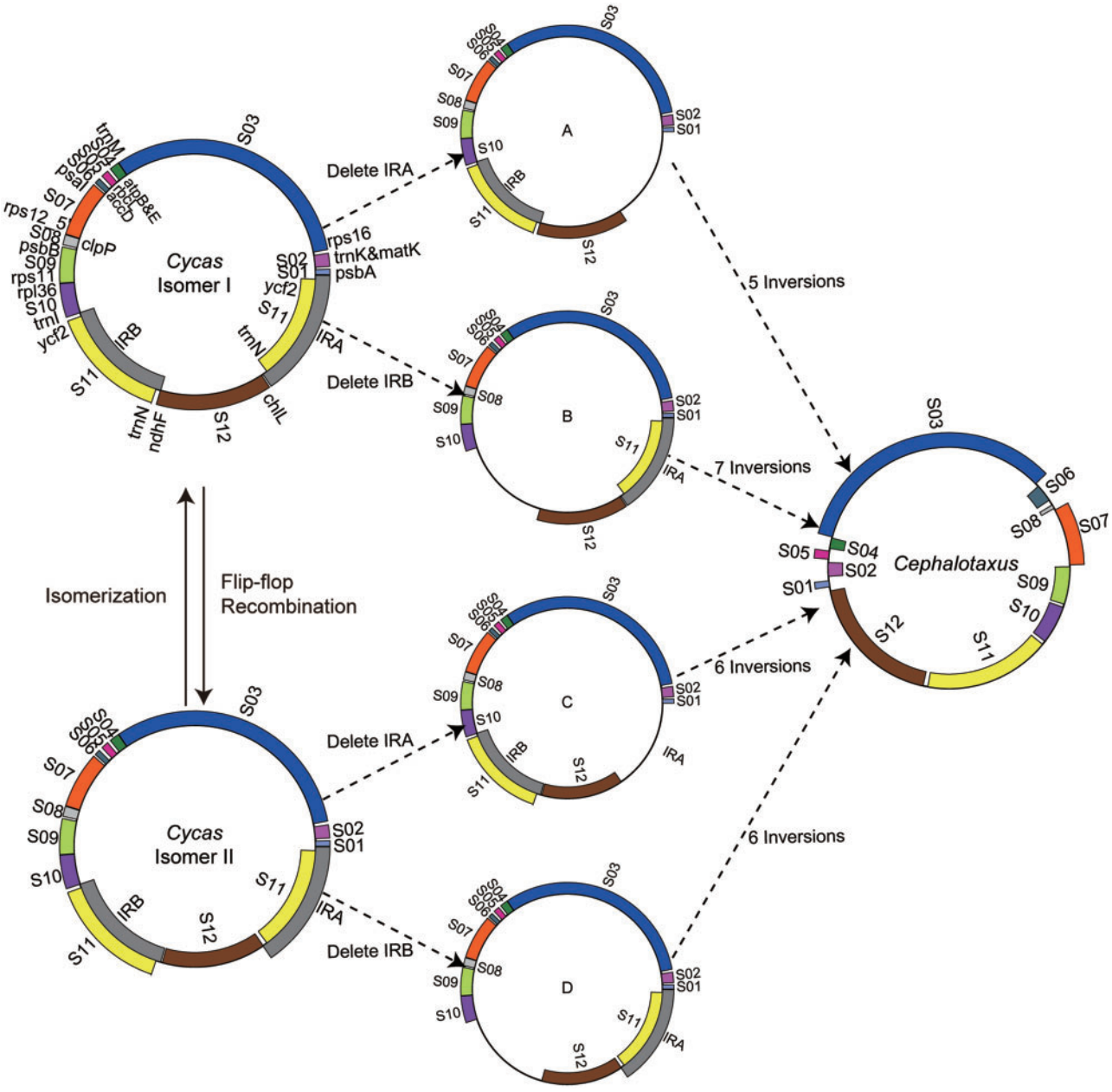

Fig. 1.—The transformation of gene order from the cpDNA of *Cycas* to that of *Cephalotaxus*. "Isomer I" represents the sequence deposited in GenBank (NC_009618), and "Isomer II" represents the reverse orientation. A total of 12 conserved blocks of genes were identified between the cpDNAs of *Cycas* and *Cephalotaxus* and denoted as S01–S12. Gene orientations have been distinguished by drawing them on either the inside or outside of the circles. The minimum number of inversions needed for the transformations from the four possible IR-lacking intermediates (A–D) to *Cephalotaxus* gene order was calculated using GRIMM (Tesler 2002).

including 4 ribosomal RNA genes, 27 transfer RNA genes, and 82 protein genes. It has equivalent numbers of protein-coding, tRNA, and rRNA genes with the *C. wilsoniana* cp genome (AP012265). The cp genome of *C. oliveri* is 1,859 bp shorter than that of *C. wilsoniana*, which can be chiefly ascribed to deletions in both coding and noncoding regions (table 1). Nonetheless, the two genomes share identical gene content and gene order. Of the 82 protein-coding genes, 14 show exactly the same sequence, which include 12 photosynthetic apparatus genes (*petL*, *petN*, *psaC*, *psaI*, *psaM*, *psbE*, *psbH*, *psbM*, *psbN*, *psbT*, *atpH*, and *ycf3*) and two ribosomal protein genes (*rpl33* and *rps12*). For the 27 tRNA genes, 24 are identical, with three (*trnS-GCU*, *trnK-UUU*, and *trnA-UGC*) each having only a single-nucleotide change. In the four rRNA

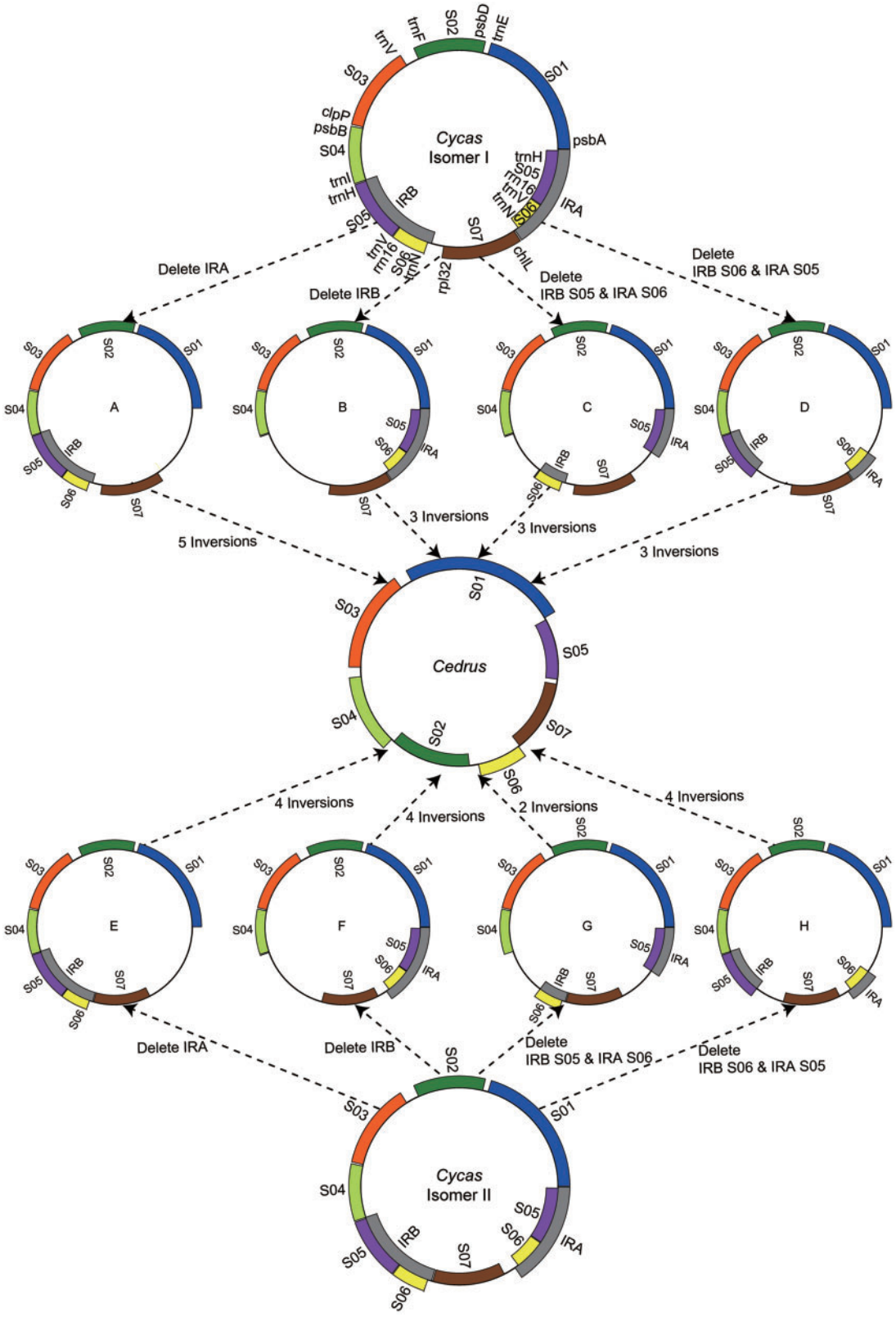

Fig. 2.—The transformation of gene order from the cpDNA of *Cycas* to that of *Cedrus*. "Isomer I" represents the sequence deposited in GenBank (NC_009618), and "Isomer II" represents the reverse orientation. A total of seven conserved blocks of genes were identified between the cpDNAs of *Cycas* and *Cedrus* and denoted as S01–S07. Gene orientations have been indicated on the inside or outside of the circles. The minimum number of inversions needed for the transformations from the eight possible IR-lacking intermediates (A–H) to *Cedrus* gene order was calculated using GRIMM (Tesler 2002).

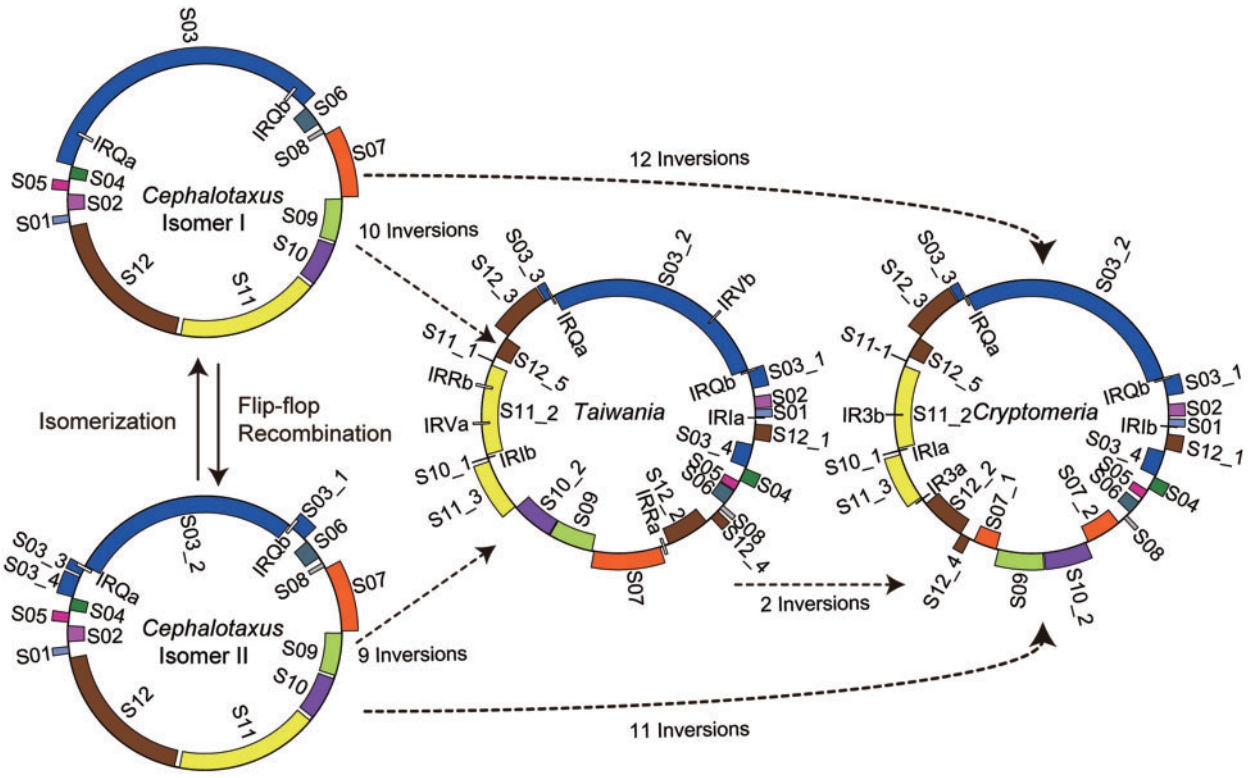

Fig. 3.—The occurrence of IRs and diversity of cp genome organization in cupressophytes. A number of gene blocks in the *Cephalotaxus* cp genome have been broken in *Taiwania cryptomerioides* and *Cryptomeria japonica*. IRQa and IRQb, repeats contain *trnQ-UUG*; IRVa and IRVb, repeats contain *trnV-GAC*; IRIa and IRIb, repeats contain *trnI-CAU*; IRRa and IRRb, repeats contain 5′ partial of *rrn23*; IR3a and IR3b, repeats reside in IGSs of *rrn16* to *trnV-GAC* and *trnL-CAA* to *ycf1*, respectively.

Table 1
Comparison of General Features for cp Genomes between *Cephalotaxus oliveri* and *C. wilsoniana*

	C. oliveri	*C. wilsoniana*
Length (bp)		
Total	134,337	136,196
Protein-coding genes	76,305	77,268
tRNA genes	2,110	2,110
rRNA genes	4,457	4,457
Introns	11,949	11,990
IGSs	39,516	40,371
GC content (%)		
Total	35.24	35.08
Protein-coding genes	36.23	36.04
tRNA genes	52.71	52.97
rRNA genes	52.02	52.14
Introns	35.29	35.08
IGSs	30.54	30.55

genes, *rrn5* has identical sequence, with *rrn4.5*, *rrn23*, and *rrn16* each showing a slight dissimilarity of 1%. The *C. oliveri* and *C. wilsoniana* cp genomes share 113 IGSs. Eleven of the IGS sequences are identical between them (*psbL/psbF*, *psbF/psbE*, *rpoC1/rpoB*, *psaB/psaA*, *ndhJ/ndhK*, *atpB/atpE*, *chlL/chlN*, *ndhH/ndhA*, *ccsA/trnL-UAG*, *rpl22/rps3*, and *rpl14/rps8*).

The GC contents of the protein coding regions in *C. oliveri* and *C. wilsoniana* cp genome are 36.23% and 36.04%, respectively. *Cephalotaxus wilsoniana* has higher GC content than *C. oliveri* in first codon position but lower in second and third positions. The lower GC content at the third codon position both in *C. oliveri* and *C. wilsoniana* cp genome reflects a strong codon usage bias for A and T, which may also contribute to the high AT content of the *Cephalotaxus* cp genomes. This pattern has been attributed to codon usage bias (Shimda and Sugiuro 1991; Chaw et al. 2004; Kim and Lee 2004; Liu and Xue 2005). Guisinger et al. (2008) has suggested that codon usage in cp genes is generally driven by selection and not GC content.

Mutational Dynamics in *Cephalotaxus* cpDNAs

Because information remains limited on the nature and organization of repeated elements in cp genomes, we have therefore analyzed their occurrence, nature, organization, and distribution in the *C. oliveri* cp genome. In total, 24 SSRs have been identified in the *C. oliveri* cp genome (supplementary table S2, Supplementary Material online). Of them,

Table 2
Statistical Significance of Colocation of Indels versus **Substitutions, Repeats** versus **Indels, and Repeats** versus **Substitutions in the Aligned** cp **Genomes of *Cephalotaxus oliveri* and *C. wilsoniana***

	Indels and Substitutions	Repeats and Indels	Repeats and Substitutions
Correlation value (r)	0.103	0.132	0.303
Coefficient of determination (r^2)	0.010	0.018	0.064
Significance of correlation (t)	56.064*	72.489**	141.442**
Significance (two tailed)	0.016	0.002	1.965E-09

*Correlation was significant at 0.05_{α} and 543 degree of freedom.
**Correlation was significant at 0.01_{α} and 543 degree of freedom.

homopolymers are the most common SSRs (15), whereas tripolymers (5) and dipolymers (4) rank second and third most frequent SSRs. This is in accordance with the previous findings that the most common type of cp repeat is the mononucleotide repeat, whereas di- or trinucleotide repeats are rare (Cozzolino et al. 2003). All the homopolymers and dipolymers are composed of multiple A or T bases, and four of all five tripolymers are also composed of A and T. In contrast, only one C/G-containing SSR has been found in the CDS of *infA*, whose type is ATG*5. Eleven of the 24 SSR loci are located within IGSs, 10 in coding regions, and 3 in introns. A total of 11 SSRs are shared between *C. oliveri* and *C. wilsoniana* cp genomes, of which seven are present in the gene coding regions (*ndhC*, *rps8*, *ycf1*, *rpoc2*, and *ycf2*).

Medium size tandem repeats, with a sequence identity of more than 90%, have been also examined (supplementary table S3, Supplementary Material online). With the help of the online software Tandem repeats finder (Benson 1999), we have identified 17 tandem repeats in the *C. oliveri* cp genome, of which 12 are located in coding regions of *ycf1* (6), *accD* (3), *rps18* (1), *ycf2* (1), and *rps3* (1). The other five tandem repeats are distributed in the IGS of *trnI*-GAU/*rrn16* (3), *atpE/rbcL* (1), and *psbE/petL* (1). In comparison, 7 out of the 17 tandem repeats are found to be shared in the *C. wilsoniana* cp genome; five are located in the coding region of *ycf1* (1), *accD* (3), and *rps3* (1), whereas two are present in the IGS of *trnI*-GAU/*rrn16* (1) and *atpE/rbcL* (1). Tandem repeats with repeat size over 10 bp have also been described from cp genomes of other gymnosperms, for example, *Pseudotsuga* (Hipkins et al. 1995). The medium size repeats may be used to develop a cp-genome-specific gene introduction vector by using them as specific recombination sites (Yi et al. 2012).

We have further explored the extent of genome-wide association between repeats, indels, and substitutions in *Cephalotaxus* cp genomes as Ahmed et al. (2012) have done in the monocot family Araceae. As presented in table 2, correlations are significant in the pairwise comparisons between the three types of mutations: substitutions and indels, repeats and indels, and repeats and substitutions. The highest correlation is found between repeats and substitutions, followed by correlations between repeats and indels and then substitutions and indels. These results lend further support for the hypothesis that repeat sequences play a role in the generation of substitutions and indels (McDonald et al. 2011; Ahmed et al. 2012).

IR Losses in *Cephalotaxus* and Other Gymnosperms

Extant gymnosperms consist of five major groups: cycads, *Ginkgo*, Pinaceae, gnetophytes, and cupressophytes. The cupressophytes are non-Pinaceae conifers, including six families sensu lato: Araucariaceae, Ceplalotaxaceae, Cupressaceae, Podocarpaceae, Sciadopityaceae, and Taxaceae. Recently, Wu, Wang, et al. (2011) have conducted a comparative genomic analysis of cpDNAs of *Cyc. taitungensis*, *E. equisetina*, *Ced. deodara*, and *C. wilsoniana*, concluding that the regions encompassing the *ycf2* and the adjoined *psbA* or *rpl*23 gene in the cpDNAs of *Ced. deodara* and *C. wilsoniana* are the retained IRs corresponding to either the IR_A or IR_B in those of *Cyc. taitungensis* and *E. equisetina*. They have further suggested that Pinaceae and cupressophytes retain different residual IR copies; namely the former saves IR_A, whereas the latter IR_B.

Here, we have more comprehensively compared the gymnosperm cp genomes with 15 representative species of each of the five groups (supplementary fig. S3, Supplementary Material online). Our results are in agreement with earlier studies suggesting that the cp genomes of cupressophytes and Pinaceae have lost one of the IRs that are found in those of other plants. The regions including *ycf2* and *psbA* are the retained IRs (Wu, Wang, et al. 2011). Nevertheless, we are cautious of the claim that the two groups have preserved different IR copies.

The following analyses that we have conducted have specifically considered the potential role of cpDNA isomers, which can be produced by IR-facilitated homologous recombination (HR) (Palmer 1983). We focus on the cpDNAs of *Cycas*, *Cephalotaxus*, and *Cedrus* as previous studies have shown that they each have the most ancestral gene order for seed plants (Jansen and Ruhlman 2012), cupressophytes (Wu, Wang, et al. 2011), and Pinaceae (Wu, Lin, et al. 2011), respectively. The cpDNA sequence of *Cyc. taitungensis* deposited in GenBank (NC_009618) is represented by the Cycas Isomer I, whereas the same sequence whose LSC or SSC has been manually reversed is represented by the Cycas Isomer II (figs. 1 and 2). Theoretically, after the loss of one copy of IRs, the two types of isomer may generate in total four distinct forms of IR-lacking intermediates (A–D) in the transformation of cp genome architecture from *Cycas* to *Cephalotaxus* (fig. 1). We have used GRIMM (Tesler 2002) to estimate the minimum number of inversions (reversal distance) required to achieve the transformation from each of the intermediates into the *Cephalotaxus* gene order.

As shown in figure 1, starting from *Cycas* Isomer I, the transformation via Intermediate A (the loss of IR_A) appears more parsimonious than via Intermediate B (the loss of IR_B). This is because the former requires a minimum of five inversions, whereas the latter requires seven. In contrast, when transforming from the *Cycas* Isomer II, at least six inversions are needed via either the loss of IR_A or IR_B.

As for the *Cycas–Cedrus* transformation, it should be more complex than the *Cycas–Cephalotaxus* one, because the *Cedrus* regions corresponding to the ancestral IR segments exist not as a single block but as two distantly separated blocks (conservative gene blocks S05 and S06 in fig. 2). The loss of one IR copy in the *Cedrus* cp genome can be due to 1) a single deletion of the entire IR_A or IR_B, 2) two separate deletions of S05 and S06 from the same IR copy, or 3) the deletion of S05 from IR_A but S06 from IR_B, and vice versa (fig. 2). To simplify the analysis, we have only shown the situation wherein the IR deletion(s) occur before inversions. A total of eight IR-lacking intermediates (A–H in fig. 2) may be generated after one IR copy or the S05/S06 blocks have been deleted from each of the *Cycas* isomers. If the *Cedrus* gene order is derived from *Cycas* Isomer I, the most parsimonious pathway is the loss of IR_B as a single block (Intermediate B) and then followed by merely three inversions. Nevertheless, if *Cycas* Isomer II pattern is the actual ancestor, the pathway that includes two separate deletions of S05 from IR_B and S06 from IR_A (Intermediate G) and followed by two inversions is the most parsimonious. Because both of the two most parsimonious pathways undergo the same number of genomic changes, one deletion plus three inversions for the former and two deletions plus two inversions for the latter, it remains open which pathway reflects the true process leading to the loss of IR in the Pinaceae cpDNA. Currently, we cannot rule out the possibility that the same IR copy has been lost from all conifers.

IRs and the Diversity of cpDNA Organizations in Cupressophytes

It is known that one copy of the canonical rRNA-containing IR has been lost in conifers. Wu, Lin, et al. (2011) proposed that this high IR reduction may cause IRs to lose the ability to induce HR. They have also shown that certain Pinaceae-specific short repeats can replace the reduced IRs to mediate HR in Pinaceae. Here, we explore whether a similar scenario exists in cupressophytes. The *C. oliveri* cp genome has a pair of 544 bp IRs between *trnT-UGU* and *trnL-UAA*, forming a duplication of full length of *trnQ-UUG* and 5′ partial of *chlB* (fig. 3). As well, this pair of IRs is present in the *C. wilsoniana* cp genome with a slightly shorter length of 530 bp. In the *T. cryptomerioides* and *Cry. japonica* cp genomes, the repeats are shortened to 277 bp and 284 bp, respectively. Following Wu, Lin, et al. (2011), we conducted PCRs to test whether the *trnQ*-related repeats, like those type 1 and 3 repeats in Pinaceae, may represent substrates for HR (supplementary fig. S4, Supplementary Material online). We have detected two isomeric cpDNA forms in *C. oliveri* but only one in both *T. cryptomerioides* and *Cry. japonica*.

Besides the *trnQ*-related IRs, *T. cryptomerioides* and *Cry. japonica* cp genomes also contain a pair of small *trnI-CAU*-containing repeats and other two species-specific repeats in *T. cryptomerioides*, one in *Cry. japonica*, respectively (fig. 3). However, all these repeats fail to induce HR (supplementary fig. S5, Supplementary Material online). Our results imply that the length of the repeats is crucial for them to be substrates for HR.

Diversification of the *accD* Gene in Gymnosperm cp Genomes

The *accD* gene encodes the β-carboxyl transferase subunit of plastidic acetyl-CoA carboxylase (ACCase, EC 6.4.1.2). ACCase catalyzes the formation of malonyl-CoA from acetyl-CoA and is considered to be the rate-limiting enzyme in the regulation of de novo fatty acid biosynthesis (Madoka et al. 2002). *AccD* is widely distributed in plants, including nonphotosynthetic parasitic plants. However, in gymnosperms, the gene is found to be totally lost in Gnetales, and its reading frame length in other gymnosperms varies significantly. The *accD* reading frame length of the *C. oliveri* cp genome is 936 codons, which is shorter than that of *C. wilsoniana* (1,056 codons) but longer than those of *Cry. japonica* (700 codons) and *T. cryptomerioides* (800 codons). In contrast, the reading frame lengths of cycads, *Ginkgo*, and Pinaceae range from 320 to 359 codons, less than half the size of that of cupressophytes. Furthermore, the *accD* reading frame length in cupressophytes is also larger than those whose entire cp genome sequence are available in liverworts, mosses, lycophytes, ferns, and angiosperms excluding Fabaceae. These results support the hypothesis of Hirao et al. (2008) that the *accD* reading frame has displayed a tendency toward enlarging sizes in cupressophytes.

We have aligned the deduced amino acid sequences of accD in seven gymnosperm cp genomes (fig. 4). Because the *accD* genes of examined Pinaceae lineages are conserved with 92.38% sequence identity, we chose that of *Ced. deodera* as the representative for the reasons mentioned earlier. The alignment shows that the amino acid sites near the N and C terminal regions are more conserved than the middle region; and it is the large insertion in the middle of the sequence that causes the length expansion of the *accD* CDS in cupressophytes (fig. 4 and supplementary data S1, Supplementary Material online). Compared with *Cyc. taitungensis*, *G. biloba*, and *Ced. deodera*, *C. oliveri* has a 620-codon insertion in the middle region of accD, in which nine repeats of SDIE/DE/SD/F motif have been observed. In contrast, *C. wilsoniana* has a 740-codon insertion, which contains 17 repeats of the same motif. In *T. cryptomerioides*, a 464-codon large

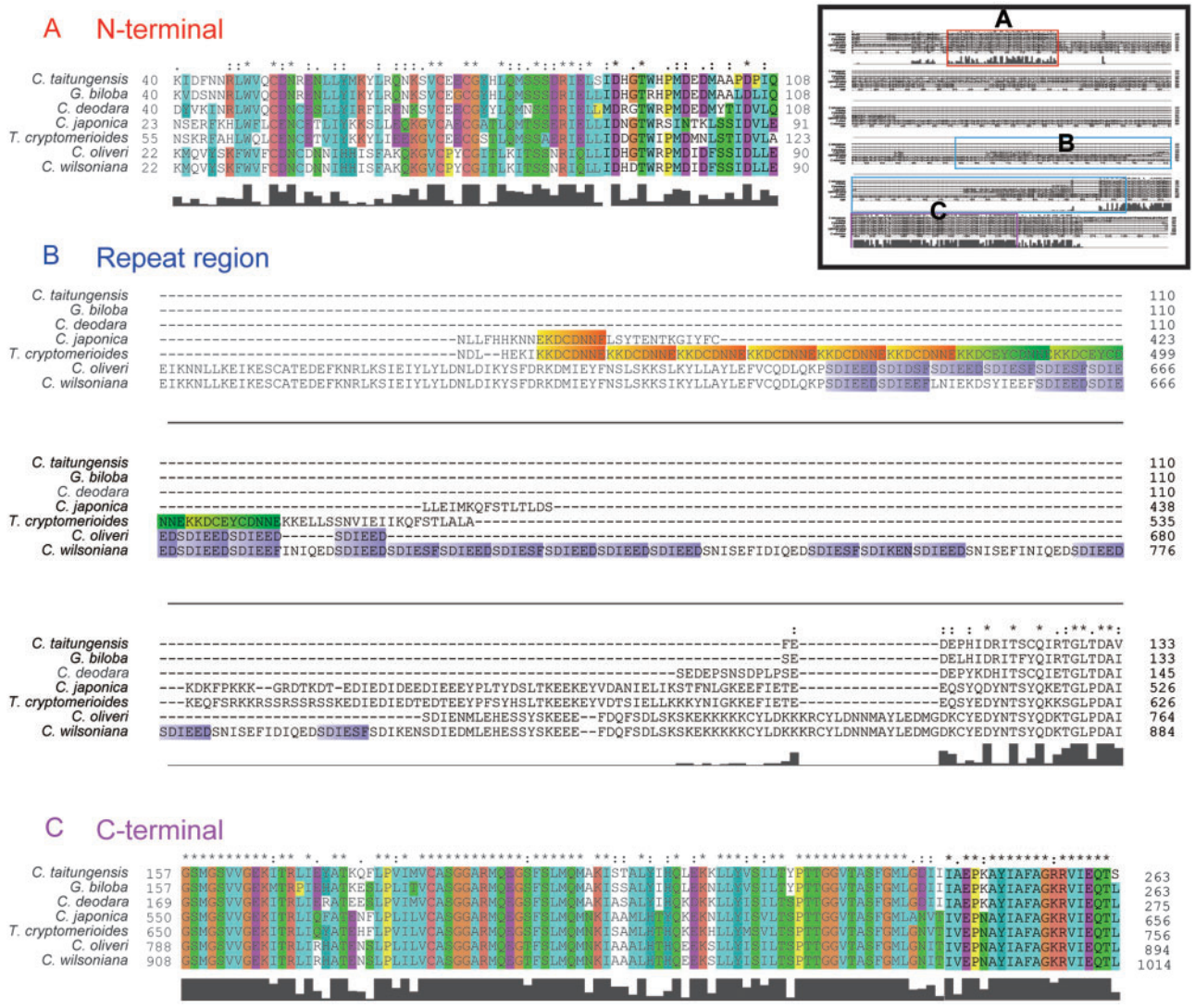

Fig. 4.—Comparison of amino acid sequences of accD from seven gymnosperms. (*A*) Section of the N terminal of accD with 69 conserved amino acids. (*B*) Repeat region; repeat units are marked with gradient-colored boxes. (*C*) The 107 conserved amino acids close to the C terminal. The histogram below the sequences represents the degree of similarity. Peaks indicate sites of high similarity and valleys of sites of low similarity. Numbers on either side of the sequences indicate the relative position in each accD. Above the amino acid, "*" indicates positions that have a single, fully conserved residue; ":" indicates that one of the following strong groups is fully conserved; and "." indicates that one of the following weaker groups is fully conserved. The thumbnail view of the alignment is shown in the top right corner.

insertion has occurred, containing six repeats of KKDCDNNE and three repeats of KKDCEYCDNNE. Of note, in *Cry. japonica,* only one copy of EKDCDNNF has been detected. In the previous study, Magee et al. (2010) have also partly attributed the expansion of the *accD* ORF to the presence of tandemly repeated sequences in the legume cpDNA. The repetitive insertions may be beneficial, possibly fostering new phenotypic variants (Erixon and Oxelman 2008). Our RT-PCR results show that the *accD* gene has been actively transcribed in the leaves of *C. oliveri* (supplementary fig. S6, Supplementary Material online), suggesting that it is functional.

Conclusion

We provide here the complete cp genome sequence of *C. oliveri*, a cupressophyte endemic to China. Availability of this sequence and the recently determined *C. wilsoniana* cp genome sequence enables us to assess genome-wide mutational dynamics within the genus *Cephalotaxus*. The cp genome of *C. oliveri* is 1,859 bp shorter than that of *C. wilsoniana*, which can be chiefly ascribed to deletions in genic and IGS regions. Both cp genomes, however, share identical gene content, gene order, and intron content; they all lack typical IR regions. The number and pattern of SSRs, tandem

repeats, and their shared loci for *C. oliveri* and *C. wilsoniana* cp genomes have been revealed. Of note, significant correlations have been observed between repeats, indels, and substitutions in the *Cephalotaxus* cp genomes, which provide further evidence for the hypothesis that repeats are crucial in inducing substitutions and indels. Pinaceae and cupressophytes are thought to retain different IR copies in their cp genomes (Wu, Wang, et al. 2011). Nevertheless, after examining the role that cpDNA isomers may play during the IR loss process, we cannot exclude that it is the same IR copy that has been conserved (figs.1 and 2). In the Pinaceae cp genomes, Wu, Lin, et al. (2011) have found that the highly reduced IRs are replaced by certain short repeats to mediate HR. Our analyses have gained the similar results in cupressophyte cp genomes. In addition, there is an increasing tendency toward large size for the *accD* reading frame in cupressophytes. These length expansions are caused by large insertions consisting of tandemly repeated motifs.

Supplementary Material

Supplementary data S1, tables S1–S3, and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Jin-Qing Wu and Shou-Jun Zhang for providing samples, Jia Li and Yuan Zhou for experimental assistance, and Gong Xiao for bioinformatics assistance. They also gratefully acknowledge the valuable comments by two anonymous reviewers. This work was supported by National Natural Science Foundation of China (nos. 30970290 and 31070594), Knowledge Innovation Program of the Chinese Academy of Sciences (nos. KSCX2-EW-J-20 and KSCX2-YW-Z-0940), Basic Research Project of Department of Science and Technology of Zhuhai City, China (no. 2012D0401990031), and the CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences.

Literature Cited

Ahmed I, et al. 2012. Mutational dynamics of aroid chloroplast genomes. Genome Biol Evol. 4:1316–1323.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Chaw SM, Chang CC, Chen HL, Li WH. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol. 58:424–441.

Conant GC, Wolfe KH. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24:861–862.

Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A. 2003. Molecular evolution of a plastid tandem repeat locus in an orchid lineage. J Mol Evol. 57:S41–S49.

Erixon P, Oxelman B. 2008. Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene. PLoS One 3:e1386.

Fu LK. 1984. A study on the genus *Cephalotaxus* Sieb. et Zucc. Acta Phytotax Sin. 22:277–288.

Fu LK, Li N, Mill RR. 1999. Cephalotaxaceae. In: Wu ZY, Raven PH, editors. Flora of China. St. Louis (MO): Missouri Botanical Garden. p. 85–88.

Gillham NW, Boynton JE, Harris EH. 1985. Evolution of plastid DNA. In: Cavalier-Smith T, editor. DNA and evolution: natural selection and genome size. New York: Wiley. p. 220–351.

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

Gu ZJ, Zhou QX, Yue ZS. 1998. A karyomorphological study of Cephalotaxaceae. Acta Phytotax Sin. 36:47–52.

Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci U S A. 105:18424–18429.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 41:95–98.

Hipkins VD, Marshall KA, Neale DB, Rottmann WH, Strauss SH. 1995. A mutation hotspot in the chloroplast genome of a conifer (Douglas-fir: *Pseudotsuga*) is caused by variability in the number of direct repeats derived from a partially duplicated tRNA gene. Curr Genet. 27:572–579.

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. 2008. Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 8:70.

Hu YS, Shao W. 1986. Comparative anatomy of the secondary phloem of the stem in *Cephalotaxus*. Acta Phytotax Sin. 24:423–427.

Jansen RK, Ruhlman TA. 2012. Plastid genomes in seed plants. In: Bock R, Knoop V, editors. Genomics of chloroplasts and mitochondria. Dordrecht (The Netherlands): Springer. p. 103–126.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Kim KJ, Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 11:247–261.

Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci U S A. 76:41–45.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K. 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res. 31:2417–2423.

Kurtz S, et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29:4633–4642.

Li Y, Wang FH, Chen ZR. 1986. An embryological investigation and systematic position of *Cephalotaxus oliveri* Mast. Acta Phytotax Sin. 24:411–422.

Lin CP, Wu CS, Huang YY, Chaw SM. 2012. Complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. Genome Biol Evol. 4:374–381.

Liu QP, Xue QZ. 2005. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. J Genet. 84:55–62.

Madoka Y, et al. 2002. Chloroplast transformation with modified *accD* operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. Plant Cell Physiol. 43:1518–1525.

Magee AM, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700–1710.

McDonald MJ, Wang WC, Huang HD, Leu JY. 2011. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biol. 9:e1000622.

Palmer JD. 1983. Chloroplast DNA exists in two orientations. Nature 301: 92–93.
Shimda H, Sugiura M. 1991. Fine-structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res. 19:983–995.
Su YJ, Wang T, Yang WD, Huang C, Fan GK. 1998. DNA extraction and RAPD analysis of *Podocarpus*. Acta Sci Nat Univ Sunyatseni. 37:13–18.
Tesler G. 2002. GRIMM: genome rearrangements web server. Bioinformatics 18:492–493.
Wang T, Su YJ, Zheng B, Li XY. 2002. Cladistic analysis of sequences of chloroplast *rbcL* gene and *trnL-trnF* intergenic spacer in Taxaceae and related taxa. Acta Sci Nat Univ Sunyatseni. 41:70–74.
Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 6:273–297.
Wu CS, Lin CP, Hsu CY, Wang RJ, Chaw SM. 2011. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. Genome Biol Evol. 3:309–319.
Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. 3:1284–1295.
Wu CS, Wang YN, Liu SM, Chaw SM. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol. 24:1366–1379.
Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.
Xi YZ. 1993. Studies on pollen morphology and exine ultrastructure in cephalotaxaceae. Acta Phytotax Sin. 31:425–431.
Yi DK, Lee HL, Sun BY, Chung MY, Kim KJ. 2012. The complete chloroplast DNA sequence of *Eleutherococcus senticosus* (Araliaceae); comparative evolutionary analyses with other three asterids. Mol Cells. 33: 497–508.

Associate editor: Bill Martin