Mol. Cells *33*, 497-508, May 31, 2012
DOI/10.1007/s10059-012-2281-6

Molecules and Cells

The Complete Chloroplast DNA Sequence of *Eleutherococcus senticosus* (Araliaceae); Comparative Evolutionary Analyses with Other Three Asterids

Dong-Keun Yi, Hae-Lim Lee, Byung-Yun Sun[1], Mi Yoon Chung[2], and Ki-Joong Kim*

This study reports the complete chloroplast (cp) DNA sequence of *Eleutherococcus senticosus* (GenBank: JN 637765), an endangered endemic species. The genome is 156,768 bp in length, and contains a pair of inverted repeat (IR) regions of 25,930 bp each, a large single copy (LSC) region of 86,755 bp and a small single copy (SSC) region of 18,153 bp. The structural organization, gene and intron contents, gene order, AT content, codon usage, and transcription units of the *E. senticosus* chloroplast genome are similar to that of typical land plant cp DNA. We aligned and analyzed the sequences of 86 coding genes, 19 introns and 113 intergenic spacers (IGS) in three different taxonomic hierarchies; *Eleutherococcus* vs. *Panax*, *Eleutherococcus* vs. *Daucus*, and *Eleutherococcus* vs. *Nicotiana.* The distribution of indels, the number of polymorphic sites and nucleotide diversity indicate that positional constraint is more important than functional constraint for the evolution of cp genome sequences in Asterids. For example, the intron sequences in the LSC region exhibited base substitution rates 5-11-times higher than that of the IR regions, while the intron sequences in the SSC region evolved 7-14-times faster than those in the IR region. Furthermore, the Ka/Ks ratio of the gene coding sequences supports a stronger evolutionary constraint in the IR region than in the LSC or SSC regions. Therefore, our data suggest that selective sweeps by base collection mechanisms more frequently eliminate polymorphisms in the IR region than in other regions. Chloroplast genome regions that have high levels of base substitutions also show higher incidences of indels. Thirty-five simple sequence repeat (SSR) loci were identified in the *Eleutherococcus* chloroplast genome. Of these, 27 are homopolymers, while six are di-polymers and two are tri-polymers. In addition to the SSR loci, we also identified 18 medium size repeat units ranging from 22 to 79 bp, 11 of which are distributed in the IGS or intron regions. These medium size repeats may contribute to developing a cp genome-specific gene introduction vector because the region may use for specific recombination sites.

INTRODUCTION

The plant chloroplast (cp) genome has maintained a relatively conserved structure and gene content throughout evolution. Therefore, the genome is widely used to trace the evolutionary history of the plant kingdom. Complete cp DNA sequences have been generated from 104 species of plants, including liverwort (*Marchantia polymorpha*; Ohyama et al., 1986), black pine (*Pinus thunbergii*; Wakasugi et al., 1994), rice (*Oryza sativa*; Hiratsuka et al., 1989), ginseng (*Panax schinseng*; Kim and Lee, 2004), carrot (*Daucus carota*; Ruhman et al., 2006), tobacco (*Nicotiana tabacum*; Shinozaki et al., 1986) and many others (Bausher et al., 2006; Chung et al., 2006; Daniell et al., 2006; Jansen et al., 2006; Jo et al., 2011; Kim et al., 2006; 2009; Kuang et al., 2011; Samson et al., 2007; Yang et al., 2010). However, the information currently available on cp genomes is insufficient for elucidating the general evolutionary mechanisms of the genomes or for defining the evolutionary patterns of the plant kingdom.

Land plant cp genomes usually contain two inverted repeat (IR) regions between a small single copy (SSC) region and a large single copy (LSC) region (Palmer, 1990; 1991). Approximately 110-120 genes are located along the circular genome structure. Although the number of genes is conserved across the plant kingdom, except in some parasitic plants (Funk et al., 2007; McNeal et al., 2007; Wolfe et al., 1992), a wide range of genomic size variations exist (from 72 to 217 kb) due to contraction and expansion of the IR regions. Length variation is the most common mutation encountered in plant cp genomes (Palmer, 1987). In addition, small variations in length are prevalent in cp genomes, and may be the result of the slippage-mispairing mechanisms active during cp genome replication processes (Bowman and Dyer, 1986; 1988; Cosner et al., 1997; Morton and Clegg, 1993; Wolfson et al., 1991). Gene order in cp genomes is frequently altered and provides important infor-

School of Life Sciences, Korea University, Seoul 136-701, Korea, [1]Division of Biological Sciences, Jeonbuk National University, Jeonju 561-756, Korea, [2]Department of Biology, Gyeongsang National University, Jinju 660-701, Korea
*Correspondence: kimkj@korea.ac.kr

Received December 14, 2011; revised March 11, 2012; accepted March 14, 2012; published online April 24, 2012

Keywords: chloroplast genome, *Eleutherococcus senticosus*, indels, nucleotide diversity, positional effect

mation for understanding the evolution of some plant groups. The evolutionary correlations between cp genome rearrangement and plant group diversification are exemplified in a number of flowering plant groups, including Fabaceae, Poaceae, Asteraceae, Gesneriaceae, and Oleaceae (Cosner et al., 1997; 2004; Doyle et al., 1992; Hachitel et al., 1991; Hiratsusuka et al., 1989; Hoot and Palmer, 1994; Jansen and Palmer, 1987; Kim et al., 2005; Lee et al., 2007; Saski et al., 2005).

Chloroplast genomes also show uneven distribution of mutation events along the genome. Therefore, evolutionary hot spots have been identified in a number of different plant groups (Guo and Terachi, 2005; Hipkens et al., 1995; Morton and Clegg, 1993; Raubeson and Jansen, 2005). These hot spot regions are characterized by high incidences of indels and rearrangements. However, elucidating the evolutionary mechanisms in hot spot regions is difficult due to several complicated factors and insufficient data.

Eleutherococcus senticosus (Rupr. & Maxim.) Maxim. (*Acanthopanax senticosus*) is commonly known as Siberian ginseng and belongs to the family Araliaceae, along with Korean ginseng (*Panax schinseng*). The plant grows up to 5 m tall and has small spines with palmate compound leaves. *E. senticosus* is listed as an endangered plant and is protected by wildlife protection acts in Korea. Distribution ranges extend to the far eastern regions of the Russian taiga and the northern regions of Korea, Japan and China. *E. senticosus* is cultivated in Korea, China and Japan as an important herbal medicine. The primary active ingredients of *E. senticosus* are typically concentrated in the root and consist mainly of chemically distinct glycosides called eleutherosides A-M. Eleutherosides I, K, L, and M have also been identified and isolated from the leaf of the plant. These active ingredients are reported to increase stamina, immune deficiency, resistance to a variety of physical, chemical, and biological stressors and to act as a general stabilizer/ normalizer (Kumura and Sumiyoshi, 2004; Kurkin, 2003). In addition to its anti-fatigue and anti-stress effects, the plant also exhibits immunomodulatory activity (Davydov and Krikorian, 2000).

The family Araliaceae is a member of the Asterid clade in flowering plants. The Asterid clade is the largest subgroup of the flowering plants, made up of more than 110 families and 100,000 species (APG III, 2009). Many familiar flowers, shrubs, and trees belong to this group, including a number of important food crops. However, only 18 complete Asterid cp DNA sequences are available. Furthermore, most of these sequences are from food crop species (e.g., tobacco, tomato, potato, lettuce, sunflower, etc.) which belong to two families, Solanaceae and Asteraceae. To understand the evolutionary mechanisms of the cp genome in the Asterid lineage, information on cp genomes are required from more of the diverse Asterid families.

As a part of this project, we first generated complete cp genome sequences from *E. senticosus* (Araliaceae) and analyzed several evolutionary parameters of the genome. Second, we also comparatively analyzed the evolutionary parameters of two closely-related cp genomes, *E. senticosus* and *P. schinseng*. Finally, to elucidate the general evolutionary mechanisms active in the cp genomes, we expanded the comparative evolutionary analyses to *Daucus carota* (Apiaceae) and *Nicotiana tabacum* (Solanaceae).

MATERIALS AND METHODS

Plants materials and DNA extraction

Approximately 2 grams of flesh leaves from *Eleutherococcus senticosus* were collected from a single individual in a natural forest habitat at the border between China and North Korea. The voucher specimen was deposited in the Korea University herbarium (KUS). Total DNA was extracted from the liquid nitrogen ground leaf powders using CTAB extraction methods. The DNA was purified using ultra-centrifugation in a cesium chloride/ethidium bromide gradient and then further purified by dialysis (Palmer, 1986).

Table 1. Genes contained in the *Eleutherococcus senticosus* cp genome (total 114 genes)

Category for genes	Group of genes	Name of genes
Self replication	rRNA genes	*rrn16*(x2), *rrn23*(x2), *rrn4.5*(x2), *rrn5*(x2)
	tRNA genes	*30* trn genes (6 contain an intron, 7 in the IR regions)
	Small subunit of ribosome	*rps2, rps3, rps4, rps7*(x2), *rps8, rps11, rps12*(*), *rps14, rps15, rps16*, rps18, rps19*
	Large subunit of ribosome	*rpl2**(x2), *rpl14, rpl16**, *rpl20, rpl22, rpl23*(x2), *rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2*
Genes for photosynthesis	Subunits of NADH-dehydrogenase	*ndhA*, ndhB**(x2), *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ, ycf3***
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Subunits of cytochrome b/f complex	*petA, petB*, petD*, petG, petL, petN*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF*, atpH, atpI*
	Large subunit of rubisco	*rbcL*
Other genes	Translational initiation factor	*infA*
	Maturase	*matK*
	Protease	*clpP***
	Envelope membrane protein	*cemA*
	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA*
Genes of unknown functions	Open reading frames (ORF, ycf)	*yef1, ycf2*(x2), *ycf4, ycf5, ycf15*(x2)

One and two asterisks after gene names reflect one- and two-intron containing genes, respectively. Genes located in the IR regions are indicated by the (x2) symbol after the gene name.

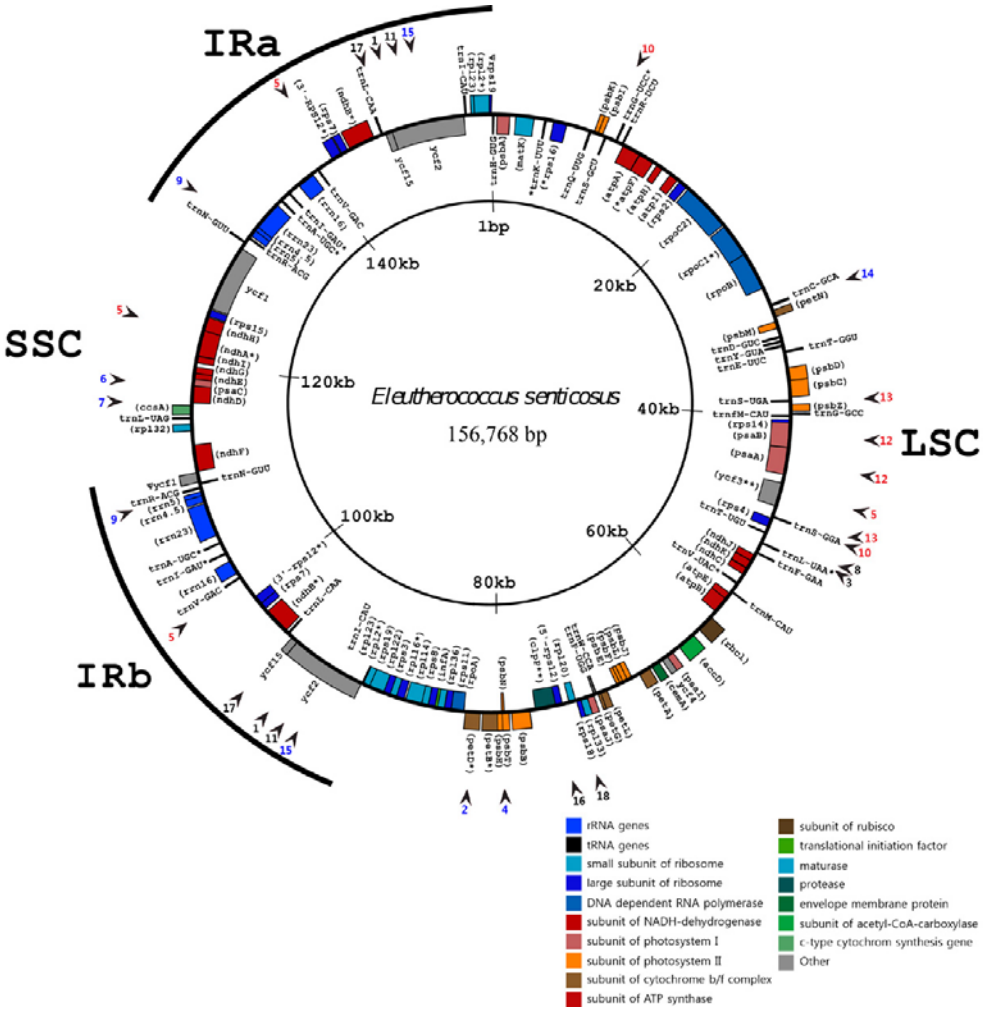

Fig. 1. The gene map of *Eleutherococcus senticosus* cp genome. A pair of thick lines at the outmost circle represents the inverted repeats (IRa and IRb; 25,930 bp each), which separate the large single copy region (LSC; 86,755 bp) from the small single copy region (SSC; 18,153 bp). Genes drawn inside the circle are transcribed clockwise, while those drawn outside the circle are transcribed counterclockwise. Intron-containing genes are marked by asterisks. The numbers at the outmost circle indicate the locations of 18 repeats including direct (black number), palimdromic (blue number), and dispersed repeats (red numbers), respectively (cf. Table 5).

PCR amplification and sequencing

Purified cpDNAs were amplified using a series of primer sets developed from the complete sequences of the *Panax schinseng* chloroplast genome (Kim and Lee, 2004). To obtain accurate sequences, each region of the chloroplast genome was amplified 3-15 times. The PCR products were purified with the MEGAquick-spin kit (iNtRON, Korea) and the cleaned products were sequenced in both directions using an ABI 3730X1 automatic sequencer.

Chloroplast gene annotation and sequence analyses

Sequence fragments were assembled using Sequencher 4.7 (Gene Code Corporation, USA). Gene annotations and comparative analysis were performed using the BLAST (BLASTN, PHI-BLAST, BLASTX), ORF finder program from the National Center for Biotechnology Information (NCBI) and DOGMA (Wyman et al., 2004). Codon usage and A-T contents were analyzed using MEGA4 (version 4.1). Repeating sequences were searched using REPuter (Kurtz et al., 2001) and further analyzed by Tandem Repeats Finder, ver. 4.0 (Benson, 1999). The locations and secondary structure of tRNA, rRNA, intron regions and other parts of genomes were evaluated using tRNAscan-SE (version 1.21; Lowe and Eddy, 1997) and mFOLD (version 3.3; Zuker, 2003). For sequence comparison, the genes, introns and gene spacer regions of cp genomes of different species were aligned using ClustalX (Thompson et al., 1994) and adjusted by hand. The spacer regions of chloroplast genomes from related species were aligned using the MUSCLE program (Edgar, 2004). mVISTA program was used to compare the overall similarities of *E. senticosus* cp genome to other cp genomes (Mayor et al., 2000). Nucleotide diversity and Ka/Ks values were analyzed using DnaSP program with K2P substitution model (version 4.50; Librado and Rozas, 2009).

RESULTS

The general features of *Eleutherococcus senticosus* cp genome

The complete chloroplast sequence of *E. senticosus* is 156,768 bp in length. It harbors a pair of inverted repeat regions (IRa and IRb) consisting of 25,930 bp each. The two IR regions divide the genome into an LSC region of 86,755 bp and an SSC region of 18,153 bp (Fig. 1). The 114 individual genes in the *E. senticosus* cp genome, including 80 peptides, 30 tRNAs, and four rRNA coding genes, are presented in Fig. 1 and summarized in Table 1. Ten protein coding and seven tRNA genes are duplicated and located in the IR regions. The LSC region contains 62 protein coding genes and 22 tRNA genes, while there are 12 protein coding genes and one tRNA gene located in the SSC region. The *E. senticosus* chloroplast genome con-

Table 2. The lengths of intron and exon splitting genes in the *Eleutherococcus senticosus* cp genome

Gene	Exon I	Intron I	Exon II	Intron II	Exon III
trn K	37	2515	35		
rps 16	40	891	197		
trn G	23	698	48		
atp F	145	724	410		
rpo C1	453	762	1617		
ycf 3	124	716	230	748	153
trn L	49	507	35		
trn V	39	590	35		
rps 12	114*	-	232	536	26
clp P	71	768	292	648	228
pet B	6	791	642		
pet D	8	754	475		
rpl 16	9	953	399		
rpl 2	391	660	434		
ndh B	777	679	756		
trn I	37	945	35		
trn A	38	810	35		
ndh A	553	1071	539		

The *rps12* gene is divided: the 5′-*rps12* is located in the LSC region and the 3′-*rps12* in the IR region.

Table 3. Base compositions in the *Eleutherococcus senticosus* cp genome

		T(U)	C	A	G	Sequence lengths (bp)
LSC region		32.5%	18.5%	31.4%	17.6%	86,755
IRa region		28.7%	22.3%	28.2%	20.7%	25,930
IRb region		28.2%	20.7%	28.7%	22.3%	25,930
SSC region		33.8%	16.8%	34.2%	15.2%	18,153
Total		31.3%	19.3%	30.7%	18.6%	156,768
		T(U)	C	A	G	Sequence lengths (bp)
Protein coding genes (CDS)		31.4%	17.6%	30.5%	20.5%	69,504
	1st position	24.0%	18.5%	30.4%	27.2%	23,168
	2nd position	32.0%	20.3%	29.2%	18.1%	23,168
	3rd position	38.0%	13.9%	32.0%	16.2%	23,168

tains 55% coding DNA and 45% non-coding DNA. Eighteen of the genes in the genome include one or two introns (Table 2). Among them, *rps12*, *clpP* and *ycf3* have two introns. The *rps12* gene is a unique divided gene in which the 5′ end exon is located in the LSC region and the 3′ exon and intron are duplicated and located in the IR regions.

The overall GC and AT contents of the *E. senticosus* chloroplast genome are 38% and 62%, respectively. The AT contents in the IR regions total 57% of the genome, whereas the AT contents in the LSC and SSC regions are 64% and 68%, respectively. The low AT content of the IR regions is attributed to the low number of A and T bases in the four rRNA (*rrn16*, *rrn23*, *rrn4.5*, *rrn5*) genes in these regions. The AT content of the protein coding regions is 62%; 54% in the first codon position, 61% in the second position, and 70% in the third position. The high AT content at the third codon position (Table 3) reflects a codon usage bias for A or T. The codon usage frequencies of stop codons are also biased to A or T at both the second and third codon positions (Supplementary data 1). Thirty tRNA genes representing 20 amino acids were identified in the *E. senticosus* cp genome by similarity search and computer prediction. Codon usage in the *E. senticosus* cp genome is summarized in Supplementary data 1. Of the 30 tRNA genes, *trnK-UUU*, *trnG-UCC*, *trnL-UAA*, *trnV-UAC*, *trnI-GAU* and *trnA-UGC* contained intervening sequences in the anticodon stem/loop or D-stem regions.

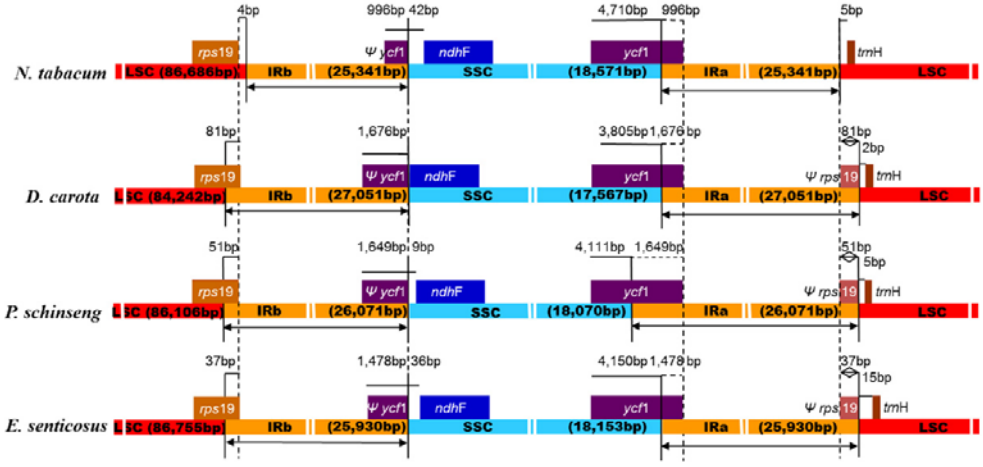

Fig. 2. Comparison of the LSC, IR and SSC border regions among four cp genomes.

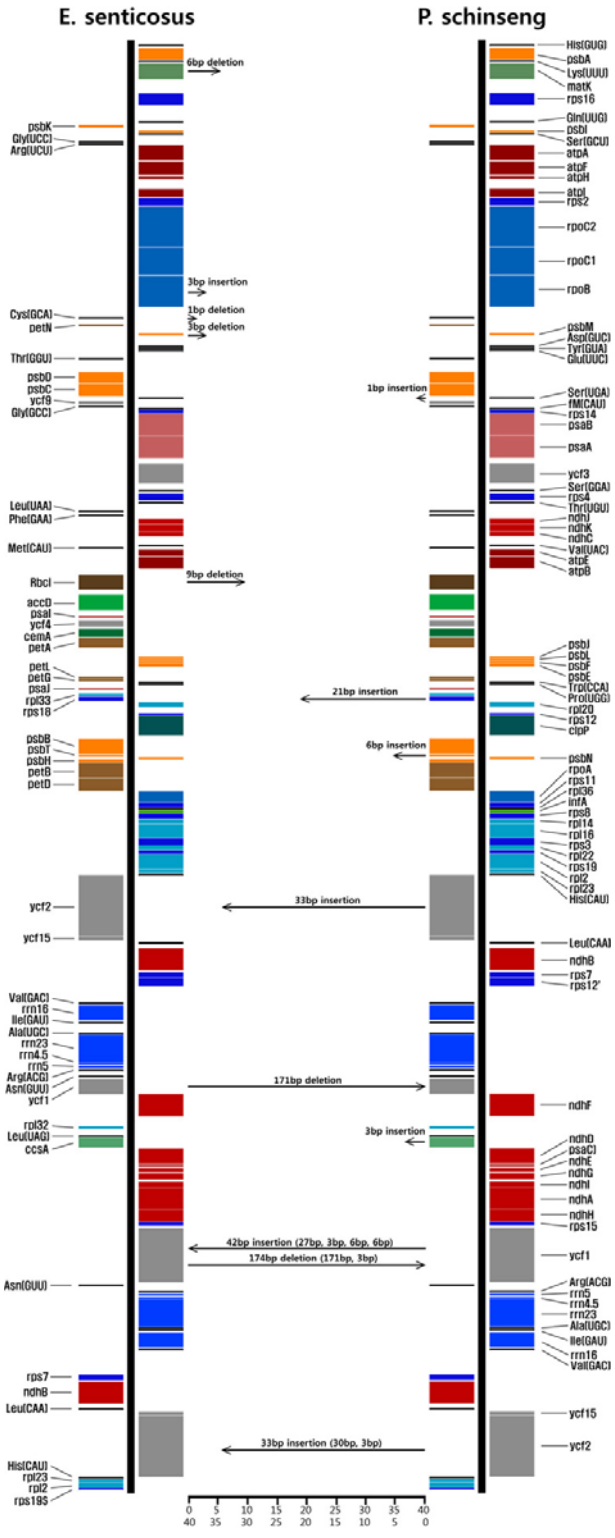

Fig. 3. The events and the lengths of indel mutations in the CDS regions of cp genomes between *Eletherococcus* and *Panax*.

Detailed comparison of the IR/single copy (SC) boundaries between four representative Asterids (*Eleutherococcus*, *Panax, Daucus,* and *Nicotiana*) are presented in Fig. 2. Pseudogenes

Table 4. Distribution of simple sequence repeat (SSR) loci in the *Eleutherococcus senticosus* cp genome

Unit	Lengths	No. SSRs	Coodinated basepairs*
A	10	4	4454-4463, 34056-34065, 46970-46979, 73278-73287
	11	4	12884-12894, **17663-17673**, 48793-48803, 57572-57582
	12	3	5616-5627, 6552-6563, 9660-9671
	13	2	**14218-14230, 23935-23947**
C	10	1	69754-69763
	11	1	23924-23934
	12	2	5457-5468, **137427-137438**
G	12	1	**106086-106097**
T	10	4	**27585-27594(*rpoB*), 56961-56970 (*atpB*), 80707-80716(*rpoA*), 82641-82650**
	11	3	**19876-19886(*rpoC2*)**, 73115-73125, **129107-129117(*ycf1*)**
	12	1	13596-13607
	13	1	**83750-83762**
AT	12	1	22269-22280(*rpoC1*)
	14	1	29881-29894
TA	10	2	29917-29926, **33619-33628**
	12	1	70390-70401
	16	1	**86479-86494**
ATA	12	1	**57015-57026**
TTC	12	1	**70069-70080**

The coordinated basepairs are the nucleotide number positions starting at the IRa/LSC junction (Fig. 1). The underline represents the SSR in the CDS and the bold numbers represent the shared SSR with *Panax*.

of *rps19* and *ycf1* of various lengths are located at the IR/LSC and IR/SSC boundaries, respectively. In the *Eleutherococcus* cp genome, the IR extends into the *rps19* gene and inserts a short *rps19* pseudogene (38 bp) at the IRa/LSC border, while the IR extends into the *ycf1* gene and inserts the *ycf1* pseudogene (1,478 bp) at the IRs/LSC border.

The repeat units and distributions in *E. senticosus* cp genome and small inversions

Simple sequence repeats (SSR) in which the same nucleotide sequence unit is repeated more than 10 times were identified in 35 different locations in the *Eleutherococcus* cp genome (Table 4). The majority (27) of the 35 SSR loci are homopolymers, while six are dipolymers and two are tri-polymers. Of the 27 homopolymer loci, 23 were composed of multiple A or T bases, while four were composed of multiples of C or G. All di-polymer loci were composed of AT or TA multiples. These SSR loci alsocontribute to the A-T richness of the *Eleutherococcus* cp genome. Twenty-nine SSR loci occur in the intergenic spacers, while only six SSR loci are located on the gene coding regions of *atpB, rpoA, rpoB, rpoC1, rpoC2* and *ycf1*. A total of 16 SSR loci are shared with the *Panax* cp genome, including five SSR loci in the gene coding region.

Repeats of 30 bp or longer with a sequence identity of more than 90% were also examined. REPuter initially identified 24 direct and 25 reverse repeats ranging from 30 to 79 bp in length. Many of these repeats, however, occurs tandem repeated

Table 5. Distribution of large repeat loci in the *Eleutherococcus senticosus* cp genome

Repeat number	Size (bp)	Repeat	Location	Region
1	79	direct	CDS (*ycf2*)	IRa, b
2	52	palindromic	Intron (*petD*)	LSC
3	45	direct	IGS (*trnT-UGU/trnL-UAA*)	LSC
4	44	palindromic	IGS (*psbT/psbN*)	LSC
5	42	dispersed	IGS (*rps12/trnV-GAC*), Intron (*ndhA*), IntronII (*ycf3*)	IR, SSC, LSC
6	41	palindromic	IGS (*ndhD/psaC*)	SSC
7	40	palindromic	IGS (*ccsA/ndhD*)	SSC
8	33	direct	IGS (*trnT-UGU/trnL-UAA*)	LSC
9	31	palindromic	IGS (*rrn5/trnR-ACG*)	IRa, b
10	30	dispersed	CDS (*trnS-GCU*), CDS (*trnS-GGA*)	LSC
11	30	direct	CDS (*ycf2*)	IRa, b
12	30	dispersed	CDS (*psaA*), CDS (*psaB*)	LSC
13	28	dispersed	CDS (*trnS-UGA*), CDS (*trnS-GGA*)	LSC
14	26	palindromic	IGS (*trnC-GCA/petN*)	LSC
15	26	palindromic	CDS (*ycf2*)	IRa, b
16	24	direct	IGS (5′-*rps12/clpP*)	LSC
17	24	direct	IGS (*ycf15/trnL-CAA*)	IRa, b
18	22	direct	CDS (*rps18*)	LSC

The repeat units larger than 22 bp are presented in this table and the locations are presented on the Fig. 1. The underline represents the SSR in the CDS and the bold numbers represent the shared SSR with *Panax*.

patterns in the same location of DNA sequences. Therefore, the repeating units, the number of repeats, the location of the repeats, and the total length were evaluated using the Tandem Repeat Finder. A total of 23 repeats were located, including eight direct tandem repeats, nine direct IRs, and six dispersed repeats (Fig. 1, Table 5). The repeats ranged from 18 to 42 bp in length, and were repeated from two to five times. Some dispersed repeats occurred in different regions of the cp genome, including the IR and the SSC regions or the IR and the LSC regions.

The gene order of the cp genome in *E. senticosus* is collinear to the *Panax* cp genomes (Fig. 3). Detailed comparisons of the two cp genomes, however, revealed several small inversion mutations, all of which were related to direct inverted repeats and stem-loop DNA sequence structures (Fig. 4).

Base substitution patterns of coding sequences (CDS) and intron regions in the *E. senticosus* cp genome compared to that of other Asterids

The overall indel and base pair similarities of *Eleutherococcus* were compared to 10 different published chloroplast genomes using the mVISTA program (Supplementary data 2). The *Eleutherococcus* cp genome has identical gene content and order as the cp genomes of *Panax, Daucus* and members of the Solanaceae family (Fig. 3). Therefore, the patterns of base substitutions at three different hierarchical levels were compared in detail using the chloroplast genomes of *Panax, Daucus*, and *Nicotiana*.

Several protein coding sequences were identical between *Eleutherococcus* and *Panax* (Supplementary data 3). Approximately 83% (8/11) of the genes were identical in the IR region, while only 18% of the genes (11/61) were identical in the LSC region, and no genes were identical in the SSC region between the two genera. The ycf1 gene showed the highest divergence (3.19%), followed by *rpl14* (2.71%), *rps19* (2.51%), *rpl22* (2.07%), *clpP* (2.03%), *rpl22* (2.07%), *rps15* (1.83%), and *matK* (1.59%). None of these genes are located in the IR region. The average sequence divergences of the gene coding sequences in the IR, LSC, and SSC regions were 0.26, 0.86, and 1.97%, respectively. The Ka/Ks ratios were larger than 1.00 in the *clpP* (1.61), *ycf2* (5.20) and *ycf1* (1.05) genes.

Only three gene sequences (*psbF, rps7* and *rrn5*) were identical between the *Eleutherococcus* and *Daucus* cp genomes (Supplementary data 4). Five out of 63 genes (*psbF* - 0.00%, 5′*rps12* - 0.38%, *atpH* - 0.41%, *psaJ* - 0.79%, *psbN* - 0.98%) had less than 1% sequence divergence from the LSC region, and seven out of 12 genes had less than 1% sequence diver-

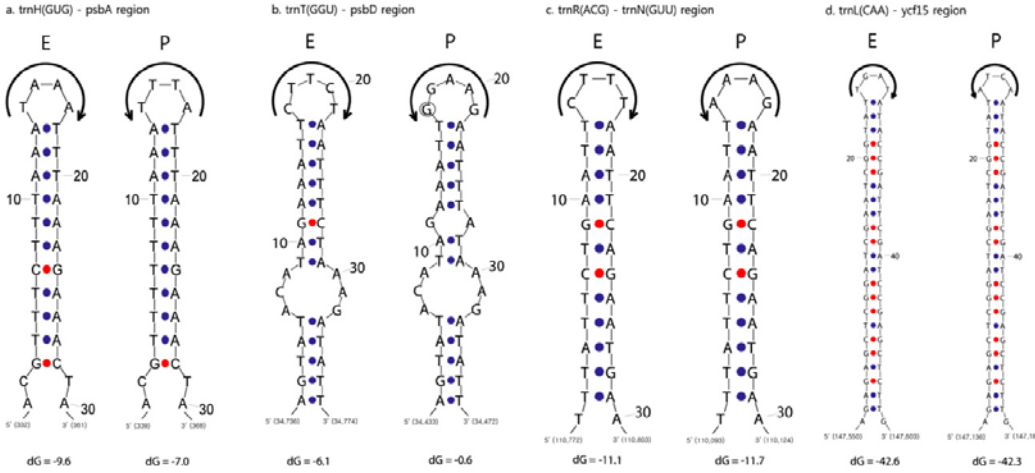

Fig. 4. Small inversion mutations and associated secondary structures between the cp genomes of *Eletherococcus* (E) and the cp genome of *Panax* (P).

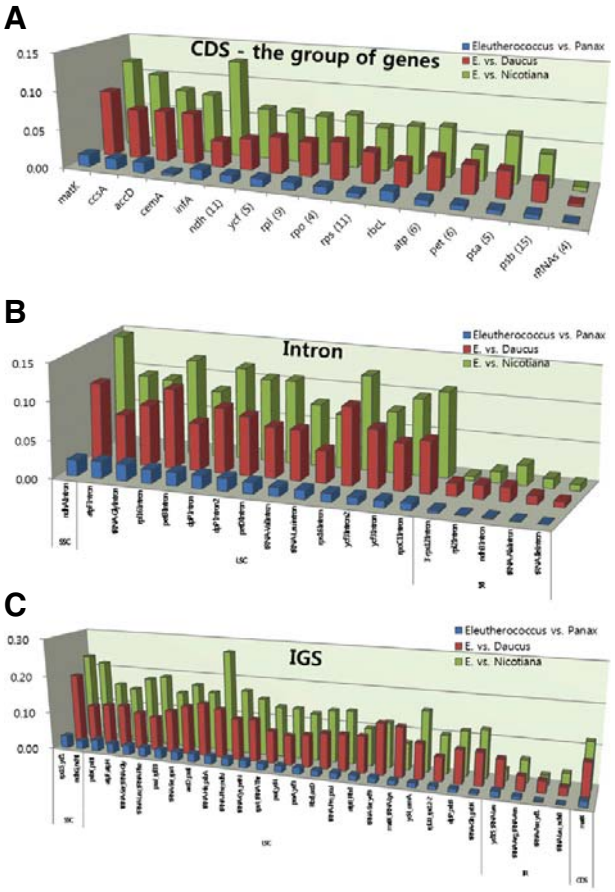

Fig. 5. Comparisons of protein coding genes (CDS), introns, and intergenic spacers (IGS) of the chloroplast genomes in the three different comparisons of *Eleutherococcus vs. Panax*, *Eleutherococcus vs. Daucus*, and *Eleutherococcus vs. Nicotiana*. Y asis indicate the sequence divergences. For the CDS comparisons (top), 86 gene coding regions except trn genes are classified into 16 functional groups (Table 1) and their average sequence diversity is given in the figure. In the intron region comparisons (middle), the low levels of sequence divergences are distinct in the introns that are located on the IR regions. For the IGS region comparisons (bottom), the IGS between the 300 to 800 bp in length are summarized in this figure (Supplementary datas 3-5).

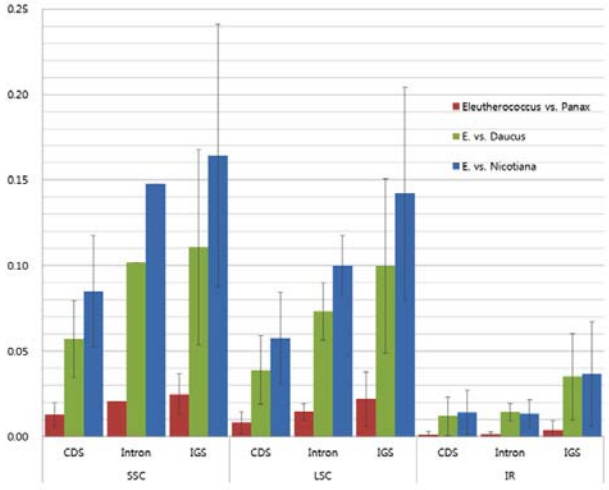

Fig. 6. The levels of evolutionary divergences among the SSC, LSC, and IR regions of cp genomes. Y-axis represents the sequence divergences. The IR region evolves slower than the SSC or the LSC regions regardless the CDS, intron, and IGS.

gence from the IR region. In contrast, all genes in the SSC region showed more than 2.6% divergence between the two families. The *ycf1* gene showed the highest divergence (11.85%), followed by *rpl22* (9.38%), *matK* (8.73%), *rps16* (8.15%), *ndhF* (7.37%), *cemA* (6.70%), *accD* (6.67%), *ccsA* (6.54%) and *atpE* (6.52%). None of these genes are located in the IR region. The average sequence divergences of the gene coding sequences in the IR, LSC, and SSC regions were 1.37, 4.28, and 8.04%, respectively.

Only the *rrn5* gene sequence was identical between the *Eleutherococcus* and *Nicotiana* cp genomes (Supplementary data 5). Most genes in the IR region (for example, *rrn5* - 0.00%; *rrn16* - 0.13%; *rps7* - 0.86%; *rrn4.5* - 0.97%, *rpl2* - 0.97%) had low levels of sequence divergence. All genes in the IR region had less than 2.7% sequence divergence. By contrast, all genes in the SSC region showed more than 4.3% divergence between the two orders. The *ycf1* gene in the SSC region showed the highest divergence (15.46%), followed by *rpl22* (13.64%), *psaI* (13.00%), *matK* (11.82%), *rps32* (11.73%), *ndhF* (11.43%), *ccsA* (6.70%), *accD* (10.22%), *atpE* (9.52%), *psbM* (8.82%). None of these genes are located in the IR region. The average sequence divergences of the gene coding sequences in the IR, LSC, and SSC regions were 1.72%, 6.14%, and 11.49%, respectively.

The sequence divergence data for 84 coding genes were summarized according to 16 different functional groups of genes, as depicted in Fig. 1 and Table 1. The *matK*, *ccsA*, *accD*, *cemA*, and *infA* genes showed high levels of sequence divergence compared to other gene groups. In contrast, the rRNA gene group in the IR region showed the lowest sequence divergence (Fig. 5A). The sequence divergence values of 19 introns were also categorized by IR and SSC regions. The majority of these introns are located on the LSC regions, while there are five and one introns in the IR and the SSC regions, respectively. The five introns in the IR region showed distinctively lower levels of sequence variation than introns in the SSC and LSC regions (Fig. 5B). Therefore, the sequence divergence levels of CDSs, introns, and IGSs were grouped according to the IR, SSC, and LSC regions, respectively (Fig. 6).

Base substitution patterns of IGS regions in the *E. senticosus* cp genome compared to other Asterids

Intergenic spacers (IGS) longer than 11 bp were compared to identify base substitutions. Several of the IGS sequences were identical between *Eleutherococcus* and *Panax* (Supplementary data 3). Approximately 47% (9/19) of the IGSs were identical in the IR region between the two genera, while only 5% of IGSs (4/80) were identical in the LSC region. By contrast, there were no identical IGSs in the SSC region. The sequence divergence of IGSs ranged from 0.00 to 1.88% in the IR region, 0.00 to 8.1% in the LSC region, and 1.1 to 4.27% in the SSC region. The average sequence divergence of IGSs in the IR, LSC, and SSC regions was 0.44, 2.63, and 2.77%, respectively.

Only four short IGS areas were identical between the cp genomes of *Eleutherococcus* and *Daucus* (Supplementary data 4). Two of these are located in the IR region, and another two in the LSC region. By contrast, no sequences were identical in the SSC region. The sequence divergence of IGSs ranged from

Table 6. Comparisons of protein coding genes (CDS), introns, and intergenic spacers (IGS) at the IR, LSC, and SSC regions of the chloroplast genomes

Region		*Eleutherococcus/Panax*							*Eleutherococcus/Daucus*							*Eleutherococcus/Nicotiana*						
		NG	LD (IE)	NP	ND	Ks	Ka	Ka/Ks	NG	LD (IE)	NP	ND	Ks	Ka	Ka/Ks	NG	LD (IE)	NP	ND	Ks	Ka	Ka/Ks
CDS	LSC	62	33 (7)	379	0.0086	0.0216	0.0047	0.22	62	-69 (27)	1843	0.0429	0.0933	0.0302	0.32	62	-252 (39)	2683	0.0614	0.1046	0.0525	0.50
	IR	12	-138 (3)	41	0.0026	0.0006	0.0034	5.67	12	-114 (19)	217	0.0137	0.0292	0.0143	0.49	12	43 (28)	263	0.0172	0.0234	0.0171	0.73
	SSC	12	-129 (7)	287	0.0197	0.0323	0.0162	0.50	12	144 (71)	1151	0.0804	0.1591	0.0646	0.41	12	-96 (66)	1659	0.1149	0.2587	0.0897	0.35
	TOTAL	86	-234 (17)	707	0.0095	0.0182	0.0070	0.38	86	-39 (117)	3211	0.0439	0.0721	0.0375	0.52	86	-401 (133)	4605	0.0627	0.0904	0.0582	0.64
Intron	LSC	13	39	141	0.0149	-	-	-	13	10	691	0.0749	-	-	-	13	18	928	0.1019	-	-	-
	IR	5	3	5	0.0014	-	-	-	5	0	50	0.0139	-	-	-	5	333	46	0.0140	-	-	-
	SSC	1	48	21	0.0208	-	-	-	1	-25	106	0.1019	-	-	-	1	-77	155	0.1478	-	-	-
	TOTAL	19	90	167	0.0118	-	-	-	19	-15	847	0.0610	-	-	-	19	274	1129	0.0840	-	-	-
IGS	LSC	82	678	729	0.0263	-	-	-	80	2135	3046	0.1132	-	-	-	81	169	4331	0.1508	-	-	-
	IR	19	8	25	0.0044	-	-	-	19	-959	258	0.0471	-	-	-	19	208	189	0.0356	-	-	-
	SSC	12	-36	105	0.0277	-	-	-	12	232	476	0.1401	-	-	-	12	194	672	0.2065	-	-	-
	TOTAL	113	650	859	0.0233	-	-	-	111	1408	3780	0.1056	-	-	-	112	571	5192	0.1392	-	-	-
TOTAL		218	506	1733	0.0141	-	-	-	216	1354	7838	0.0638	-	-	-	217	444	10926	0.0879	-	-	-

This is a summary tables of each calculation from three different comparisons of *Eleutherococcus vs. Panax*, *Eleutherococcus vs. Daucus*, and *Eleutherococcus vs. Nicotiana*. The *rps 12* gene is included in the LSC region. Abbreviations: NG, The numbers of genes; LD, the length differences; ID, the indel events; NP, the numbers of polymorphic sites; ND, the nucleotide differences; Ks, the synonymous substitution differences; and Ka, the nonsynonymous substitution differences.

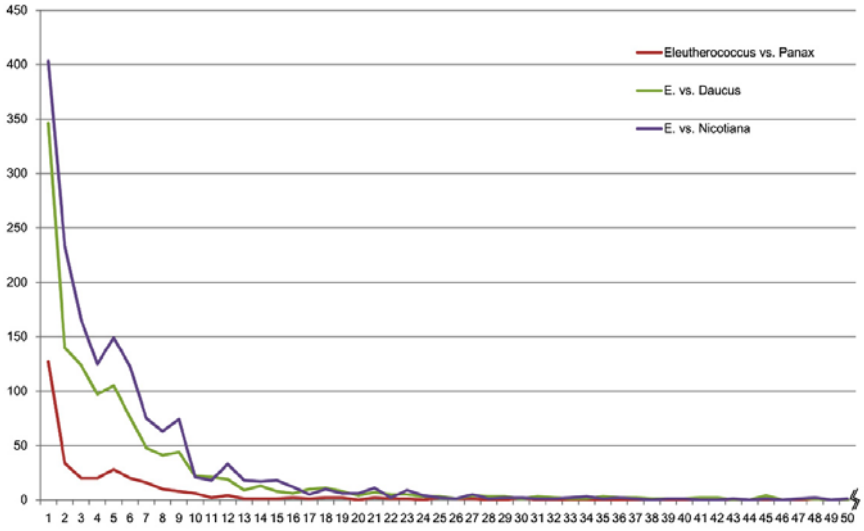

Fig. 7. Indel size and indel number distribution pattern among three cp genomes. The X-axis and Y-axis represent the indel size in base pair and indel numbers, respectively.

0.00 to 8.27% in the IR region, 0.00 to 27.87% in the LSC region, and 6.59 to 18.85% in the SSC region. The average sequence diversity of IGSs in the IR, LSC, and SSC regions was 4.71, 11.32, and 14.01%, respectively.

No IGS regions were identical between the cp genomes of *Eleutherococcus* and *Nicotiana* (Supplementary data 5). The sequence divergence of IGSs ranged from 0.00 to 9.99% in the IR region, 3.90 to 50.0% in the LSC region, and 8.70 to 24.20% in the SSC region. The average sequence diversity of IGSs in the IR, LSC, and SSC regions was 3.56, 15.08, and 20.65%, respectively.

We summarized the sequence divergence values of IGSs whose length ranged from 300 to 800 bp. The spacers of *rps15-ycf1*, *psbK-I*, *atpF-H*, *trnH-psbA*, and rpl20-rps12/2 showed high levels of sequence variation (Fig. 5C). By contrast, IGSs in the IR showed low levels of sequence variation.

Base substitution data from the chloroplast genome of *E. senticosus* and other Asterids clearly indicate that the IR region

of the cp genome evolved at a slower rate than the SSC or the LSC regions, regardless of the function of the sequence. Therefore, the base substitution patterns of the CDS, intron, and IGS can be summarized by the genome regions of SSC, LSC, and IR, respectively (Fig. 6).

The indel patterns of *E. senticosus* cp genome compared to those of other Asterids

The indel patterns of the cp genome of *E. senticosus* were compared to the cp genomes of *Panax, Daucus*, and *Nicotina*. A total of 14 indels from nine genes were identified as indel mutations between the *Eleutherococcus* and *Panax* cp genomes (Fig. 3). Twelve of the 14 indels are triple repeats. Two other frame-shift indels are located at the 3′ ends of the genes. The mutational directions of the indels were determined based on other related sequences. One half of the indels were interpreted as insertion mutations and the other half were determined to be deletion mutations. Indels of more than 12 bp were detected in the *accD*, *ycf2*, and *ycf1* genes (Fig. 3). Two large indels, which corresponded to a 47-bp insertion and a 174-bp deletion, were recorded in the *ycf1* gene. The majority of genes (62/85) were identical in length between the *Eleutherococcus* and *Daucus* cp genomes (Supplementary data 4). Twenty-one genes contained one or two simple indel events that were triple repeats. The genes *accD*, *ycf1*, and *ycf2* have frame-shift mutations resulting from these indel events. Similar patterns of indel mutation were also observed between the cp genomes of *Eleutherococcus* and *Nicotiana* (Supplementary data 5). The majority of genes (68/85) were of identical length, and 11 genes contained simple indels that were triple repeats. However, eight genes (*accD*, *rpoC1*, *rpoC2*, *infA*, *ycf1*, *ycf2*, *ndhF*, and *ycf15*) showed frame-shift mutations.

Indel mutations are widespread in the IGS regions of the *E. senticosus* cp genome. Out of 111 IGS areas, 68 showed length variations between *Eleutherococcus* and *Panax* (Supplementary data 3). The majority (74%, 16/19) of the IGS areas in the IR region were of identical length, while only 29% (27/92) of the IGS areas in the LSC and the SSC regions had conserved lengths. Only 13 of the 113 IGS areas showed no length variation between *Eleutherococcus* and *Daucus* (Supplementary data 4). Six of these are located in the IR region, while the other seven are located in the LSC regions. No IGSs in the SSC regions were identical in length. The majority of the IGS areas of variable length contained multiple indel events. Finally, nine IGS areas showed no length variation between *Eleutherococcus* and *Nicotiana* (Supplementary data 5). Four are located in the IR region, while the other five are located in the LSC regions. No IGSs in the SSC regions were conserved in length.

The indel patterns of chloroplast genomes from the three different hierarchical comparisons are summarized in Fig. 7. The data suggest that similar indel patterns are observed, regardless of the taxonomic hierarchies. The data also indicate that large indels are relatively rare and that the majority of indels are less than 10 bp in length. Furthermore, mapping of indel and base substitution events also indicates that two kinds of mutations occur in the same regions of the cp genomes. When the same comparisons were performed for the intron regions of splicing genes at the same the hierarchical levels, similar base substitution and indel patterns were also observed in the analyses of the intron regions (Supplementary data 3-5).

DISCUSSION

Comparative analysis of cp genomes

Eleutherococcus senticosus is rare in the wild and is listed as an endangered species in Korea due to its over-collection for medicinal use. Although approximately 100 completed cp DNA sequences are available in seed plants (NCBI database), most of these sequences are limited to crop plants or common wild species due to the need for a relatively large amount of material for the isolation of cp DNA. We were able to complete the chloroplast genome sequences of *Eleutherococcus* from small amounts of leaf materials using the combined methods of a series of overlapping long range PCRs (Cheng et al., 1994a; Sambrook and Russell, 2001) with extensive primer designs. The availability of the complete cpDNA sequences from this endangered species provides a rare opportunity for understanding not only the evolutionary history of the chloroplast genome itself, but also the genetic diversity of a rare and endangered species.

The complete chloroplast genome of *Eleutherococcus* is 156,768 bp long with an LSC region of 86,755 bp, an SSC region of 18,153 bp, and two IR regions of 25,930 bp each (Fig. 1). Overall, the genome size, genome structure, gene and intron contents and AT composition of *Eleutherococcus* cp DNA is similar to that of typical land plant cp genomes (Kim and Lee, 2004; Palmer and Stein, 1986; Ruhlman et al., 2006; Shinozaki et al., 1986). We compared the detailed features of the *Eleutherococcus* cp genome to the previously sequenced cp genomes of *Panax, Daucus,* and *Nicotiana.* This comparison represents three different taxonomic hierarchies. For example, *Eleutherococcus* and *Panax* represent the two main generic groups in the same family, Araliaceae (Wen et al., 1998; 2001; APG III, 2009). *Eleutherococcus* and *Daucus* belong to the two core family groups in the same order, Apiales (Plunkett et al., 1996; 1997). Finally, *Eleutherococcus* and *Nicotiana* represent the two main groups (Campanulidae and Lamiidae, respectively) in the Asterid clade (Jansen et al., 2007). Therefore, our comparative data illustrates cp genome evolution patterns in a hierarchy of the Asterid clade. Separate comparative analyses were performed for 86 gene sequences, 19 intron sequences, and 113 IGS sequences.

Chloroplast base substitution patterns

We aligned the sequences of 86 coding genes, 19 introns, and 113 IGSs in three different comparative pairs; *Eleutherococcus* vs. *Panax, Eleutherococcus* vs. *Daucus,* and *Eleutherococcus* vs. *Nicotiana.* The lengths and event numbers of the indels, number of polymorphic sites, nucleotide diversity, synonymous (Ks) and non-synonymous (Ka) substitutions, and Ka/Ks ratios were calculated for each aligned gene region (Supplementary datas 3-5).

The pairwise nucleotide difference (%) data were partitioned into two different ways. First, the data were partitioned according to the functional constraint of the sequences, such as the CDS, introns, and IGS regions. As expected, the CDS region showed the most conserved rate of base substitution in all three comparisons. The base differences of CDS regions were only 0.95% in the *Eleutherococcus/Panax* comparison (same family), 4.39% in the *Eleutherococcus/Daucus* comparison (same order, different family), and 6.27% in the *Eleutherococcus/Nicotiana* comparison (different orders). The nucleotide difference ratio among the CDS, intron, and IGS regions was consistent among the different taxonomic levels. The ratio was approximately 1:1.24:2.31. These data indicate that intron sequences change at a slightly faster rate than the CDSs but slower than IGS sequences.

Second, the data were partitioned according to the positional constraint of sequences, such as the IR, LSC and SSC regions of the cp genomes. As expected, the IR region showed the

most conserved rate of base substitution in all three comparisons. The base differences in the IR region were only 0.28% in the *Eleutherococcus/Panax* comparison (same family), 2.10% in the *Eleutherococcus/Daucus* comparison (same order, different family), and 2.08% in the *Eleutherococcus/Nicotiana* comparison (different orders). The nucleotide difference ratios between the IR, LSC, and SSC regions were somewhat different for the comparisons of different taxonomic levels. The average ratio was approximately 1:4.56:6.12. These data indicate that the SC regions of the cp genome evolve 4-6-times faster than the IR region. Similar base substitution patterns were also reported in previous studies (Kim et al., 2007; Wolfe et al., 1989). The conserved nature of the IR region is primarily due to frequent recombination and the subsequent base collection mechanisms (Maier et al., 1995; Wolfe et al., 1989). To identify the detailed positional constraint, we compared nucleotide diversity values within the same functional constraint. For example, CDSs showed the most conserved nature in the IR region. In the LSC region, CDSs evolved 3-4-times faster than those in the IR region, while the CDSs in the SSC regions evolved 6-8-times faster than those in the IR region (Fig. 6, Table 6). Base substitution patterns showed more differences in nonfunctional regions, such as introns and IGS regions (Figs. 5B and 5C). Intron sequences in the LSC region showed base substitution rates 5-11-times higher than that of the IR regions, while intron sequences in the SSC region evolved 7-14-times faster than sequences in the IR regions. Our data suggest that positional constraint is more critical than functional constraint for the evolution of cp genome sequences. The CDSs encoding *matK, ccsA, accD, cemA* and *infA* genes showed relatively higher nucleotide differences than the CDSs of other functional classes, but the differences are relatively minor compared to the positional effects of the gene. Four rRNA genes located in the IR regions showed distinctively slow base substitution patterns compared to other genes (Fig. 5A). Furthermore, the Ka/Ks ratio of CDS sequences supports a stronger evolutionary constraint in the IR region than in the LSC or SSC regions (Supplementary datas 3-5). Therefore, our data indicate that selective sweeps by the base collection mechanism more frequently eliminate polymorphisms in the IR regions than in other regions.

Evolution of repeated sequences

The function and origin of SSRs in the chloroplast genome are not fully understood. However, SSR loci contained in the plant cp genome can provide useful information about plant population genetics (Echt et al., 1998; Powell, 1995a). SSRs in the cp genome were initially reported in studies of *Pinus radiata* and *Oryza sativa* (Cato and Richardson, 1996; Powell et al., 1995b; Provan et al., 1996). Kim and Lee (2004) also reported 18 SSR loci in the *Panax* cp genome and 29 in the *Nicotiana* cp genome.

Thirty-five SSR loci were identified in the *Eleutherococcus* cp genome (Table 4). Twenty-seven were homopolymers, while six loci were di-polymers and two loci were tri-polymers. Of the 27 homopolymer loci, 22 were composed of multiple A or T bases, while only five homopolymer loci were composed of multiples of C or G. All di- or tri-polymer loci were composed of multiples of AT, TA, ATA or TTA. In the *Eleutherococcus* cp genome, 29 SSR loci are distributed in the IGS region while the other six SSR loci occur in the CDSs. Eleven of the SSR loci in the IGS region and five SSR loci in the CDSs are also found in the *Panax* cp genome. With the exception of these conserved SSR loci, length variation in SSR loci provide useful markers for identifying varieties of plant species and understanding population genetics (Bryan et al., 1999; Garris et al., 2005; Powell et al., 1995b; Xu et al., 2002). The variable SSR loci in *Eleutherococcus* will be useful for elucidating the genetic structures and the genetic variability of this rare and endangered plant species, and will be useful tools for the identification of cultivars and determination of the geographical origin of the medicinally important *E. senticosus*.

In addition to the SSR loci, we also located 18 medium size repeats ranging from 22 to 79 bp in length. Eleven of these repeats are distributed in the IGS or intron regions. Four are dispersed repeats and the others are direct repeats. Half of the direct repeats are palindromic repeats and the other half are tandem repeats. Some of these medium repeats are conserved in different plant groups. For example, the 30-bp repeats on the *trnS* gene are also found in the *Vitis* and *Gossypium* cp genomes. These medium size repeats may aid in cp genome specific gene introduction because the region may use specific recombination sites (Daniell, 1993; Daniell et al., 1998).

Evolution of indel mutations

Indel mutations were analyzed according to the CDSs, introns, and IGSs as well as to the IR, SSC, and LSC regions in the three different comparisons, respectively. A total of 284 indels were identified between *Eleutherococcus* and *Panax*, which belong to the same family, Araliaceae. A majority (75.0%) of these indels occur in the IGS region, while 19.4 and 5.6% of indels occur in the intron and CDS regions, respectively. A total of 1,245 indels were counted between *Eleutherococcus* and *Daucus*, which are members of two different families within a single order. The majority (73.6%) of these indels occur in the IGS region, while 13.9 and 8.9% of the indels occur in the intron and CDS regions, respectively. A total of 1,649 indels were counted between *Eleutherococcus* and *Nicotiana*, which are members of two different orders within the Asterid I clade. The majority (79.0%) of these indels occur in the IGS region, while 12.9 and 8.1% of the indels occur in the intron and CDS regions, respectively. Therefore, more indels are distributed in the IGS regions than in the CDS or intron regions. The majority of indels (85%) are less than 10 bp in length and are concentrated in the LSC and SSC regions of the cp genomes (Fig. 7 and Supplementary datas 3-6). They probably originated from slippage-mispairing mechanisms active during cpDNA replication. To identify the positional effects of indel mutation in the cp genomes, we also calculated the indel frequency per 1 kb cp genome according to the cp genome regions. The indel frequencies per 1 kb in the CDS, intron and IGS regions in the *Eleutherococcus/Panax* comparison were 0.22, 3.86, and 5.15, respectively. The frequencies per 1 kb of CDS, intron and IGS regions in the *Eleutherococcus/Daucus* comparison were 1.44, 12.13, and 23.26, respectively. In addition, the frequencies per 1 kb of CDS, intron and IGS regions in the *Eleutherococcus/Nicotiana* comparison were 1.78, 14.94, and 31.53, respectively. Our data indicate that indel mutations are more negatively selective in the CDS regions than in intron and IGS regions. The intron regions accumulated 8-10-times more indels than the CDS region, and the IGS region accumulates 2-times more indels than the intron regions. Indels are 3-4-times more frequent in the LSC and SSC regions compared to the IR region if the same functional regions are compared. Indel mutation, therefore, also shows some degree of positional constraint, even the though the functional constraint has a more robust effect because of the strong negative selection in the CDS region.

Note: Supplementary information is available on the Molecules

and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This research was supported by a research grant (#KRF-2007-0053690 and KRF-2010-0011796) from Korea Research Foundation to Ki-Joong Kim and by grants (#KEITI 062-091-078 and #KEITI 052-08-071) from Korea Environmental Industry and Technology Institute to Ki-Joong Kim and Byung Yun Sun. We thank three anonymous reviewers for their helpful suggestions for improving the manuscript. All DNA materials used in this study are deposited in Plant DNA Bank of Korea (PDBK) and available from the PDBK.

REFERENCES

APG III. (2009). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. *161*, 105-121.

Bausher, M.G., Singh, N.D., Lee, S.B., Jansen, R.K., and Daniell, H. (2006). The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biol. *6*, 21.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucl. Acids Res. *27*, 573-580.

Bowman, C.M., and Dyer, T.A. (1986). The location and possible evolutionary significance of small dispersed repeats in Wheat ctDNA. Curr. Genet. *10*, 931-941.

Bowman, C.M., Barker, R.F., and Dyer, T.A. (1988). In Wheat ctDNA, Segments of ribosomal-protein genes are dispersed repeats, probably conserved by nonreciprocal recombination. Curr. Genet. *14*, 127-136.

Bryan, G.J., McNicoll, J., Ramsay, G., Meyer, R.C., and De Jong, W.S. (1999). Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. Theor. Appl. Genet. *99*, 859-867.

Cato, S.A., and Richardson, T.E. (1996). Inter- and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* D Don. Theor. Appl. Genet. *93*, 587-592.

Cheng, S., Chang, S.Y., Gravitt, P., and Respess, R. (1994). Long PCR. Nature *369*, 684-685.

Chung, H.J., Jung, J.D., Park, H.W., Kim, J.H., Cha, H.W., Min, S.R., Jeong, W.J., and Liu, J.R. (2006). The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. Plant Cell Rep. *25*, 1369-1379.

Cosner, M.E., Jansen, R.K., Palmer, J.D., and Downie, S.R. (1997). The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr. Genet. *31*, 419-429.

Cosner, M.E., Raubeson, L.A., and Jansen, R.K. (2004). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol. Biol. 4, 27.

Daniell, H. (1993). Foreign gene-expression in chloroplasts of higher-plants mediated by tungsten particle bombardment. Method Enzymol. *217*, 536-556.

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., Lee, S.B., Grevich, J., Saski, C., Quesada-Vargas, T., Guda, C., Tomkins, J., and Jansen, R.K. (2006). Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theor. Appl. Genet. *112*, 1503-1518.

Davydov, M., and Krikorian, A.D. (2000). *Eleutherococcus senticosus* (Rupr. & Maxim.). Maxim. (Araliaceae) as an adaptogen: a closer look. J. Ethnopharmacol. *72*, 345-393.

Doyle, J.J., Davis, J.I., Soreng, R.J., Garvin, D., and Anderson, M.J. (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proc. Natl. Acad. Sci. USA *89*, 7722-7726.

Echt, C.S., DeVerno, L.L., Anzidei, M., and Vendramin, G.G. (1998). Chloroplast microsatellites reveal population genetic diversity in red pine, *Pinus resinosa* Ait. Mol. Ecol. *7*, 307-316.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 1-19.

Funk, H.T., Berg, S., Krupinska, K., Maier, U.G., and Krause, K. (2007). Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Biol. *7*, 45.

Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S., and McCouch, S. (2005). Genetic structure and diversity in *Oryza sativa* L. Genetics *169*, 1631-1638.

Guo, C.H., and Terachi, T. (2005). Variations in a hotspot region of chloroplast DNAs among common wheat and Aegilops revealed by nucleotide sequence analysis. Genes Genet. Syst. *80*, 277-285.

Hachtel, W., Neuss, A., and Vomstein, J. (1991). A Chloroplast DNA inversion marks an evolutionary split in the genus Oenothera. Evolution *45*, 1050-1052.

Hipkins, V.D., Marshall, K.A., Neale, D.B., Rottmann, W.H., and Strauss, S.H. (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 transfer-rna gene. Curr. Genet. *27*, 572-579.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.R., Meng, B.Y., et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome - intermolecular recombination between distinct transfer-rna genes accounts for a major plastid dna inversion during the evolution of the cereals. Mol. Gen. Genet. *217*, 185-194.

Hoot, S.B., and Palmer, J.D. (1994). Structural rearrangements, including parallel inversions, within the chloroplast genome of anemone and related genera. J. Mol. Evol. *38*, 274-281.

Jansen, R.K., and Palmer, J.D. (1987). A chloroplast dna inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Natl. Acad. Sci. USA *84*, 5818-5822.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S.B., Tomkins, J., Alverson, A.J., and Daniell, H. (2006). Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol.Biol. *6*, 32.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. USA *104*, 19369-19374.

Jo, Y.D., Park, J., Kim, J., Song, W., Hur, C.G., Lee, Y.H., and Kang, B.C. (2011). Complete sequencing and comparative analyses of the pepper (*Capsicum annuum* L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. Plant Cell Rep. *30*, 217-229.

Kim, K.J., and Lee, H.L. (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.

Kim, K.J., Choi, K.S., and Jansen, R.K. (2005). Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol. Biol. Evol. *22*, 1783-1792.

Kim, J.S., Jung, J.D., Lee, J.A., Park, H.W., Oh, K.H., Jeong, W.J., Choi, D.W., Liu, J.R., and Cho, K.Y. (2006). Complete sequence and organization of the cucumber (*Cucumis sativus* L. cv. Baekmibaekdadagi) chloroplast genome. Plant Cell Rep. *25*, 334-340.

Kim, Y.K., Park, C.W., and Kim, K.J. (2009). Complete chloroplast DNA sequence from a Korean endemic genus, *Megaleranthis saniculifolia*, and its evolutionary implications. Mol. Cells *27*, 365-381.

Kimura, Y., and Sumiyoshi, M. (2004). Effects of various *Eleutherococcus senticosus* cortex on swimming time, natural killer activity and corticosterone level in forced swimming stressed mice. J. Ethnopharmacol. *95*, 447-453.

Kuang, D.Y., Wu, H., Wang, Y.L., Gao, L.M., Zhang, S.Z., and Lu, L. (2011). Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome *54*, 663-673.

Kurkin, V.A. (2003). Phenylpropanoids from medicinal plants: Distribution, classification, structural analysis, and biological activity.

Chem. Nat. Compd. *39*, 123-153.
Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl. Acids Res. *29*, 4633-4642.
Lee, H.L., Jansen, R.K., Chumley, T.W., and Kim, K.J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Mol. Biol. Evol. *24*, 1161-1180.
Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451-1452.
Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. *25*, 955-964.
Maier, R.M., Neckermann, K., Igloi, G.L., and Kossel, 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.
Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. (2000). VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics *16*, 1046-1047.
McNeal, J.R., Kuehl, J.V., Boore, J.L., and de Pamphilis, C.W. (2007). Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biol. *7*, 57.
Morton, B.R., and Clegg, M.T. (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.
Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., et al. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature *322*, 572-574.
Palmer, J.D. (1986). Isolation and structural analysis of chloroplast DNA. In Methods in Enzymology, A. Weissbach, and H. Weissbach, eds. Vol. 118 (New York: Academic Press), pp. 167-186.
Palmer, J.D. (1987). Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. Am. Nat. *130*, S6-S29.
Palmer, J.D. (1990). Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet. *6*, 115-120.
Palmer, J.D. (1991). Plastid chromosomes: structure and evolution. In Cell Culture and Somatic Cell Genetics in Plants, Vol. 7A, The Molecular biology of Plastids, I.K. Vasil, and L. Bogorad, eds. (San Diego: Academic Press), pp. 5-53.
Palmer, J.D., and Stein, D.B. (1986). Conservation of Chloroplast Genome Structure among Vascular Plants. Curr. Genet. *10*, 823-833.
Plunkett, G.M., Soltis, D.E., and Soltis, P.S. (1996). Higher level relationships of Apiales (Apiaceae and Araliaceae) based on phylogenetic analysis of rbcL sequences. Am. J. Bot. *83*, 499-515.
Plunkett, G.M., Soltis, D.E., and Soltis, P.S. (1997). Clarification of the relationship between Apiaceae and Araliaceae based on matK and rbcL sequence data. Am. J. Bot. *84*, 565-580.
Powell, W., Morgante, M., Andre, C., Mcnicol, J.W., Machray, G.C., Doyle, J.J., Tingey, S.V., and Rafalski, J.A. (1995). Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Curr. Biol. *5*, 1023-1029.
Powell, W., Morgante, M., Mcdevitt, R., Vendramin, G.G., and Rafalski, J.A. (1995). Polymorphic simple sequence repeat regions in chloroplast genomes - applications to the population-genetics of pines. Proc. Natl. Acad. Sci. USA *92*, 7759-7763.
Provan, J., Corbett, G., Waugh, R., McNicol, J.W., Morgante, M., and Powell, W. (1996). DNA fingerprints of rice (*Oryza sativa*) obtained from hypervariable chloroplast simple sequence repeats. Proc. Roy. Soc. Lond. B. Bio. *263*, 1275-1281.
Raubeson, L.A., and Jansen, R.K. (2005). Chloroplast genomes of plants. In diversity and evolution of plants-genotypic variation in higher plants, R. Henry, eds. (Oxfordshire, UK: CABI Publishing)., pp. 45-68.
Ruhlman, T., Lee, S.B., Jansen, R.K., Hostetler, J.B., Tallon, L.J., Town, C.D., and Daniell, H. (2006). Complete plastid genome sequence of *Daucus carota*: Implications for biotechnology and phylogeny of angiosperms. BMC Genomics *7*, 222.
Sambrook, J., and Russell, D. (2001). Molecular cloning: a laboratory manual. Vol. 2, 3rd edition (New York: Cold Spring Harbor), pp. 8.77-8.85.
Samson, N., Bausher, M.G., Lee, S.B., Jansen, R.K., and Daniell, H. (2007). The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. Plant Biotechnol. J. *5*, 339-353.
Saski, C., Lee, S.B., Daniell, H., Wood, T.C., Tomkins, J., Kim, H.G., and Jansen, R.K. (2005). Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol. Biol. *59*, 309-322.
Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchishinozaki, K., et al. (1986). The complete nucleotide-sequence of the tobacco chloroplast genome - its gene organization and expression. EMBO J. *5*, 2043-2049.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. *22*, 4673-4680.
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.
Wen, J., Shi, S.H., Jansen, R.K., and Zimmer, E.A. (1998). Phylogeny and biogeography of Aralia sect. Aralia (Araliaceae). Am. J. Bot. *85*, 866-875.
Wen, J., Plunkett, G.M., Mitchell, A.D., and Wagstaff, S.J. (2001). The evolution of Araliaceae: A phylogenetic analysis based on ITS sequences of nuclear ribosomal DNA. Syst. Bot. *26*, 144-167.
Wolfe, K.H., Gouy, M.L., Yang, Y.W., Sharp, P.M., and Li, W.H. (1989). Date of the monocot dicot divergence estimated from chloroplast DNA-sequence data. Proc. Natl. Acad. Sci. USA *86*, 6201-6205.
Wolfe, K.H., Morden, C.W., and Palmer, J.D. (1992). Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA *89*, 10648-10652.
Wolfson, R., Higgins, K.G., and Sears, B.B. (1991). Evidence for replication slippage in the evolution of Oenothera chloroplast DNA. Mol. Biol. Evol. *8*, 709-720.
Wyman, S.K., Jansen, R.K., and Boore, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. Bioinformatics *20*, 3252-3255.
Xu, D.H., Abe, J., Gai, J.Y., and Shimamoto, Y. (2002). Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. Theor. Appl. Genet.*105*, 645-653.
Yang, M., Zhang, X.W., Liu, G.M., Yin, Y.X., Chen, K.F., Yun, Q.Z., Zhao, D.J., Al-Mssallem, I.S., and Yu, J. (2010). The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS One *5*, 9.
Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucl. Acids Res. *31*, 3406-3415.