Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(4): 124
https://doi.org/10.31083/j.fbl2704124

Original Research

Structural Characterization and Comparative Analyses of the Chloroplast Genome of Eastern Asian Species *Cardamine occulta* (Asian *C. flexuosa* With.) and Other *Cardamine* Species

Gurusamy Raman[1], SeonJoo Park[1,*]

[1]Department of Life Sciences, Yeungnam University, 38541 Gyeongsan, Gyeongsangbuk-do, Republic of Korea
*Correspondence: sjpark01@ynu.ac.kr (SeonJoo Park)
Academic Editor: Kevin Cianfaglione
Submitted: 29 December 2021 Revised: 3 March 2022 Accepted: 11 March 2022 Published: 2 April 2022

Abstract

Background: *Cardamine flexuosa* is considered to be two separate species in the *Cardamine* genus based on their geographical distribution: European *C. flexuosa* and Eastern Asian *C. flexuosa*. These two species have not shown any morphological differences to distinguish each other. Recently, the Eastern Asian species has been regarded as *Cardamine occulta* by their ecological habitats. Therefore, we are interested in analyzing the *C. occulta* chloroplast genome and its characteristics at the molecular level. **Methods**: Here, the complete chloroplast (cp) genome of *C. occulta* was assembled de novo with next-generation sequencing technology and various bioinformatics tools applied for comparative studies. **Results**: The *C. occulta* cp genome had a quadripartite structure, 154,796 bp in size, consisting of one large single-copy region of 83,836 bp and one small single-copy region of 17,936 bp, separated by two inverted repeats (IRa and IRb) regions of 26,512 bp. This complete cp genome harbored 113 unique genes, including 80 protein-coding genes, 29 tRNA, and four rRNA genes. Of these, six PCGs, eight tRNA, and four rRNA genes were duplicated in the IR region, and one gene, *infA*, was a pseudogene. Comparative analysis showed that all the species of *Cardamine* encoded a small variable number of repeats and SSRs in their cp genome. In addition, 56 divergences (Pi $>$ 0.03) were found in the coding (Pi $>$ 0.03) and non-coding (Pi $>$ 0.10) regions. Furthermore, KA/KS nucleotide substitution analysis indicated that thirteen protein-coding genes are highly diverged and identified 29 amino acid sites under potentially positive selection in these genes. Phylogenetic analyses suggested that *C. occulta* has a closer genetic relationship to *C. fallax* with a strong bootstrap value. **Conclusions**: The identified hotspot regions could be helpful in developing molecular genetic markers for resolving the phylogenetic relationships and species validation of the controversial *Cardamine* clade.

Keywords: *Cardamine*; *C. occulta*; chloroplast genome; sequence divergence; phylogeny; hotspot regions

1. Introduction

The chloroplast genomes have a stable and straightforward genetic structure, haploid, and are generally uniparentally transmitted [1]. This organelle is involved in plant cells for nitrogen fixation, photosynthesis, biosynthesis of starch, fatty acids, essential amino acids, and pigments [2,3]. The cp genomes of flowering plants usually have a typical circular structure, 107–280 kb in length, that consists of a large single-copy (LSC) and a small single-copy (SSC) region, which are separated by two large, inverted repeats (IRs) region [4,5]. Owing to the maternal inheritance characteristics, the nucleotide substitution rate of the cp genes is lower than that of nuclear genes but higher than that of the mitochondrial genes. Nevertheless, the rate of plastome genome evolution appears to be taxon and gene-dependent [6]. Therefore, techniques for analyzing the molecular phylogeny of plants are strongly dependent on plastome genome sequence data [7]. Thus far, more than 6100 land plant chloroplast genomes are available at the NCBI organellar genome database, which can be used for comparative studies to resolve the phylogenetic implications of the controversial clade. In addition, recent studies showed that the cp genome encompasses various polymorphic regions at both coding and non-coding regions generated through genomic expansion, contraction, inversion, indel, or genome rearrangement that could be used widely as an effective tool for plant phylogenomic analyses [8].

The genus *Cardamine* (bittercress) is one of the largest genera of the family Brassicaceae and is distributed widely across all the continents except Antarctica [9]. This genus comprises more than 200 exceptionally complex species and remains controversial and unresolved in many circumstances [10]. Among the *Cardamine* taxa, *Cardamine flexuosa* With. is distributed in Europe and Eastern Asia [10]. Moreover, the two taxa have not shown any morphological differences that can be used to distinguish these species [11]. Until 2006, the Eurasian taxa, *C. flexuosa,* is considered a single species [10]. From 2006 onwards, the *C. flexuosa* was considered two different species based on their location [12]. Recent studies showed that these two species differed by their ecological habitats and reported that the Eastern Asian *C. flexuosa* species should be considered *C. occulta* [9]. Differences were also found in the ploidy level. The tetraploid species *C. flexuosa* originated from Europe, whereas the octoploid *C. occulta* Hornem. is from Eastern

Asia and introduced to other continents [11,13]. The diploid species *C. amaraeformis* and *C. hirsuta* are the parental species for *C. flexuosa*. In contrast, the tetraploidy *C. scutata* (Diploid species *C. amaraeformis* and *C. parviflora* as the parents) and *C. kokaiensis* (Diploid species *C. parviflora* as the parent) are the parental for *C. occulta* [9].

Lihova *et al.* [10] reported that the populations of Eastern species *C. occulta* distinguished from the European species *C. flexuosa* based on the phylogenetic studies of internal transcribed spacer (ITS) region of rDNA and the trnL-trnF region of cpDNA. From the biogeographical perspective of *C. flexuosa*, the diploid parental species *C. amaraeformis* is currently absent from Eastern Asia, whereas the tetraploids *C. scutata* and *C. kokaiensis* are distributed in Eastern Asia [14]. Previous studies suggested that the morphologically close species of *C. amaraeformis* are *C. torrentis* Nakai, *C. amariformis* Nakai, and *C. valida*, present in Eastern Asia [15,16]. Therefore, the diploid *C. amaraeformis* may have had a significantly broader dispersal area in the past, reaching easternmost Asia and contributing to multiple polyploidization events there [17]. A previous study characterized the cp genome of the parental species *C. amaraeformis* for *C. occulta* [18]. Therefore, the present study is interested in characterizing the complete chloroplast genome sequence of *C. occulta* (Asian *C. flexuosa* With.), and phylogenetic studies were carried out to resolve this issue. Moreover, there are no extensive comparative studies of the *Cardamine* genera at the whole plastome level. Therefore, this study compared the cp genome of *C. occulta* with other fourteen species of the *Cardamine* genomes and identified hotspot regions that could help develop the molecular markers to distinguish the controversial *Cardamine* species. Overall, this study will provide valuable information for understanding the evolutionary relationship of *C. occulta* in the *Cardamine* clade.

2. Materials and Methods

2.1 DNA Extraction and Sequencing of Cardamine occulta

The fresh young leaves of *Cardamine occulta* were collected from Cheongok Mountain, Bonghwa-gun, South Korea (geospatial coordinates: N37°4′9″, E128°57′47″) and a voucher specimen (YNUH21C064) was deposited in the Yeungnam University Plant Herbarium (http://lifesciences.yu.ac.kr), Gyeongsan, South Korea (Prof. SeonJoo Park, sjpark01@ynu.ac.kr). The entire plant genomic DNA was extracted using a modified cetyltrimethylammonium bromide method [19]. Whole-genome sequencing was performed using a pair-end library (150 × 2), and an insert size of 350 base pairs (bp) using Illumina HiSeq 2500 sequencing system at LabGenomics, Seongnam, South Korea. The read quality was examined with FastQC v0.11.9 [20], and low-quality reads were discarded with Trimmomatic 0.40 [21]. The clean reads were filtered using the GetOrganelle v1.7.4.1 pipeline (https://github.com/Kinggerm/GetOrganelle) to procure plastid-like reads, and the filtered reads were assembled *de novo* method using SPAdes v3.15.2 [22]. The complete chloroplast genome sequence of *C. occulta* and their gene annotation were submitted to GenBank (MZ043777).

2.2 Annotation of C. Occulta Chloroplast Genome

The online program Dual Organeller GenoMe Annotator (DOGMA) was accomplished to annotate the chloroplast genome sequence of *C. occulta* [23]. The initial annotation, putative starts, stops, and intron positions of homologous genes were improved by comparing with the closely related species of *Cardamine*. The transfer RNA genes were confirmed using the tRNAscan-SE version1.21 with default settings [24]. A circular cp genome map of the *C. occulta* was produced using the OrganellarGenome DRAW (OGDRAW) program [25].

2.3 Comparative Chloroplast Genome Analysis of Cardamine Genus

The mVISTA program in the Shuffle-LAGAN model was applied to analyze the cp genome of *C. occulta* with 14 other closely related cp genomes of *Cardamine* genus, applying *C. occulta* annotation as a reference [26]. The boundaries between the IR and SC regions of all the genera of *Cardamine* were also compared and investigated.

2.4 Analysis of the Genetic Divergence in the Cardamine cp Genomes

The genetic divergence was investigated by extracting and aligning the protein-coding genes, intergenic and intron-containing regions of 15 *Cardamine* species cp genome individually using Geneious Prime (Biomatters, New Zealand). The genetic divergence among the *Cardamine* species was estimated using nucleotide diversity (π) and the whole number of polymorphic sites by DnaSP v5 [27]. In this analysis, gaps and missing data were excluded.

2.5 Characterization of the Substitution Rates of Cardamine cp Genomes

The cp genome of *C. occulta* was compared with the other 14 species of *Cardamine* cp genomes to determine the synonymous (K_S) and non-synonymous (K_A) substitution rates. The specific individual functional protein-coding gene exons of these genomes were extracted and aligned separately using Geneious Prime (Biomatters, New Zealand). The aligned sequences were translated into protein sequences and evaluated using DnaSP for K_A and K_S substitution rates without stop codon [27].

2.6 Positive Selection Analysis

Positive selection analysis was carried out based on the substitution analysis of the *Cardamine* species. The site-specific model was applied to estimate the non-synonymous (K_A) and synonymous substitution (K_S) ratio of thirteen protein-coding genes (*atpB*, *ccsA*, *cemA*, *matK*,

ndhA, *ndhD*, *ndhF*, *ndhG*, *ndhJ*, *petA*, *petD*, *rps16*, and *ycf2*) of all *Cardamine* species using EasyCodeML [28]. The sequence of all the thirteen protein-coding genes was aligned separately using the MAFFT program, and the maximum likelihood phylogenetic tree was constructed using RAxML v. 7.2.6 [29]. The codon substitution models M0, M1a, M2a, M3, M7, M8, and M8a were analyzed. The likelihood ratio test was performed to detect the positively selected sites: M0 (one-ratio) vs. M3 (discrete), M1a (neutral) vs. M2a (positive selection) and M7 (β) vs. M8 (β and $\omega > 1$) and M8a (β and $\omega = 1$) vs. M8, which were compared using a site-specific model [28]. The likelihood ratio test (LRT) of the comparison was achieved to evaluate the selection strength. The *p*-values of a Chi-square (χ^2) < 0.05 were considered significant. If the LRT *p*-values were significant (<0.05), the Bayes Empirical Bayes (BEB) method was implemented to identify the codons under positive selection. BEB values higher than 0.95 and 0.99 indicate the sites possibly under positive selection and highly positive selection, which is implied by asterisks and double asterisks, respectively.

2.7 Repeat Sequences and Single Sequence Repeats (SSR) Analysis of Cardamine Genus

The program REPuter was used to determine the presence of repeat sequences in the *Cardamine* cp genomes, including forward, reverse, palindromic, and complementary repeats [30] The following parameters were used to detect repeats in REPuter: (1) Hamming distance 3, (2) minimum sequence identity of 90%, (3) and a repeat size of more than 30 bp. In addition, Phobos software v1.0.6 was used to find the SSRs in *Cardamine* cp genomes; parameters for the match, mismatch, gap, and N positions were set at 1, –5, –5, and 0, respectively [31]. Only one IR region was used in the repeat and SSR marker analyses.

2.8 Phylogenetic Tree Analysis of Brassicaceae

This study used the cp genomes of 38 Brassicaceae species and two outgroup species for phylogenetic analysis based on 68 homologous CDs, LSC, SSC, and IR regions and the whole genomes separately. The 39 completed cp genome sequences were downloaded from the NCBI Organelle Genome Resource database (**Supplementary Table 1**). For ML analysis, the aligned protein-coding gene sequences were saved in PHYLIP format using Clustal X v2.1. Phylogenetic analysis was analyzed using the maximum likelihood (ML) method and the GTRGAMMA model using RAxML v. 8.2.X with 1000 bootstrap replications [29]. The same five individual data sets were also performed by Bayesian Markov chain Monte Carlo (MCMC) inference using the MrBayes v3.2.6 [32,33] phylogenetic tree in Geneious Prime v2022.0.2. The gamma model of rate variation and the HKY85 substitution model were used for this analysis.

3. Results

3.1 General Characteristics of the Cardamine occulta Chloroplast Genome

The complete *Cardamine occulta* chloroplast genome showed a quadripartite structure comprised of 154,796 bp, including a small single-copy (SSC) region of 17,936 bp and a large single-copy (LSC) region of 83,836 bp, which were separated by a pair of inverted repeats (IRa and IRb) of 26,512 bp (Fig. 1; Table 1). The average GC content of the cp genome was 36.3%. The IR regions had the highest GC content (42.4%), followed by the LSC (34%) and SSC regions (29.2%). The *C. occulta* cp genome encoded 113 unique genes: 80 protein-coding genes, 29 tRNAs and four rRNAs. Among the 113 genes, fourteen contained one intron (eight protein-coding and six tRNA genes), and three encoded two introns (*clpP*, *ycf3*, and *rps12*). The *rps12* gene was a trans-spliced gene with its 5'- end exon located in the LSC region and its intron 3'-end exon duplicated in IR regions. In addition, 18 genes were duplicated in the IR regions (**Supplementary Table 2**).

Table 1. General characteristics of the *Cardamine occulta* chloroplast genome.

Genome features	*Cardamine occulta*
Total length (bp)	154,796
LSC length (bp)	83,836
SSC length (bp)	17,936
IR length (bp)	26,512
GC content (%)	36.3
Total genes	131
Genes duplicated in the IR region	18
Protein-coding genes	80
tRNA genes	29
rRNA genes	4

3.2 Comparative Analysis of the Species of Cardamine Genera

The cp genome border LSC-IRb and SSC-IRa of *C. occulta* were compared with the other fourteen species of *Cardamine* genera (Fig. 2). The intact copy of the *rps19* gene was distributed in the LSC/IRb border of all *Cardamine* species and dividends 106 bp to 135 bp in the IRb region resulting in the *rpl2* gene being situated in the IRb region. Similarly, the pseudogene, *ycf1,* and *ndhF* are present in the IRa/SSC border of all the *Cardamine* cp genomes that exhibit overlap. The overlap of these two coding regions was conserved from 30–192 bp in the border of IRa/SSC of their cp genomes. In all the species of *Cardamine* genera cp genomes, the SSC-IRb junction contains the full-length *ycf1* genes, whereas the IRa/LSC junction encodes the fragmented *rps19* and *trnH* genes.

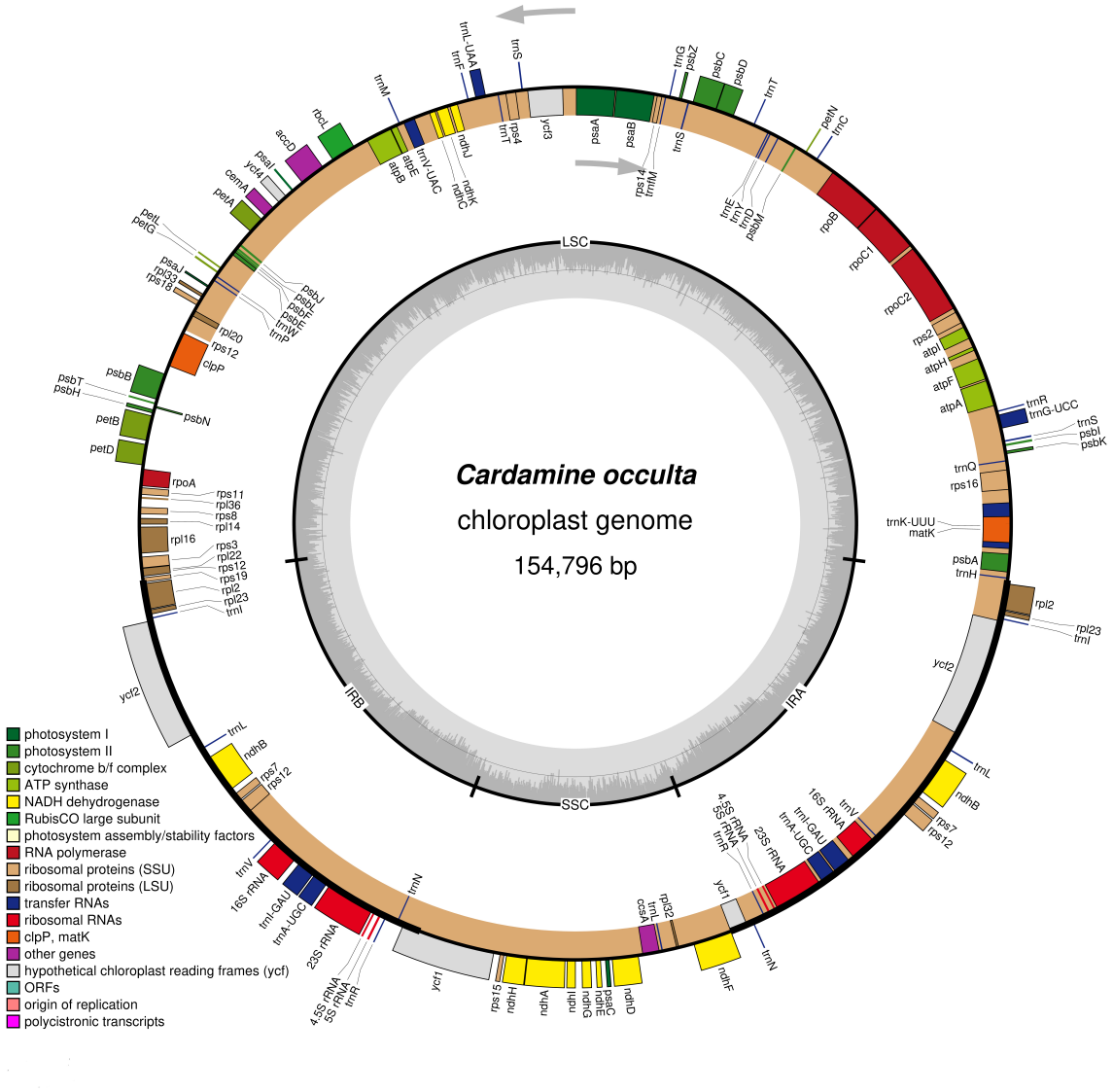

Fig. 1. Gene map of *Cardamine occulta*. Genes lying outside the outer circle are transcribed in a counter-clockwise direction, and genes inside this circle are transcribed in a clockwise direction. The colored bars indicate known protein-coding genes, transfer RNA genes, and ribosomal RNA genes. The dashed, dark grey area in the inner circle represents the GC content, and the light grey area implies the genome AT content. LSC, large single-copy; SSC, small single-copy; IR, inverted repeat.

3.3 Divergence Analysis of cp Sequence and High Variation Region of Cardamine Genera

Genome-wide comparative analyses of the fifteen *Cardamine* cp genomes were achieved using mVISTA to estimate the level of sequence divergence. The cp genomes displayed strong sequence similarity, indicating that the plastomes are highly conserved (Fig. 3). Compared to the non-coding regions and single copy, the coding regions and IR were more conserved, with low variation among *Cardamine*.

3.4 Nucleotide Diversity Analysis

The nucleotide diversity of 204 regions was evaluated using DnaSP software, including 74 protein-coding genes and 128 intergenic and intron regions among fifteen cp genomes of *Cardamine* genera. The results showed that the maximum variable regions (>0.03) in the protein-coding genes that were associated with the photosynthetic process (*cemA*, *clpP*, *ndhA*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *petD*, *petL*, *psaC*, *psbM*, and *ycf4*), transcription and translational process (*rpl14*, *rpl16*, *rpl22*, *rpl33*, *rpoC2*, *rps8*, *rps16*, *rps19*, and *rpl32*) and other processes (*accD*, *ccsA*, and *matK*) (Fig. 4a; Table 2). In addition, the variable regions of the intron and intergenic regions were 0–0.240 Pi (Fig. 4b). Among these divergence hotspots (>0.100), the *rpl32-trnL* region had the highest Pi values (Pi = 0.240) (Table 2).

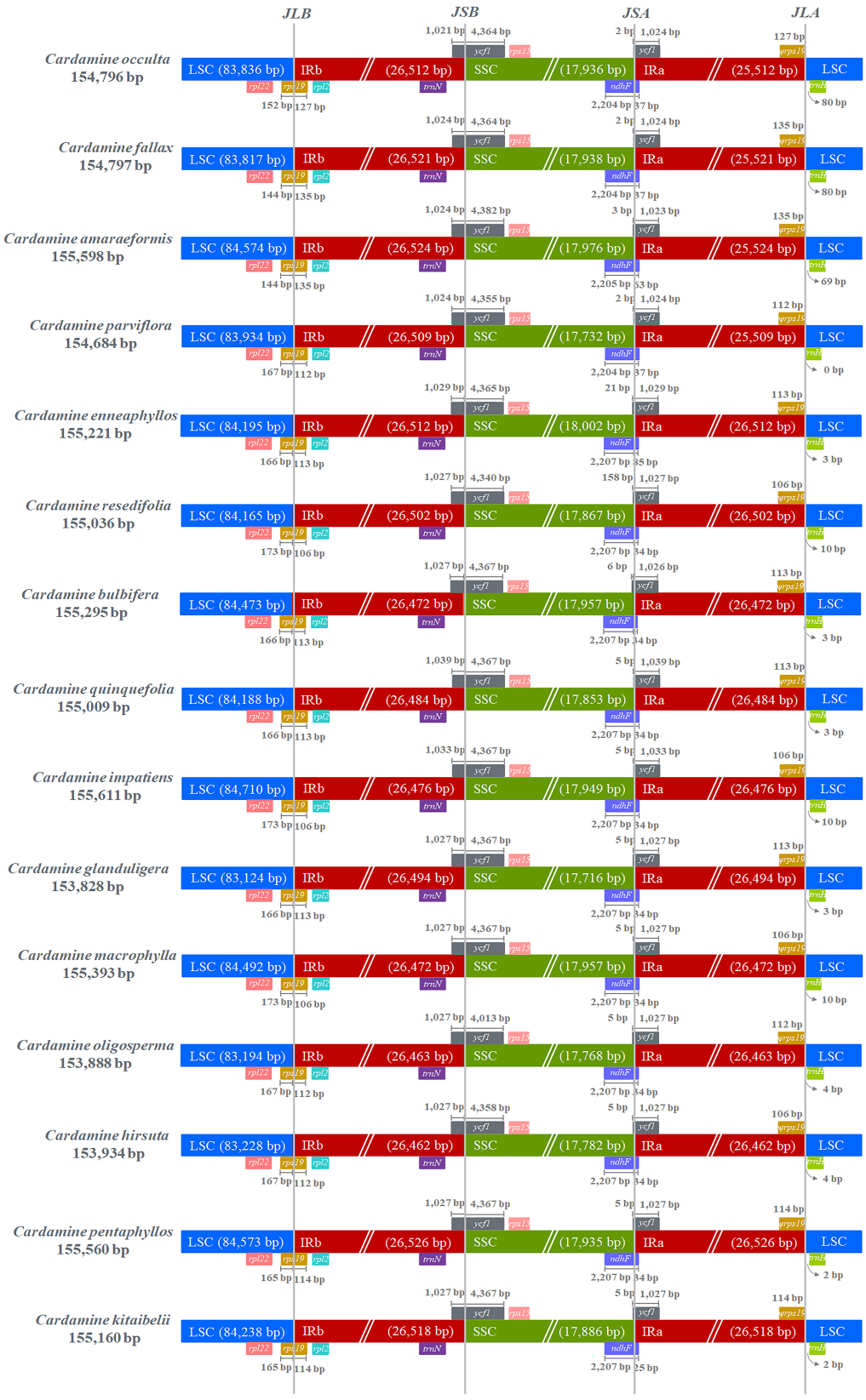

Fig. 2. Evaluation of the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR) border regions of fifteen species of *Cardamine* genera chloroplast genomes. Ψ indicates a pseudogene. The figure is not drawn to scale.

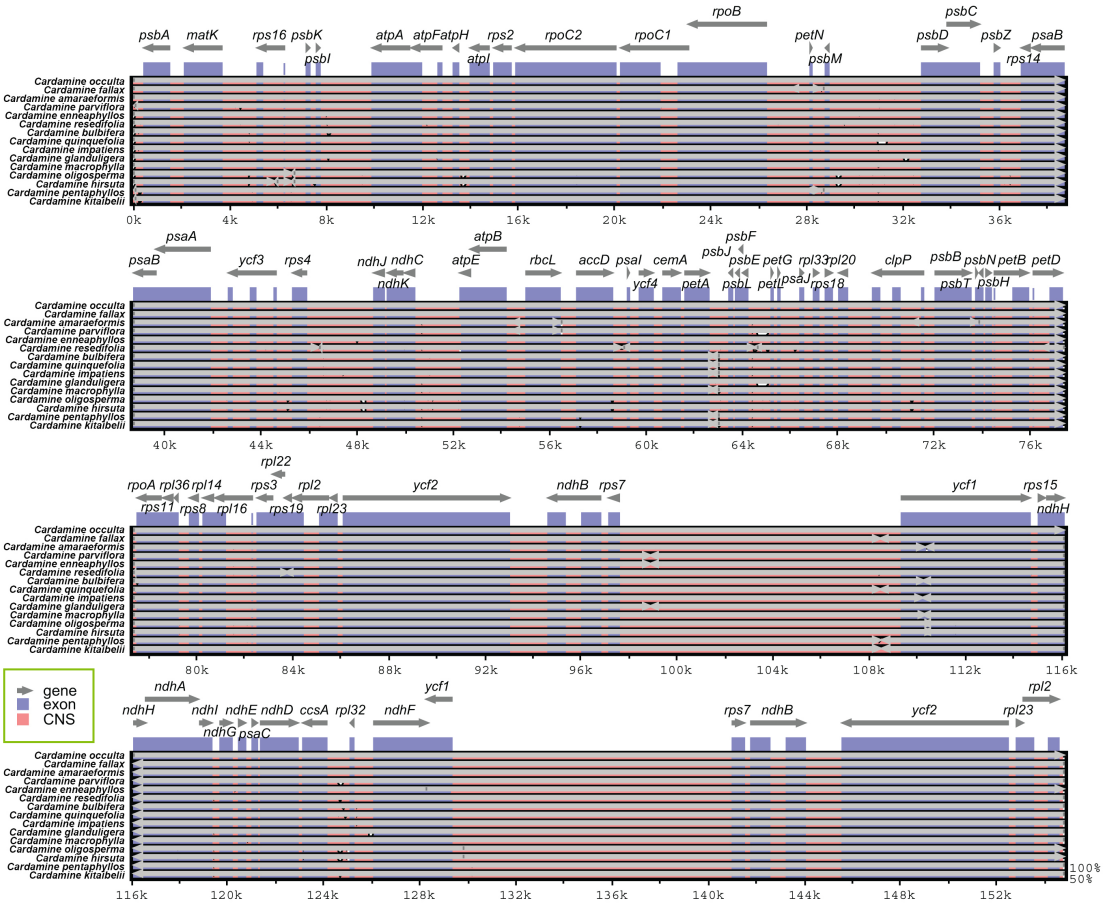

Fig. 3. Sequence alignment of fifteen species of *Cardamine* genera chloroplast genomes performed using the mVISTA program with *Cardamine occulta* as a reference. The top grey arrow shows genes in order (transcriptional direction) and the position of each gene. A 70% cut-off was used for the plots. The y-axis denotes a percent identity of between 50 and 100%, and the red and blue areas imply intergenic and genic regions, respectively.

3.5 Synonymous (K_S) and Non-Synonymous (K_A) Substitution Rate Analysis

Synonymous and non-synonymous substitution rates were calculated for 74 protein-coding genes of fifteen *Cardamine* genera cp genomes. The K_A/K_S ratio of most of the protein-coding all the genes was less than 1, except for the protein-coding genes: *accD* ranged from 0 to 1.3235, *atpB* (0–1.107), *ccsA* (0–1.130), *cemA* (0–1.162), *matK* (0–1.85), *ndhA* (0–1.175), *ndhD* (0–1.25), *ndhG* (0–1.312), *ndhI* (0–1.22), *petA* (0–1.25), *petD* (0–2.61), *rps16* (0–1.96) and *ycf2* (0–2.43). The average ratio of each protein-coding gene for all fifteen *Cardamine* species was calculated individually and plotted in Fig. 5.

3.6 Selective Pressure Events in the cp Genome of Cardamine Genera

The selective pressure of thirteen protein-coding genes, such as four NADH-dehydrogenase subunit genes (*ndhA*, *ndhD*, *ndhG*, and *ndhI*), two subunits of cytochrome (*petA* and *petD*), one ribosome small subunit genes (*rps16*), one subunit of ATP synthase (*atpB*), and *accD*, *ccsA*, *cemA*, *matK*, and *ycf2* of fifteen species of *Cardamine* genera were analyzed based on the substitution rate. If the substitution rate is >1.0 of the individual protein-coding genes between two cp genomes or all the genomes, these genes are considered as under positive selection. The ω_2 values of thirteen genes ranged from 1.0–234.47818 in the M2a model (**Supplementary Table 3**). Furthermore, Bayes empirical Bayes (BEB) analysis was applied to evaluate the location of the consistent selective sites in the thirteen protein-coding genes using M7 vs. M8 model and identified that seven sites under potentially positive selection in the four protein-coding genes (*ccsA* – 2; *matK* – 2; *ndhF* – 2 and *petA* – 1) with posterior probabilities more than 0.95 and 22 sites (*ccsA* – 2; *cemA* – 1; *matK* – 1; *ndhA* – 2; *ndhF* – 2; *ndhG* – 1; *ndhI* – 10; *petA* – 1; *petD* – 1; *ycf2* – 2) greater than 0.99 and the 2ΔLnL value ranged from 0.328019–455.6721 (Table 3). On the other hand, the *atpB*, *ndhD,* and *rps16* did not encode any positively selected sites in their genes.

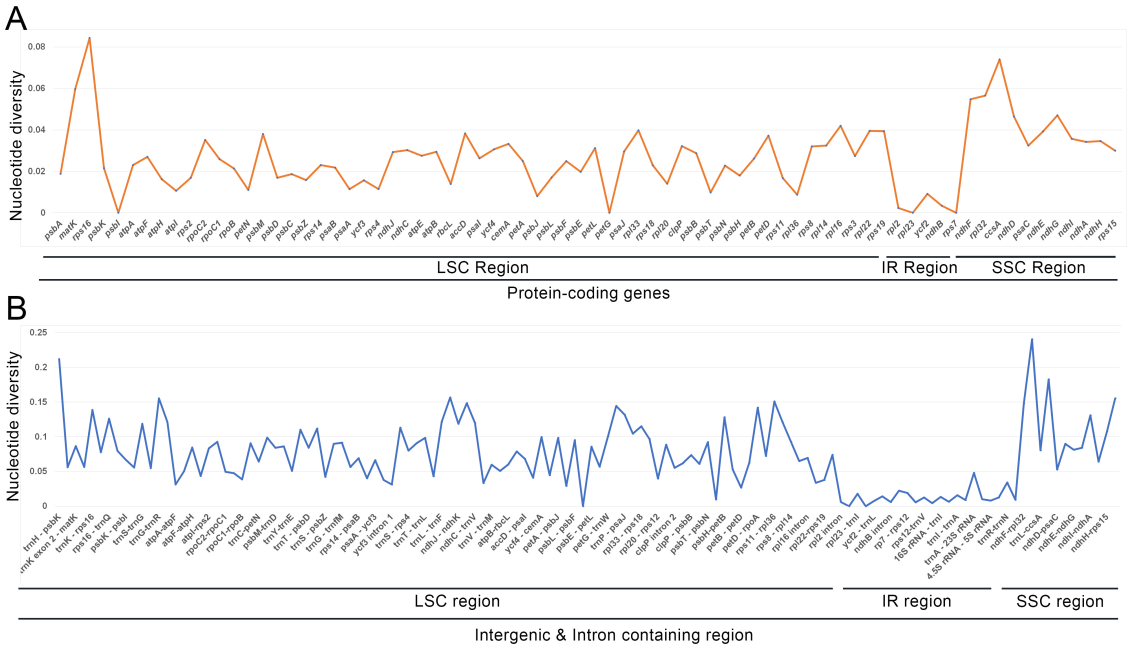

Fig. 4. Genetic diversity based on Kimura's two-parameter model. (A) The P-distance value of protein-coding genes. (B) The P-distance value of intron and intergenic regions.

Fig. 5. Comparison of the ratio of non-synonymous (K_A) to synonymous (K_S) substitutions of 74 protein-coding genes of fifteen species of *Cardamine* cp genomes. (A) Synonymous substitution analysis. (B) Non-synonymous substitution analysis. (C) Non-synonymous vs. synonymous substitution analysis.

Table 2. Mutational hotspot regions in the fifteen species of the *Cardamine* cp genomes.

Protein-coding regions	Nucleotide diversity (Pi)	Aligned length (bp)	No. of variable sites	IGS & Intron regions	Nucleotide diversity (Pi)	Aligned length (bp)	No. of variable sites
matK	0.059642	1509	90	*trnH-psbK*	0.211765	170	36
rps16	0.084388	237	20	*trnK-rps16*	0.138756	418	58
rpoC2	0.03514	4041	142	*rps16-trnQ*	0.125891	421	53
psbM	0.038095	105	4	*trnS-trnG*	0.118677	514	61
ndhC	0.030303	363	11	*trnG-trnR*	0.155405	148	23
accD	0.038298	1410	54	*trnR-atpA*	0.121107	289	35
ycf4	0.030631	555	17	*trnE-trnT*	0.11	400	44
cemA	0.033333	690	23	*psbC-trnS*	0.111732	179	20
petL	0.03125	96	3	*ycf3-trnS*	0.112971	239	27
rpl33	0.039801	201	8	*trnL-trnF*	0.121302	338	41
clpP	0.032149	591	19	*trnF-ndhJ*	0.156334	371	58
petD	0.037267	483	18	*ndhJ-ndhK*	0.118812	101	12
rps8	0.032099	405	13	*ndhK-ndhC*	0.148148	54	8
rpl14	0.03252	369	12	*ndhC-trnV*	0.12	800	96
rpl16	0.041975	405	17	*petG-trnW*	0.100775	129	13
rpl22	0.039583	480	19	*trnW-trnP*	0.144385	187	27
rps19	0.039427	279	11	*trnP-psaJ*	0.131498	327	43
ndhF	0.054807	2226	122	*psaJ-rpl33*	0.104326	393	41
rpl32	0.056604	159	9	*rpl33-rps18*	0.114833	209	24
ccsA	0.073961	987	73	*psbH-petB*	0.12782	133	17
ndhD	0.046481	1506	70	*petD-rpoA*	0.141935	155	22
psaC	0.03252	246	8	*rps11-rpl36*	0.150943	106	16
ndhE	0.039216	306	12	*rpl36-rps8*	0.121212	429	52
ndhG	0.047081	531	25	*ndhF-rpl32*	0.147826	575	85
ndhI	0.035714	504	18	*rpl32-trnL*	0.240175	458	110
ndhA	0.034164	1083	37	*ccsA-ndhD*	0.182692	208	38
ndhH	0.034687	1182	41	*ndhI-ndhA*	0.130952	84	11
				ndhH-rps15	0.107843	102	11
				rps15-ycf1	0.155063	316	49

3.7 Repeat Structure and SSRs Analysis

Repeat sequences were examined in the fifteen *Cardamine* plastomes. Six hundred and ninety-one repeat sequences containing forward, reverse, complement, and palindromic repeats, were observed among the fifteen *Cardamine* plastomes. Three hundred and twenty-two forward (46.6%) and 327 palindromic repeats (47.3%) are relatively common among the detected repeats, whereas 21 (3.04%) of each reverse and complement repeats are comparatively rare (Fig. 6a). The complement repeats were absent in the species of *C. amaraeformis*, *C. enneaphyllos*, *C. hirsuta*, and *C. parviflora*. Similarly, the reverse repeats were absent in the *C. occulta* cp genome. In addition, both reverse and complement repeats were absent in the *C. impatients* and *C. oligosperma* species. In addition, the length of the repeats (>30 bp) was analyzed, and the sizes of the repeats among the fifteen plastomes varied from 30 to 87 bp. Most repeats (472; 70.97%) are limited to 30–39 bp in size (Fig. 6b).

A total of 14,298 simple sequence repeats (SSR) were identified in the fifteen species of *Cardamine* cp genomes with an average of 953 SSRs/genome. The SSRs detected ranged from 938 (*C. parviflora*) to 965 (*C. macrophylla*). The majority of the SSRs were mononucleotide repeats, which accounted for 48.87% of SSRs, followed by hexanucleotide repeats (~15.28%) and pentanucleotide repeats (~10.78%), tetranucleotide repeats (~7.46%), trinucleotide repeats (~6.29%) and dinucleotide repeats (~5.78%) and other repeat lengths from seven-nucleotide to 10-nucleotide repeats comprised 5.53% (Fig. 6c).

3.8 Phylogenetic Analysis

ML and MrBayes analyses were performed separately to determine the phylogenetic position and distance of *C. occulta* precisely. The five individual data sets of a combined total of 68 protein-coding genes, LSC, SSC, and IR regions and whole-genome of 40 cp genome sequences were used to imply the phylogenetic relationships between the closely related species of Brassicaceae. All five of both ML (Fig. 7; **Supplementary Figs. 1–4**) and Bayesian analyses (**Supplementary Figs. 5–9**) yielded similar trees. All the phylogenetic tree analyses showed that the species of *Cardamine* genera formed a monophyletic group. The topology of the phylogenetic tree showed that *C. occulta* has a close relationship with the species of *C. fallax* with a strong bootstrap value (100% for ML and 1.0 for MrBayes) (Fig. 7). Among the *Cardamine* clade, *C. pentaphyllos* and *C. kitaibelii* are the basal groups. The *Cardamine* clade was divided into two clades; *C. bulbifera*, *C. quinquefolia*, *C. impatiens*, *C. glanduligera*, *C. macrophylla*, *C. oligosperma*, and *C. hirsuta* formed one clade, and another clade consisted of *C. occulata*, *C. fallax*, *C. amaraeformis*, *C. parviflora*, *C. enneaphyllos*, and *C. resedifolia* with a 78% bootstrap value.

4. Discussion

The species *Cardamine occulta* is distributed predominantly in Eastern Asia [9]. This species is quite similar to the European species, *C. flexuosa*, and is considered a single species [11] because these two species have not shown any morphological differences. Since 2006, these two species have been differentiated based on their ecological habitats [9–11,13,14]. *Cardamine* is a large genus in the Brassicaceae family of flowering plants that contains more than 200 species of annuals and perennials [9]. Thus far, fourteen chloroplast genomes have been sequenced and analyzed. On the other hand, no extensive and comparative studies of *Cardamine* genera have been carried out. Therefore, the present study sequenced the whole plastid genome of *C. occulta* using Illumina HiSeq 2500 platform and characterized the controversial species from South Korea. Comparative studies were carried out with fourteen other species of the *Cardamine* genera. The length of the complete chloroplast genome sequence of *C. occulta* is 154,796 bp and contains 131 individual genes, which is in the range of other species of *Cardamine* genera. The GC content of *C. occulta* is 36.3%, which is similar to all other species of *Cardamine* genera, suggesting that the distribution of the GC contents in the *Cardamine* cp genomes are consistent and highly conserved. Although the overall genomic structure, such as the gene order and gene number of the *C. occulta*, is identical to other *Cardamine* and Brassicaceae species except for the length of the *atpB* gene in the *C. amaraeformis*. The *Cardamine* plastomes were conserved, and no rearrangement events were found. All the species of *Cardamine* genera lost the protein-coding gene, initiation factor A (*infA*), in their cp genomes. Most of the angiosperms were lost independently from multiple angiosperm lineages, including other species within the Brassicaceae. This gene loss might have been due to an interruption of the nuclear-encoded DNA replication, recombination, and repair machinery that controls the cp genome and the evolution of the plant organelle genome [34].

The results of mVISTA analyses revealed high levels of similarity among the plastomes, indicating that the divergence of the *C. occulta* plastome is lower than that in other species of the *Cardamine* genera. Furthermore, lower sequence divergence in the IR region was detected compared to SC regions, which has been previously reported [8,35–37]. One conceivable reason is that in the cp genome, which has multiple copies per cell, gene conversion with a slight bias in the contradiction of new mutations would reduce the mutation load in the two IR regions much more competently than in the single-copy regions because of the duplicative characteristics of the IRs [38–41]. The expansion and contraction of the IR and single-copy convergence regions are considered the leading mechanism in driving the variation in the size of angiosperm plastomes, playing a vital role in their evolution [38,42–44]. The present study did not identify any significant expansion and shrinkage in the IR/SC

Table 3. Comparison of likelihood ratio test (LRT) statistics of positive selection models against their null models (2△LnL) for fifteen species of *Cardamine* genera.

Protein-coding genes	Comparison between models	2ΔLnL	*d.f.*	*p*-value
atpB	M0 vs M3	3.883652	4	0.421980775
	M1 vs M2A	0.472744	2	0.789486930
	M7 vs M8	1.856066	2	0.395330561
	M8a vs M8	0.473268	1	0.491487565
ccsA	M0 vs M3	99.613282	4	0
	M1 vs M2A	73.543134	2	0
	M7 vs M8	77.341918	2	0
	M8a vs M8	73.073330	1	0
cemA	M0 vs M3	85.525976	4	0
	M1 vs M2A	85.526090	2	0
	M7 vs M8	85.531562	2	0
	M8a vs M8	85.538130	1	0
matK	M0 vs M3	43.252730	4	0
	M1 vs M2A	11.527882	2	0.003138718
	M7 vs M8	7.7505199	2	0.020748942
	M8a vs M8	7.3216859	1	0.006812747
ndhA	M0 vs M3	81.093012	4	0
	M1 vs M2A	81.093202	2	0
	M7 vs M8	81.104690	2	0
	M8a vs M8	81.113370	1	0
ndhD	M0 vs M3	5.764528	4	0.217437134
	M1 vs M2A	0	2	1.0
	M7 vs M8	0.328019	2	0.848733535
	M8a vs M8	0.037706	1	0.846034685
ndhF	M0 vs M3	42.00404	4	0.000000017
	M1 vs M2A	14.27694	2	0.000793965
	M7 vs M8	14.85864	2	0.000593592
	M8a vs M8	14.24146	1	0.000160789
ndhG	M0 vs M3	71.71579	4	0
	M1 vs M2A	64.05478	2	0
	M7 vs M8	31.71765	2	0.000000130
	M8a vs M8	29.72069	1	0.000000050
ndhI	M0 vs M3	696.4341	4	0
	M1 vs M2A	619.6289	2	0
	M7 vs M8	455.6721	2	0
	M8a vs M8	455.7377	1	0
petA	M0 vs M3	93.69115	4	0
	M1 vs M2A	93.73442	2	0
	M7 vs M8	89.69042	2	0
	M8a vs M8	88.736632	1	0
petD	M0 vs M3	42.008744	4	0.000000017
	M1 vs M2A	18.072960	2	0.000118990
	M7 vs M8	28.667442	2	0.000000596
	M8a vs M8	18.07296	1	0.000021260
rps16	M0 vs M3	2.750450	4	0.600415820
	M1 vs M2A	1.880814	2	0.390468882
	M7 vs M8	2.010804	2	0.365897514
	M8a vs M8	1.879698	1	0.170368476
Ycf2	M0 vs M3	25.882614	4	0.000033417
	M1 vs M2A	20.461022	2	0.000036053
	M7 vs M8	18.727454	2	0.000085780
	M8a vs M8	17.756692	1	0.000025103

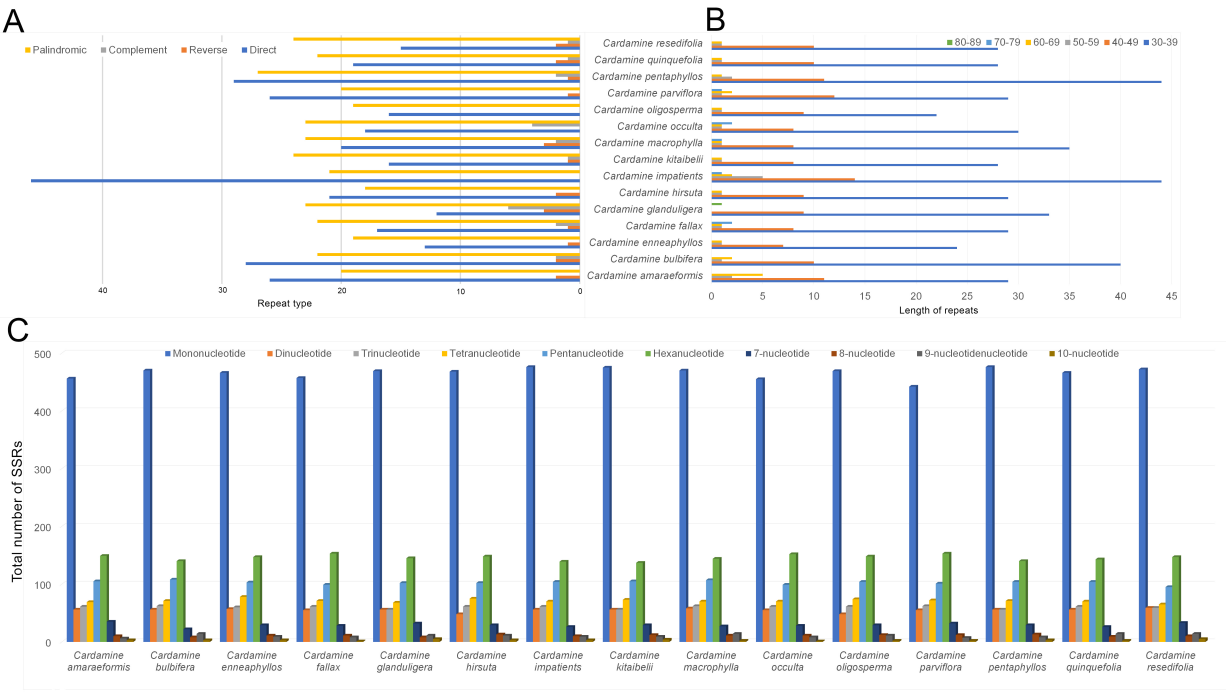

Fig. 6. Comparison of the distribution of different repeat types and SSRs in the fifteen species of *Cardamine* cp genomes. (A) The number of different types of repeats. F—forward repeats; R—Reverse repeats; P— palindromic repeats; C—complement repeats. (B) The length and the total number of repeat sequences present in their respective cp genomes. (C) Distribution of different types of SSRs.

regions. Previous studies reported that the size of the whole cp genome does not always vary with the expansion or contraction of IRs [45–50].

Comparative studies of the fifteen *Cardamine* cp genome sequences revealed several regions of sequence polymorphisms. Among these polymorphisms, most of the sequence variations were dispersed in the LSC and SSC regions, whereas the IR regions displayed relatively lower sequence variations. The lower sequence divergence of the IR region compared to the SC regions in *Cardamine* species and other plants may be due to a copy correction among the IR sequences during gene conversion. Gene mutations and rearrangements in the cp genome are not exhibited constantly throughout the genome sequence. Instead, identifying the hypervariable regions in the chloroplast genome is considered the hotspot region that serves as specific molecular markers [51]. The present study identified 27 protein-coding and 29 intron and intergenic hypervariable regions. Among these, the maximum hypervariable regions, such as the protein-coding genes (>0.050; *ccsA*, *matK*, *ndhF*, *rps16*, and *rpl32*) and intron and intergenic regions (>0.150; *rpl32-trnL*, *trnH-psbK*, *trnG-trnR*, *trnF-ndhJ*, *rpl1-rpl36*, and *rps15-ycf1*) could be used as a DNA barcoding and molecular phylogenetic studies in the *Cardamine* clade (Table 2).

This study analyzed the substitution rate in the fifteen species of the *Cardamine* genera. *C. occulta* was used as a reference genome in the present study and compared with other cp genomes. Initially, the substitution rates of all the individual protein-coding genes of fifteen species of *Cardamine* genera were averaged. The results showed that the ratio of the K_A/K_S rate of all the protein-coding genes was less than 1. In addition, the synonymous and non-synonymous substitution for all the protein-coding genes was analyzed individually (Fig. 5a,b). The results showed that IR regions had low substitution rate than the SC regions. Furthermore, these genes are considered as being under positive selection if the K_A/K_S (ω) rate ratio is >1.0 of individual protein-coding genes between the two cp genomes or all the genomes. Therefore, this study identified thirteen protein-coding genes of fifteen species of *Cardamine* genera that were under selective pressure events: *accD*, *atpB*, *ccsA*, *cemA*, *matK*, *ndhA*, *ndhD*, *ndhG*, *ndhI*, *petA*, *petD*, *rps16*, and *ycf2*. In the selective pressure events, six types of photosynthesis/transcription and translation-related gene groups were categorized: (1) Subunits of ATP synthase (*atpB*); (2) Chloroplast envelope membrane protein (*cemA*); (3) Subunits of NADH dehydrogenase (*ndhA*, *ndhD*, *ndhG*, and *ndhI*); (4) Subunits of cytochrome b/f complex (*petA* and *petD*); (5) One small subunit of the ribosomal gene (*rps16*); (6) Other genes, such as a maturase gene (*matK*), a subunit of the acetyl-coA gene (*accD*), cytochrome synthesis gene (*ccsA*), and unknown function gene (*ycf2*). This variation from unity was attributed to indel and amino acid substitution events in the protein-coding genes of *Cardamine* species. Hence,

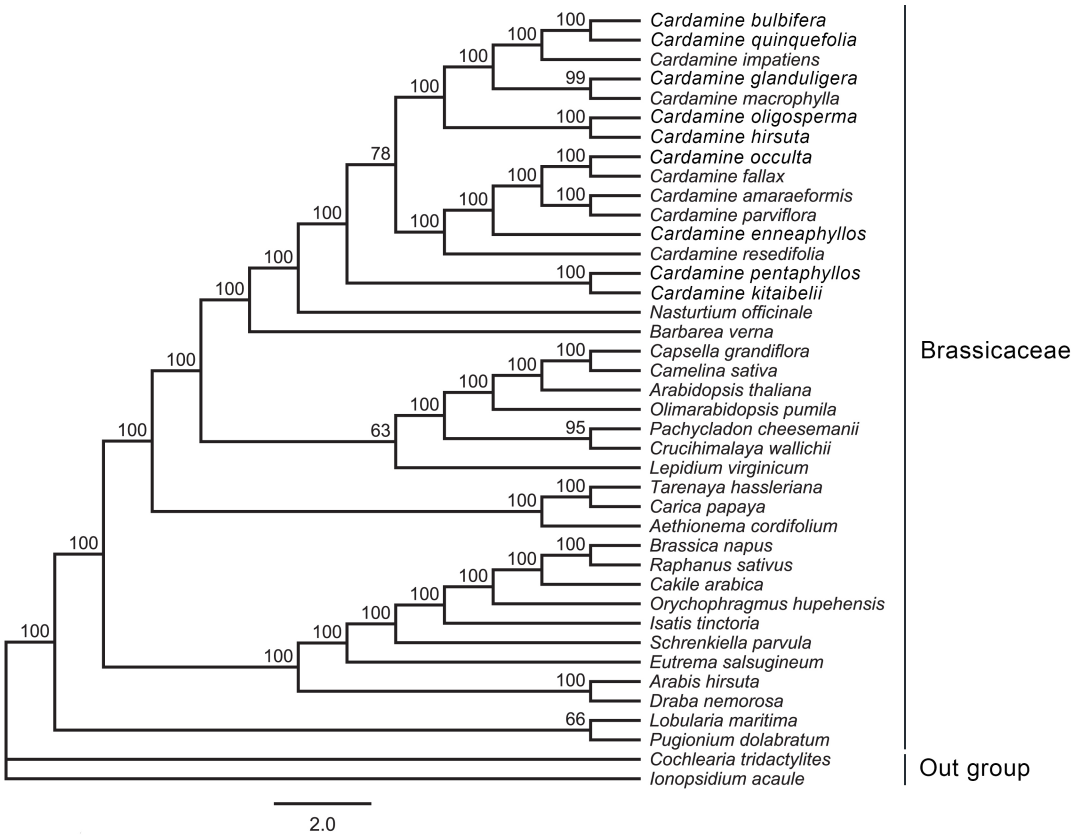

Fig. 7. Molecular phylogenetic tree based on 68 protein-coding genes of 40 Brassicales chloroplast genomes. *Ionopsidium acaule* and *Cochlearia tridactylites* were set as the outgroup. The tree was constructed by maximum likelihood analysis of the conserved regions using the RAxML program and the GTRGAMMA nucleotide model. The stability of each tree node was tested by bootstrap analysis with 1000 replicates. The bootstrap values are shown on the branches, and the branch length reflects the estimated number of substitutions per 1000 sites.

selective analysis of the exons of thirteen protein-coding genes across the publicly available *Cardamine* chloroplast genomes was performed to understand the selective pressure events using site-specific models with four comparison models and LRT values. The positive selective model, M2a, showed that the ω_2 values of seven genes ranged from 1.0–234.47818 (**Supplementary Table 3**). Furthermore, BEB analysis showed that seven sites are under potentially positive selection in the four protein-coding genes (*ccsA*, *matK*, *ndhF*, and *petA*) with posterior probabilities of more than 0.95 and 22 sites greater than 0.99 (Table 3). Nevertheless, no positively selected sites could be determined in the *atpB*, *ndhD,* and *rps16* genes even though some *Cardamine* species have higher ω values. Overall, the nucleotide diversity results show that hotspot mutations in the 11 protein-coding genes (*accD*, *atpB*, *ccsA*, *cemA*, *matK*, *ndhA*, *ndhD*, *ndhG*, *ndhI*, *petD*, and *rps16*) were acquired at a significantly higher rate than expected under neutrality, suggesting that the hotspot mutations are the result of positive selection. The occurrence of hotspot mutation is strong evidence of positive selection, showing that the substitution of a specific amino acid offers an adaptive benefit under specific conditions [52,53]. Previous studies reported that the highly positive selection genes could play a major role in the plant genetic system or photosynthesis process [54–58]. Moreover, the thirteen genes of the fifteen species of the *Cardamine* genera have undergone positive selection, which might be the consequence of adaptation to their diverse habitats. Finally, highly variable regions of both coding and non-coding and thirteen protein-coding genes that were discovered to be under positive selection in the fifteen species of the *Cardamine* genome could be used to produce potential molecular markers for phylogenetic/genomic studies or candidates for DNA barcoding in future studies.

The repeat units were distributed with high frequency and played a substantial role in the chloroplast genome evolution [59–62]. The repeat types of the fifteen *Cardamine* plastomes comprised a variable number in their genomes. Liu *et al.* [60] reported that the variation in number and variety of repeats play a significant role in the plastome structural organization, but there was no correlation between these large repeat regions and rearrangement endpoints. In addition, microsatellite repeats are primarily present in the plastomes, which exhibit a high level of

polymorphism and are used as a molecular marker in genetic studies [63,64]. Simple sequence repeats (SSRs) play a major role during genome rearrangement and recombination [65]. The content of different SSRs and their distribution on various chloroplast regions were similar in their *Cardamine* species. The distribution of SSRs in the *Cardamine* plastome does not involve any genome rearrangement process. On the other hand, the existence of repeat sequences in the cp genome of *Cardamine* genera could be helpful for developing lineage-specific markers for genetic diversity and evolutionary studies.

The whole chloroplast genome of the plant offers a significant foundation to resolve the evolutionary, taxonomic, and phylogenetic studies [58,66–71]. The molecular phylogenetic analysis of both ML and Bayesian analyses in the current study showed that the species of *Cardamine* formed a monophyletic clade. The *Cardamine* clade is subdivided into two clades, and the species *C. occulta is* clustered with *C. fallax* with a substantial bootstrap value. On the other hand, the cytogenetic studies showed that the tetraploidy *C. scutata* (Diploid species *C. amaraeformis* and *C. parviflora* as the parents) and *C. kokaiensis* (Diploid species *C. parviflora* as the parent) are the parental for *C. occulta* [9,11,17]. In contrast, the diploid species *C. amaraeformis* and *C. hirsuta* are the parental species for *C. flexuosa* [9]. Moreover, the *C. fallax* was implied to be hexaploidy. Based on the DNA sequence data of *C. fallax*, it was postulated that this species or its diploid progenitors might have influenced the origin of *C. occulta*. Nevertheless, the cp genomes of *C. flexuosa* (European species), *C. kokaiensis*, and *C. scutata* need to be included to understand the phylogenetic position and their relationship with other *Cardamine* genera in future studies.

5. Conclusions

The complete chloroplast genome *Cardamine occulta* was sequenced, assembled, and analyzed in the present study. Valuable genomic resources were provided for *Cardamine* genera. Overall, the gene contents and arrangements were similar and highly conserved in the species of the *Cardamine* genera. Comparative analyses of the chloroplast genomes identified variable regions with potential application as species-specific DNA barcodes. Furthermore, thirteen protein-coding genes have diverged widely and under potentially positive selection, resulting from adaptation to the ecosystem. Finally, phylogenetic analyses of various cp data sets of both ML and Bayesian analyses show that *C. occulta* species has a closer genetic relationship to *C. fallax*. In conclusion, this study will facilitate future research, particularly resolving the controversial *Cardamine* clade. Nevertheless, in future studies, the cp genome of *C. flexuosa* (European species), *C. kokaiensis,* and *C. scutata* needs to be incorporated to understand the phylogenetic position and their relationship with *C. occulta*.

Data Availability

The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at (https://www.ncbi.nlm.nih.gov) under the accession number MZ043777. The associated BioProject, SRA, and BioSample numbers are PRJNA738458, SRR14833115, and SAMN19729360, respectively.

Abbreviations

cp, chloroplast; LSC, large single-copy; SSC, small single-copy; IR, inverted-repeats; tRNA, transfer RNA; rRNA, ribosomal RNA; K_S, synonymous substitution; K_A, non-synonymous substitution; ω, non-synonymous vs. synonymous ratio; SSR, simple sequence repeats; LRT, likelihood ratio test; π, nucleotide diversity.

Author Contributions

GR and SJP designed the research study. GR performed the research, analyzed the data, and prepared a manuscript draft and figures. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research was funded by grants from Scientific Research (KNA1-1-13, 14-1) of the Korea National Arboretum, Republic of Korea.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2704124.

References

[1] Sun J, Sun R, Liu H, Chang L, Li S, Zhao M, *et al*. Complete chloroplast genome sequencing of ten wild *Fragaria* species in China provides evidence for phylogenetic evolution of Fragaria. Genomics. 2021; 113: 1170–1179.

[2] Neuhaus HE, Emes MJ. Nonphotosynthetic metabolism in plastids. Annual Review of Plant Physiology and Plant Molecular Biology. 2000; 51: 111–140.

[3] Liu J, Qi ZC, Zhao YP, Fu CX, Jenny Xiang QY. Complete cpDNA genome sequence of *Smilax china* and phylogenetic placement of Liliales–influences of gene partitions and taxon sampling. Molecular Phylogenetics and Evolution. 2012; 64: 545–562.

[4] Palmer JD. Comparative organization of chloroplast genomes. Annual Review of Genetics. 1985; 19: 325–354.
[5] Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology. 2016; 17: 134.
[6] Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences of the United States of America. 1987; 84: 9054–9058.
[7] Williams AV, Miller JT, Small I, Nevill PG, Boykin LM. Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia. Molecular Phylogenetics and Evolution. 2016; 96: 1–8.
[8] Raman G, Park SJ. The complete chloroplast genome sequence of the Speirantha gardenii: Comparative and Adaptive Evolutionary Analysis. Agronomy. 2020; 10: 1405.
[9] Mandáková T, Zozomová-Lihová J, Kudoh H, Zhao Y, Lysak MA, Marhold K. The story of promiscuous crucifers: origin and genome evolution of an invasive species, *Cardamine occulta* (Brassicaceae), and its relatives. Annals of Botany. 2019; 124: 209–220.
[10] Lihová J, Marhold K. Phylogenetic and diversity patterns in *Cardamine* (Brassicaceae) – a genus with conspicuous polyploid and reticulate evolution. In Sharma AK, Sharma A (eds.) Plant genome: Biodiversity and evolution (pp. 149–186). Science Publishers: Enfield. 2016.
[11] Marhold K, Šlenker M, Kudoh H, Zozomová-Lihová J. *Cardamine occulta*, the correct species name for invasive Asian plants previously classified as *C. flexuosa*, and its occurrence in Europe. Phytokeys. 2016; 57–72.
[12] Shehbaz IAl, Arai K, Ohba H. *Cardamine*. In Iwatsuki K, Boufford DE, Ohba H (eds.) Flora of Japan: Angiospermae, Dicotyledoneae, Archichlamydeae. Kodansha: Tokyo. 2006.
[13] Mandáková T, Marhold K, Lysak MA. The widespread crucifer species *Cardamine flexuosa* is an allotetraploid with a conserved subgenomic structure. The New Phytologist. 2014; 201: 982–992.
[14] Lihová J, Marhold K, Kudoh H, Koch MA. Worldwide phylogeny and biogeography of *Cardamine flexuosa* (Brassicaceae) and its relatives. American Journal of Botany. 2006; 93: 1206–1221.
[15] Lihová J, Kudoh H, Marhold K. Morphometric studies of polyploid *Cardamine* species (Brassicaceae) from Japan: solving a long-standing taxonomic and nomenclatural controversy. Australian Systematic Botany. 2010; 23: 94.
[16] Doronkin VM. Berberidaceae-Grossulariaceae. In Malyschev LI, Peschkova GA (eds.). Flora Sibiri (63-73). 7. Novosibirsk: Nauka, Russia. 1994.
[17] Šlenker M, Zozomová-Lihová J, Mandáková T, Kudoh H, Zhao Y, Soejima A, *et al*. Morphology and genome size of the widespread weed *Cardamine occulta*: how it differs from cleistogamic *C. kokaiensis* and other closely related taxa in Europe and Asia. Botanical Journal of the Linnean Society. 2018; 187: 456–482.
[18] Raman G, Park KT, Park S. The complete chloroplast genome of an endemic plant to Korea, *Cardamine amaraeformis* Nakai.: genome structure and phylogenetic analysis. Mitochondrial DNA Part B. 2021; 6: 2725–2726.
[19] Doyle JJ. Isolation of plant DNA from fresh tissue. Focus. 1990; 12: 13–15.
[20] Andrews S: FASTQC. A quality control tool for high throughput sequence data. 2010. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 25 September 2021).
[21] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120.
[22] Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics. 2020; 70: e102.
[23] Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20: 3252–3255.
[24] Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Research. 2005; 33: W686–W689.
[25] Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics. 2007; 52: 267–274.
[26] Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Research. 2004; 32: W273–W279.
[27] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25: 1451–1452.
[28] Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW. EasyCodeML: a visual tool for analysis of selection using CodeML. Ecology and Evolution. 2019; 9: 3891–3898.
[29] Stamatakis A, Hoover P, Rougemont J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Systematic Biology. 2008; 57: 758–771.
[30] Kurtz S. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research. 2001; 29: 4633–4642.
[31] Mayer C, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in *Daphnia pulex*–a comparative approach. BMC Genomics. 2010; 11: 277.
[32] Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution. 1997; 14: 717–724.
[33] Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, *et al*. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 2012; 61: 539–542.
[34] Raman G, Lee EM, Park S. Intracellular DNA transfer events restricted to the genus *Convallaria* within the Asparagaceae family: Possible mechanisms and potential as genetic markers for biographical studies. Genomics. 2021; 113: 2906–2918.
[35] Raman G, Park KT, Kim J, Park S. Characteristics of the completed chloroplast genome sequence of *Xanthium spinosum*: comparative analyses, identification of mutational hotspots and phylogenetic implications. BMC Genomics. 2020; 21: 855.
[36] Wang Y, Wang S, Liu Y, Yuan Q, Sun J, Guo L. Chloroplast genome variation and phylogenetic relationships of *Atractylodes* species. BMC Genomics. 2021; 22: 103.
[37] Song Y, Dong W, Liu B, Xu C, Yao X, Gao J, *et al*. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. Frontiers in Plant Science. 2015; 6: 662.
[38] Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. The New Phytologist. 2016; 209: 1747–1756.
[39] Wu C, Chaw S. Evolutionary Stasis in *Cycad* Plastomes and the first Case of Plastome GC-Biased Gene Conversion. Genome Biology and Evolution. 2015; 7: 2000–2009.
[40] Perry AS, Wolfe KH. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. Journal of Molecular Evolution. 2002; 55: 501–508.
[41] Li F, Kuo L, Pryer KM, Rothfels CJ. Genes Translocated into the Plastid Inverted Repeat Show Decelerated Substitution Rates and Elevated GC Content. Genome Biology and Evolution.

2016; 8: 2452–2458.
[42] Kim K, Lee H. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Research. 2004; 11: 247–261.
[43] Asaf S, Waqas M, Khan AL, Khan MA, Kang S, Imran QM, *et al*. The Complete Chloroplast Genome of Wild Rice (*Oryza minuta*) and its Comparison to Related Species. Frontiers in Plant Science. 2017; 8: 304.
[44] Hu G, Cheng L, Huang W, Cao Q, Zhou L, Jia W, *et al*. Chloroplast genomes of seven species of Coryloideae (Betulaceae): structures and comparative analysis. Genome. 2020; 63: 337–348.
[45] Xu W, Xia B, Li X. The complete chloroplast genome sequences of five pinnate-leaved Primula species and phylogenetic analyses. Scientific Reports. 2020; 10: 20782.
[46] Hu Y, Woeste KE, Zhao P. Completion of the Chloroplast Genomes of Five Chinese *Juglans* and their Contribution to Chloroplast Phylogeny. Frontiers in Plant Science. 2017; 7: 1955.
[47] Hong S, Cheon K, Yoo K, Lee H, Cho K, Suh J, *et al*. Complete Chloroplast Genome Sequences and Comparative Analysis of *Chenopodium quinoa* and *C. album*. Frontiers in Plant Science. 2017; 8: 1696.
[48] Cho K, Park T. Complete chloroplast genome sequence of *Solanum nigrum* and development of markers for the discrimination of *S. nigrum*. Horticulture, Environment, and Biotechnology. 2016; 57: 69–78.
[49] Asaf S, Khan AL, Lubna, khan A, Khan A, Khan G, *et al*. Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of *Plantago ovata*. Scientific Reports. 2020; 10: 3881.
[50] Hu G, Wang Y, Wang Y, Zheng S, Dong W, Dong N. New insight into the phylogeny and taxonomy of cultivated and related species of *Crataegus* in China, based on complete chloroplast genome sequencing. Horticulturae. 2021; 7: 301.
[51] Dong W, Liu J, Yu J, Wang L, Zhou S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS ONE. 2012; 7: e35071.
[52] Philippe N, Crozat E, Lenski RE, Schneider D. Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. BioEssays. 2007; 29: 846–860.
[53] Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106: 12412–12417.
[54] Hao DC, Chen SL, Xiao PG. Molecular evolution and positive Darwinian selection of the chloroplast maturase matK. Journal of Plant Research. 2010; 123: 241–247.
[55] Jiang P, Shi F, Li M, Liu B, Wen J, Xiao H, *et al*. Positive Selection Driving Cytoplasmic Genome Evolution of the Medicinally Important Ginseng Plant Genus *Panax*. Frontiers in Plant Science. 2018; 9: 359.
[56] Zhang Z, An M, Miao J, Gu Z, Liu C, Zhong B. The Antarctic sea ice alga *Chlamydomonas* sp. ICE-L provides insights into adaptive patterns of chloroplast evolution. BMC Plant Biology. 2018; 18: 53.
[57] Heyduk K, Moreno-Villena JJ, Gilman IS, Christin P, Edwards EJ. The genetics of convergent evolution: insights from plant photosynthesis. Nature Reviews Genetics. 2019; 20: 485–493.
[58] Li CJ, Wang RN, Li DZ. Comparative analysis of plastid genomes within the Campanulaceae and phylogenetic implications. PLoS ONE. 2020; 15: e0233167.
[59] Dong W, Xu C, Cheng T, Lin K, Zhou S. Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of saxifragales. Genome Biology and Evolution. 2013; 5: 989–997.
[60] Liu W, Kong H, Zhou J, Fritsch PW, Hao G, Gong W. Complete Chloroplast Genome of *Cercis chuniana* (Fabaceae) with Structural and Genetic Comparison to Six Species in Caesalpinioideae. International Journal of Molecular Sciences. 2018; 19: 1286.
[61] Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD, He XJ. Comparative analysis of the chloroplast genomes of the Chinese endemic genus *Urophysa* and their contribution to chloroplast phylogeny and adaptive evolution. International Journal of Molecular Sciences. 2018; 19: 1847.
[62] Shi H, Yang M, Mo C, Xie W, Liu C, Wu B, Ma X. Complete chloroplast genomes of two *Siraitia* Merrill species: Comparative analysis, positive selection and novel molecular marker development. PLoS ONE. 2019; 14: e0226865.
[63] Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, *et al*. An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. Theoretical and Applied Genetics. 1997; 95: 163–173.
[64] Kawabe A, Nukii H, Furihata HY. Exploring the history of chloroplast capture in *Arabis* using whole chloroplast genome sequencing. International Journal of Molecular Sciences. 2018; 19: 602.
[65] Mrázek J. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Molecular Biology and Evolution. 2006; 23: 1370–1385.
[66] Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD. The chloroplast ycf3 and ycf4 open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex. The EMBO Journal. 1997; 16: 6095–6104.
[67] Yang J, Tang M, Li H, Zhang Z, Li D. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evolutionary Biology. 2013; 13: 84.
[68] Raman G, Park S. The Complete Chloroplast Genome Sequence of *Ampelopsis*: Gene Organization, Comparative Analysis, and Phylogenetic Relationships to other Angiosperms. Frontiers in Plant Science. 2016; 7: 341.
[69] Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, *et al*. The Complete Chloroplast Genome Sequences of Five Epimedium Species: Lights into Phylogenetic and Taxonomic Analyses. Frontiers in Plant Science. 2016; 7: 306.
[70] Kahraman K, Lucas SJ. Comparison of different annotation tools for characterization of the complete chloroplast genome of *Corylus avellana* cv Tombul. BMC Genomics. 2019; 20: 874.
[71] Li X, Zuo Y, Zhu X, Liao S, Ma J. Complete Chloroplast Genomes and Comparative Analysis of Sequences Evolution among Seven *Aristolochia* (Aristolochiaceae) Medicinal Species. International Journal of Molecular Sciences. 2019; 20: 1045.