RESEARCH ARTICLE

Characterization of the complete chloroplast genome of *Arabis stellari* and comparisons with related species

Gurusamy Raman¹, Veronica Park², Myounghai Kwak³, Byoungyoon Lee³, SeonJoo Park¹*

¹ Department of Life Sciences, Yeungnam University, Gyeongsan, Gyeongsan-buk, Republic of Korea, ² Mcneil high school, Austin, Texas, United States of America, ³ Plant Resources Division, National Institute of Biological Resources of Korea, Incheon, Republic of Korea

* sjpark01@ynu.ac.kr

Abstract

*Arabis stellari* var. *japonica* is an ornamental plant of the Brassicaceae family, and is widely distributed in South Korea. However, no information is available about its molecular biology and no genomic study has been performed on *A. stellari*. In this paper, the authors report the complete chloroplast genome sequence of *A. stellari*. The plastome of *A. stellari* was 153,683 bp in length with 36.4% GC and included a pair of inverted repeats (IRs) of 26,423 bp that separated a large single-copy (LSC) region of 82,807 bp and a small single-copy (SSC) region of 18,030 bp. It was also found to contain 113 unique genes, of which 79 were protein-coding genes, 30 were transfer RNAs, and four were ribosomal RNAs. The gene content and organization of the *A. stellari* chloroplast genome were similar to those of other Brassicaceae genomes except for the absence of the *rps16* protein-coding gene. A total of 991 SSRs were identified in the genome. The chloroplast genome of *A. stellari* was compared with closely related species of the Brassicaceae family. Comparative analysis showed a minor divergence occurred in the protein-coding *matK*, *ycf1*, *ccsA*, *accD* and *rpl22* genes and that the *Kₐ/Kₜ* nucleotide substitution ratio of the *ndhA* genes of *A. stellari* and *A. hirsuta* was 1.35135. The genes *infA* and *rps16* were absent in the *Arabis* genus and phylogenetic evolutionary studies revealed that these genes evolved independently. However, phylogenetic analysis showed that the positions of Brassicaceae species are highly conserved. The present study provides *A. stellari* genomic information that may be found useful in conservation and molecular phylogenetic studies on Brassicaceae.

Introduction

Chloroplasts are the most noticeable feature in green plant cells and are specific to plants. The chloroplast is a semi-autonomous organelle that was derived from a cyanobacterial endosymbiont around one billion years ago [1, 2]. Plastids are involved in several critical biochemical processes other than photosynthesis, such as, starch biosynthesis, nitrogen metabolism, sulfate
reduction, fatty acid synthesis, and DNA and RNA synthesis [3]. The high copy number of plastomes in plant cells is inherited maternally in most plant cells, and the chloroplast genome varies in size from 75 to 250 kb and is highly conserved in terms of gene contents and genome structure in vascular plants [4, 5]. Chloroplasts are normally separated by two large inverted repeat regions separated by a large single-copy region (LSC) and small single-copy region (SSC) that vary in length. Currently, more than 1100 genomes are available in the chloroplast genome database. Comparative studies on these genomes have shown some infrequent structural changes, such as, gene or intron loss, large inverted repeat (IR) expression, inversions and rearrangements in many land plants [6]. For example, intron loss was observed in the clpP gene of Sileneae [7], infA gene loss in Brassicales, Cucurbitales, Fabales, Fagales, Malphighiales, Malvales, Myrtales, Rosales, Sapindales, Solanales, Dianthus, and Lychinis [8–12], rpl22 gene loss in Fabaceae and Passifloraceae [13], rpl23 loss in Dianthus, Lychnis and Spinacia [12, 14], rpl32 gene loss in Populus [15], ycf2 gene loss in rice and maize [16, 17], and ycf4 gene loss in all legume plants of angiosperms [18, 19]. Such studies provide information for plant phylogenetic tree reconstruction [20], DNA barcoding [21], and for population [22], transplastomic, and evolutionary studies [23].

The herbaceous Brassicaceae plants are distributed worldwide. They Brassicaceae family is composed of more than 3700 species, and includes vegetable and vegetable oil crops, ornamentals, and model species [6]. The ornamental plant, A. stellari var. japonica also belong to this family and is widely distributed in Russia, Taiwan, Japan, and South Korea. It grows up to a height of 30 centimeters, is sparsely to densely pilose, has erect or ascending stems, is basal and cauline, and it a popular garden plant. To the best of our knowledge, no previous molecular or genomic study has been carried out in this ornamental plant and its plastome sequence has not been reported. In the present study, we sought to determine the complete chloroplast genome sequence of A. stellari, to describe the structure of the plastome genome, and to compare its plastome genome with those of closely related Brassicaceae species. Accordingly, we sought to expand understanding of the diversity of Arabis chloroplast genomes and provide basic data for phylogenetic studies on Brassicaceae.

**Materials and methods**

**DNA extraction and sequencing**

The A. stellari plant sample was collected on Dokdo island (South Korea). DNA was extracted using a modified CTAB method [24]. Whole-genome sequencing was performed using Illumina NextSeq 500 (LabGenomics, South Korea) technology and a paired-end library of 2x101 bp and insert size of ~200 bp. About 152,770,066 raw reads were trimmed and filtered using Genious v10.1 (Biomatters, New Zealand). Filtered reads were assembled using A. alpina (NC_023367) as a reference genome. Consensus sequences were extracted and specific primers were designed based on gaps between sequences and these gaps were filled by polymerase chain reaction (PCR) amplification. PCR products were purified and sequenced using the conventional Sanger sequencing method. The chloroplast genome sequencing data and gene annotation were submitted to GenBank and assigned the accession number KY126841.

**Chloroplast genome annotation and sequence statistics**

The online program Dual Organeller GenoMe Annotator (DOGMA) was used to annotate the A. stellari cp genome [25]. The initial annotation results were checked manually and putative starts, stops, and intron positions were adjusted by comparing them with closely related homologous genes of A. alpina, A. hirsuta, and Arabidopsis thaliana. Transfer RNA genes were
verified using tRNAscan-SE version 1.21 and default settings [26]. The OGDRAW program was used to draw a circular map of the *A. stellari* cp genome [27].

**Comparative genome analysis**

The mVISTA program in Shuffle-LAGAN mode was used to compare the *A. stellari* cp genome with four other cp genomes using *A. stellari* annotation as a reference [28]. The boundaries between IR and SC regions of these species were also compared and analyzed.

**Analysis of repeat sequences and single sequence repeats (SSR)**

REPuter software was used to identify the presence of repeat sequences, including forward, reverse, palindromic, and complementary repeats in the cp genome of *A. stellari* [29]. The following conditions were used to identify repeats in REPuter: (1) Hamming distance 3, (2) minimum sequence identity of 90%, (3) and a repeat size of more than 30 bp. Phobos software v1.0.6 was used to detect SSRs of cp genome; parameters for match, mismatch, gap, and N positions were set at 1, -5, -5 and 0, respectively [30].

**Characterization of substitution rates**

To analyze synonymous (*K_ S* ) and nonsynonymous (*K_ A*) substitution rates, the *A. stellari* cp genome was compared with the cp genome sequences of *A. alpina* and *A. hirsuta*. Similar individual functional protein-coding gene exons were extracted and aligned separately using Geneious v10.1.3. Aligned sequences were translated into protein sequences and *K_ S* and *K_ A* rates were estimated using DnaSP software v5.10.01 [31].

**PCR amplification of the *rps16* gene**

The genomic DNA of *A. stellari* was used as a template to detect the *rps16* gene and gene specific primers were designed using Primer3 v0.4.0 [32]. The *rps16* gene was amplified using the primers (*rps16*F: 5'–ACCAAGCTATATACGAGTCTTTCA–3' and *rps16*R: 5'–ACGATACTGACTGAACTATGACT–3'), and the PCR product was purified using the Solgel™ Gel & PCR purification System Kit (Solgent Co., Daejeon, South Korea). Purified PCR products were sequenced with an ABI 3730XL DNA analyzer (Applied Biosystems, Foster City, USA) at Solgent. The nucleotide sequence of *rps16* was aligned using MAFFT v7 [33] in Geneious v10.1.3 ( Biomatters, New Zealand).

**Phylogenetic analysis**

A phylogenetic tree was constructed using 76 protein-coding genes of 20 cp genomes of angiosperms using the *Vitis* set as the outgroup. The 19 completed cp genome sequences were downloaded from the NCBI Organelle Genome Resource database S1 Table. *rps16*, *ycf15*, and 76 protein-coding gene sequences were aligned separately using MAFFT v7 [33] through Geneious v10.1.3. The aligned individual gene sequences and protein-coding gene sequences were saved in PHYLIP format using Clustal X v2.1 [34] and phylogenetic analysis was performed based on maximum likelihood (ML) analysis using the general time-reversible model and the gamma model site heterogeneity (GTR+GAMMA) nucleotide substitution model using default parameters in RAxML v. 7.2.6 [35]. The bootstrap probability of each branch was calculated using 1000 replications.
Results and discussion

Genome organization and features of the A. stellari cp genome

The complete chloroplast genome of A. stellari was found to have a total length of 153,683 bp, with a pair of inverted repeats (IRs) of 26,423 bp that separated a large single copy (LSC) region of 82,807 bp and a small single copy (SSC) region of 18,030 bp (Fig 1). Total GC content was 36.4%, which is similar to those of A. alpina [36], Draba nemorosa, and Brassica napus [37].
whereas GC contents are low in the species A. hirsuta (33.0%) and Arabidopsis thaliana (32.1%) [38]. These results suggest that GC contents are unevenly distributed in the genomes of the Brassicaceae family. In A. stellari, GC content was higher in the IRs region (42.4%) than in the LSC and SSC region (34.1% and 29.2%). The high GC content percentage in IR regions was attributed to the presence of high GC nucleotide percentages in the four rRNA genes rrn4.5, rrn5, rrn16, and rrn23. Identical results have been reported for other chloroplast genomes [39, 40].

The chloroplast genome of A. stellari encoded a total of 113 unique genes, of which 18 were duplicated in IR regions. Of the 113 genes, 79 were protein-coding genes, 30 were transfer genes and four were rRNA genes Table 1. Of these, 14 genes encoded one intron (eight protein-coding and six tRNA genes) and three encoded two introns (clpP, ycf3 and rps12). The rps12 gene was found to be 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 the total A. stellari cp genome, protein-coding regions accounted for 79,437 bp (51.68%), intron regions for 19,688 bp (12.82%) and tRNA and rRNA regions for 2,785 bp (1.81%) and 9,049 bp (5.89%) respectively. The remaining regions were intergenic spacers (42,724 bp, 27.8%). The pseudogene, rps16 was identified in the LSC region. Overall, the gene order and gene contents of A. stellari were identical to those of A. alpina and A. hirsuta.

Comparisons of the A. stellari cp genome and those of other Brassicaceae species

The cp genome of A. stellari was compared with four closely related Brassicaceae family cp genomes, namely with those of A. alpina, A. hirsuta, Brassica napus, and A. thaliana. The organization of the Brassicaceae cp genome is highly conserved, and neither translocations nor inversions were identified in the analyses. However, two dissimilarities were identified involving the protein-coding genes rps16 and ycf15, and some differences between total genome sizes were detected. The shortest genome was that of Brassica napus (152,860 bp) and the longest that of Pugionium dolabratum (155,002 bp). These differences were largely due to variabilities in the length of the LSC region. Similar genome size variations in the LSC region were observed in rosid chloroplast genomes [12].

The overall sequence variation of five Brassicaceae family cp genomes was plotted using the mVISTA program, and the results obtained revealed that cp genomes within Brassicaceae are highly conserved (Fig 2). However, minor divergences were detected in protein-coding regions. In order to analyze divergent hotspot regions further, all coding regions of A. stellari, A. alpina, and A. hirsuta were extracted and evaluated. The most divergent regions found were in the protein coding genes matK, ycf1, ccsA, accD, and rpl22 (Fig 3), which are present in the large and single copy regions.

Due to the size variation exhibited by angiosperm chloroplast genomes, expansion and contraction at IR/SC borders are more common in chloroplast genomes [41]. In the present study, the LSC/IRb/SSC/IRa junctions of the five Brassicaceae family chloroplast genomes were compared (Fig 4). The lengths of the LSC, IR and SSC regions were similar in the cp genomes of A. stellari, A. alpina, A. hirsuta and D. nemorosa as compared with B. napus and A. thaliana; although some variances in IR expansions and contractions were detected. The rps19 gene was present in the LSC region and expanded in the IR region in all six cp genomes. Also, the pseudogene ycf1 was completely present in the IR region. Likewise, the ndhF genes of A. stellari, A. hirsuta, D. nemorosa, B. napus and A. thaliana were completely contained in the SSC region. Whereas the ndhF gene of A. alpina was extended and overlapped with pseudogene ycf1 in the IRb region. Similarly, the tRNA gene, trnH-GUG was entirely positioned in IRa region of all
chloroplast genomes except that of *A. stellari*. Nevertheless, 3 bp of the *trnH* gene in *A. stellari* overlapped the IRa region.

**Repeat and SSR analysis**

The REPuter program was used to screen repeat sequences in the *A. stellari* chloroplast genome. The results obtained showed the following were present; 30 forward repeats, 23

| Table 1. List of genes present in the cp genome of *Arabis stellari*. |
|------------------------|------------------------|------------------------|------------------------|
| **Category** | **Group of genes** | **Name of genes** | **Name of genes** |
| RNA genes | Ribosomal RNA genes | mrNA.5" | mrN5" |
| | Transfer RNA genes | trnC-GCA | tmD-GUC |
| | | trnG-GCC" | trnG-UCU |
| | | trnH-GUG" | trnL-CAA" |
| | | trnL-UAAb | trnL-UGAb |
| | | trnP-UGG | trm7-UGG |
| | | trnR-ACG" | trmS-AGA |
| | | trnS-UGA | trnT-GGU |
| | | trnT-GGU | trnW-CCA |
| | | trnY-GUA | trnY-GUA |
| Protein genes | Subunits of photosystem I | psaA | psaB |
| | | psaC | psaI |
| | | psaJ | psaJ |
| | Subunits of photosystem II | psbA | psbB |
| | | psbC | psbD |
| | | psbE | psbE |
| | | psbF | psbH |
| | | psbI | psbJ |
| | | psbK | psbK |
| | | psbL | psbM |
| | | psbN | psbT |
| | | psbZ | psbZ |
| | | petA | petB" |
| | | petD" | petG |
| | | petI | petL |
| | | petN | petN |
| | | atpA | atpB |
| | | atpE | atpF" |
| | | atpH | atpH |
| | | atpI | atpI |
| | | rbcL | rbcL |
| | | rbcL | rbcL |
| | | ndhA" | ndhB" |
| | | ndhC | ndhD |
| | | ndhE | ndhE |
| | | ndf | ndg |
| | | ndhF | ndhH |
| | | ndhI | ndhI |
| | | ndhK | ndhK |
| | | clpP" | clpP" |
| Chloroplast envelope membrane protein | cemA | cemA |
| Ribosomal proteins | Small subunit of ribosome | rps2 | rps3 |
| | | rps4 | rps7" |
| | | rps8 | rps8 |
| | | rps11 | rps12" |
| | | rps14 | rps15 |
| | | rps16" | rps16" |
| | | rps18 | rps19 |
| Transcription | Large subunit of ribosome | rpl2" | rpl14 |
| | | rpl16" | rpl20 |
| | | rpl22 | rpl22 |
| | | rpl23 | rpl32 |
| | | rpl33 | rpl36 |
| | DNA-dependent RNA polymerase | rpoA | rpoB |
| | | rpoC1" | rpoC2 |
| Other genes | Maturase | matK | matK |
| | Subunit of acetyl-CoA | accD | accD |
| | C-type cytochrome synthesis gene | ccsA | ccsA |
| | Component of TIC complex | ycf1" | ycf1" |
| | Hypothetical proteins | ycf12" | ycf15" |

©2017. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
reverse repeats, 35 palindromic repeats, and 17 complementary repeats (Fig 5A). Of these repeats, 95 (90.5%) were 30–39 bp long, 8 (7.6%) were 40–49 bp long, and 2 (1.9%) were 50–59 bp long. The longest repeat had a length of 56 bp. Simple sequence repeats (SSRs) play significant roles during genome rearrangement and recombination [42]. A total of 991 SSRs were detected in the *A. stellari* chloroplast genome (Fig 5B). Of these, 451 (45%) were mono-nucleotide repeats, 69 (7%) di-nucleotide repeats, 60 (6%) tri-nucleotide repeats, 84 (8%) tetra-nucleotide repeats, 108 (11%) penta-nucleotide repeats, 146 (15%) hexa-nucleotide repeats, and 35, 18, 16 and 4 were 7-, 8-, 9- and 10- nucleotide repeats respectively. Of the 991 SSRs, 60% (594), 21% (208), and 19% (189) SSRs were present in the LSC, IR, and SSC regions, respectively (Fig 5C). In addition, we determined number of repeats in protein-coding and intron and intergenic regions (IGS) (Fig 5D), and found 570 (58%), 329 (33%), and 92 (9%) SSRs were located in IGS, protein-coding, and intron regions, respectively. The presence of repeat sequences in the chloroplast genome of *A. stellari* may be useful for developing lineage-specific markers for genetic diversity and evolutionary studies.

![Sequence alignment of six chloroplast genomes in the Brassicaceae family](https://doi.org/10.1371/journal.pone.0183197.g002)
Pseudogenization of rps16 gene

In photosynthetic plants, chloroplast gene loss infrequently occurs, but only when nuclear and/or mitochondrial genomes encode another functional copy or acquire one from the plastome through gene transfer [43]. Although the number of genes and their order are generally conserved among angiosperm chloroplast genomes [44]. Besides, rare cases have been observed in the chloroplast genomes of Brassicaceae family [6]. Hence, the cp genome size, % GC content and total number of unique protein-coding genes, tRNA and rRNA genes of 14 Brassicaceae family genomes were compared for analysis of gene duplication, pseudogene or gene deletion in its closely related species of Arabis chloroplast genome S2 Table. However, some dissimilarity was identified in protein-coding genes of Brassicaceae. The cp genomes of Arabis genus, D. nemorosa, Arabidopsis arenicola, A. arenosa and A.cebennensis were found to encode 79 protein-coding genes, whereas Brassica genus and A. thaliana possessed 80 protein-coding genes (Fig 6). This one gene variation was caused by either pseudogenization of rps16 in the LSC region of the Arabis or, pseudogenization of ycf15 in A. arenicola, A. arenosa and A. cebennensis cp genomes.

The rps16 gene is critical for cell viability [45] and is involved in the assembly of the 30S subunit [46] in Escherichia coli. In order to analyze pseudogenization of the rps16 gene, we designed a primer and amplified the rps16 gene of A. stellari (S1 Fig). The gene sequence of rps16 confirmed that the A. stellari chloroplast genome encoded a pseudogene rps16. In addition, the rps16 gene was analyzed and compared with Brassicaceae family chloroplast genomes. Among, 14 Brassicaceae, the rps16 gene was found to be a pseudogene in A. stellari, A. hirsuta, and D. nemorosa but to be entirely missing in A. alpina (S2 Fig). The intact nucleotide sequence of rps16 is ~1,161 bp long which includes two exons (~45-bp—exon I and ~226-bp—exon II) and one intron sequence (~890-bp). In the chloroplast genomes of A. stellari, A. hirsuta, and D. nemorosa, 10-bp deletion within the first exon of rps16, leading to a framshift (S2 Fig). Although, deletion of 9-bp found in the second exon of rps16 of A. stellari, A. hirsuta, and D. nemorosa. Whereas, the rps16 gene of A. alpina encoded 21-bp only and it lost the entire second exon and part of the intron sequences. Interestingly, the expression of rps16 gene analyzed in A. thaliana cp genome and identified that the cp rps16 is a pseudogene in this species.
due to the splicing of the group II intron is defective [10]. Whereas, its closely related species *A. arenosa*, *A. lyrata* and *Crucihimalaya lasicarpa* were compared and detected that *rps16* is a functional gene in these species. These results suggested that the pseudogenization event must have occurred after the divergence of *Arabidopsis* and its close relatives of Brassicaceae.

In addition, evolution of the *rps16* gene of *A. stellari* accessed by comparing it with 13 other Brassicaceae chloroplast genomes. Phylogenetic analysis showed intron loss of *rps16* in different genus formed one clade and complete gene loss of *Arabidopsis alpina* formed another clade with *Arabidopsis* genus, suggesting independent evolutionary lineages occurred in Brassicaceae family (Fig 7A). In contrast, another phylogenetic tree was constructed without *Arabis alpina*, and pseudogene *rps16* of *A. stellari*, *A. hirsuta*, and *D. nemorosa* were observed to form one clade and remaining species containing intact *rps16* gene to form another clade (Fig 7B). However, Roy et al. [44] studied evolution of the *rps16[4] gene in the *Arabidopsis* and its closely related species, and commented phylogenetic tree construction with only one gene is unreliable and can misrepresent phylogenetic relationships, since a pseudogene does not always reflect the
phylogenetic position of species. Therefore, it is possible gene or intron loss of *rps16* might have occurred independently in each species rather than by dependent evolution, which is supported by reports of independent *rps16* loss in *Medicago truncatula* [3], *Phaseolus vulgaris* [6], *Cicer arietinum* [47], *Vigna radiata* [48], and *Populus* genus [49, 50].

Additionally, we investigated the presence of *infA* protein-coding gene in Brassicaceae. The plastome gene, *infA* was completely absent in Brassicaceae family, which might have acquired a copy of the *infA* gene from either nuclear or mitochondrial genomes. Earlier studies also suggest that the gene *infA* have been lost in the Brassicales, Cucurbitales, Fabales, Fagales, Malphighiales, Malvales, Myrtales, Rosales, Sapindales, Solanales, *Dianthus* and *Lychinis* [6, 8–12].

**Evolution of the ycf15 gene**

The plastome gene, *ycf15* encodes an ATG start codon in all species of Brassicaceae, suggesting it is probably a functional gene in this family. The genuses *Arabis, Draba, Capsella* and *Brassica* encode two intact copies of the 234-bp*ycf15* gene in their plastomes. *Pugionium* genus encoded
only 162-bp for the yfc15 gene, which may have been due to a point mutation (GAA to TAA) at the 160-bp position. Interestingly, in Arabidopsis genus, only A. thaliana encoded an intact ycf15 gene, whereas other species, such as, A. arenicola, A. arenosa and A. cebennensis encoded multiple internal stop codons, suggesting ycf15 is disabled in these species (S3 Fig). However, comparative analysis suggested the organelle-encoded gene differs within the genus Arabidopsis. Nevertheless, the pseudogene, ycf15 in these species might be transferred to the nucleus. Previous studies have also reported that internal stop codons in the ycf15 gene of many angiosperms [51] and suggested that gene transfer from plastid to nucleus occurred more frequently.
during plastid evolution [52–54]. We also studied evolution of the ycf15 gene in Brassicaceae (Fig 8). The evolutionary patterns of ycf15 showed that it evolved independently in Brassicaceae species. Also, it contained an intact, an internal stop codon, or completely disabled or absent in the Brassicaceae phylogeny. Although, the same results were obtained when evolution of the ycf15 gene was investigated in an angiosperm phylogenetic study [51].

**Synonymous (K_S) and nonsynonymous (K_A) substitution rate analysis**

Synonymous and nonsynonymous nucleotide substitution patterns are more important indicators in gene evolution studies [55]. Although nonsynonymous substitutions occur much less
frequently than synonymous substitutions, K$_A$/K$_S$ ratios are less than one in the majority of protein-coding genes [56]. In the present study, synonymous and nonsynonymous substitution rates were analyzed for 78 protein-coding genes of *A. stellari*, *A. alpina*, and *A. hirsuta* chloroplast genomes (Fig 9). The K$_A$/K$_S$ ratio of all genes was less than 1, except for *ndhA* of *A. hirsuta*. The K$_A$/K$_S$ ratio of *ndhA* of *A. stellari* vs. *A. hirsuta* was 1.35135. This deviation from unity was due to a four-amino acid change by nonsynonymous substitution and the deletion of five amino acids in the second exon of the *ndhA* gene of *A. stellari* due to silent mutation. Although, *ndhA* nucleotide identity was 98.2% vs. *A. hirsuta*. Although, the plastid genes, *atpH*, *petB*, *petG*, *petL*, *petN*, *psaB*, *psaI*, *psbE*, *psbF*, *psbH*, *psbI*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *rbcL*, *rpl23*, *rpl36*, *rps7*, *rps14*, *rps19*, *ycf3* and *ycf15* showed no synonymous or nonsynonymous changes occurred in the cp genomes of *A. stellari*, *A. alpina*, and *A. hirsuta*.

**Phylogenetic analysis of *A. stellari***

To study the phylogenetic position of *A. stellari* within the Brassicaceae family, we used 76 protein-coding genes shared by the chloroplast genomes of 20 rosids and *Vitis* using the *Liquidambar* set as outgroups. Phylogenetic analysis revealed that Brassicaceae family formed a monophyletic group (Fig 10). *A. stellari* clustered with *A. hirsuta* with a bootstrap value of 100%, and *A. stellari* and *A. hirsuta* formed a sister clade with *D. nemorosa* rather than with *A. alpina*. Ten species of the Brassicaceae family showed extremely conserved chloroplast genome structures and their phylogenetic positions remained unaltered.

Overall, in the present study, we have compared the pseudogenization of *rps16*, *ycf15* and *infA* genes of Brassicaceae family. Fig 10 showed that the pseudogenization of *rps16* occurred only in *Arabis* genus whereas *ycf15* gene lost has not occurred in the entire genus of *Arabidopsis*. It occurred only in the species of *A. arenicola*, *A. arenosa* and *A. cebennensis*. While, the *infA* gene has lost in the entire Brassicals, Malvales, Sapindales and Myrtales. Based on these analysis, it suggested that the pseudogenization or gene lost event must have occurred in the species of *A. arenicola*, *A. arenosa* and *A. cebennensis* and Brassicals, Malvales, Sapindales and Myrtales after the earliest divergence lineage of the rosids.
Conclusions

The chloroplast genome *Arabis stellari* was sequenced, analyzed, and compared with closely related species. Its total genome was found to be 153,683 bp long with a GC content of 36.4%. Overall gene contents were similar and gene arrangements were found to be highly conserved in the Brassicaceae family. Minor divergences were observed in the protein-coding genes *matK*, *ycf1*, *ccsA*, *accD*, and *rpl22* and a total of 991 SSRs were also detected in the *A. stellari* plastome genome. The Kᵈ/Kᵈ nucleotide substitution ratio of *ndhA* gene of *A. stellari* vs. *A. hirsuta* was 1.35135. Furthermore, the genes *infA* and *rps16* were completed deleted but the *ycf15* gene was retained in the *Arabis* genus, and phylogenetic evolutionary studies revealed these genes evolved independently. In addition, phylogenetic analysis showed that the Brassicaceae species are extremely highly conserved based on their phylogenetic positions. It is hoped this study will be found useful by those involved in *Arabis* species conservation and molecular phylogenetic studies of Brassicaceae.

Supporting information

S1 Fig. PCR amplification of the *rps16* gene of *Arabis stellari* var. *japonica*. (TIF)

S2 Fig. Comparisons of the *rps16* genes of Brassicaceae family. (TIF)

S3 Fig. Comparisons of the *ycf15* genes of Brassicaceae family. (TIF)

S1 Table. Accession numbers of the chloroplast genome sequences used in this study. (DOCX)
S2 Table. Comparison of cp genome size, %GC content and total number plastid genes of Brassicaceae family.

(DOCX)

Acknowledgments
This work was supported by a National Institute of Biological Resources of Korea (NBR 201631201).

Author Contributions
Conceptualization: Gurusamy Raman, SeonJoo Park.
Data curation: Gurusamy Raman.
Formal analysis: Gurusamy Raman.
Funding acquisition: SeonJoo Park.
Investigation: Gurusamy Raman, SeonJoo Park.
Methodology: Gurusamy Raman.
Resources: Myounghai Kwak, Byoungyoon Lee.
Supervision: SeonJoo Park.
Validation: SeonJoo Park.
Writing – original draft: Gurusamy Raman.
Writing – review & editing: Gurusamy Raman, Veronica Park.

References
1. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Rev Genet. 2004; 5:123–135. https://doi.org/10.1038/nrg1271 PMID: 14735123
2. Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber AP, et al. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. Science 2012; 335:843–847. https://doi.org/10.1126/science.1213561 PMID: 22344442
3. Saski C, Lee S, Fjellheim S, Guda C, Jansen RK, Luo H, et al. Complete chloroplast genome sequences of Hordeum vulgare, Sorghum bicolor and Agrostis stolonifera, and comparative analyses with other grass genomes. Theor Appl Genet. 2007; 115:571–590 https://doi.org/10.1007/s00122-007-0567-4 PMID: 17534593
4. Palmer JD. Comparative organization of chloroplast genomes. Annu Rev Genet. 1985; 19:325–54. https://doi.org/10.1146/annurev.ge.19.120185.001545 PMID: 3936406
5. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17:134. https://doi.org/10.1186/s13059-016-1004-2 PMID: 2733912
6. Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, et al Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics 2017; 18:176 https://doi.org/10.1186/s12864-017-3555-3 PMID: 28209119
7. Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol Phylogenet Evol. 2014; 72:82–89. https://doi.org/10.1016/j.ympev.2013.12.004 PMID: 2437909
8. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986; 5:2043–2049. PMID: 16453699
9. Wolfe KH, Morden CW, Ems SC, Palmer JD. Rapid evolution of the plastid translational apparatus in a non-photosynthetic plant: Loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J Mol Evol. 1992; 35:304–317. PMID: 1404416

10. Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res. 1999; 165:283–290.

11. Hupfer F, Swiatek M, Hornung S, Hermann RG, Maier RM, Chiu WL, et al. Complete nucleotide sequence of the Oenothera elata plastid chromosome, representing plastome I of the five distinguishable Euonothera plastomes. Mol Gen Genet. 2000; 165:581–585.

12. Raman G, Park S. Analysis of the Complete Chloroplast Genome of a Medicinal Plant, Dianthus superbus var. longicalycinus, from a Comparative Genomics Perspective. PLoS ONE 2015; 10(10): e0141329 https://doi.org/10.1371/journal.pone.0141329 PMID: 26513163

13. Jansen RK, Sasaki C, Lee SB, Hansen AK, Daniell H. Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol. 2011; 28:835–47. https://doi.org/10.1093/molbev/msq261 PMID: 20935065

14. Thomas F, Massenet O, Dorne AM, Mache R. Expression of the rpi23, rpl2, and rps19 genes in spinach chloroplasts. Nucleic Acids Res. 1988; 16:2461–2472. PMID: 33627671

15. Ueda M, Fujimoto M, Arimura SI, Murata J, Tsutsumi N, Kadowaki KI. Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene. 2007; 402:51–6 https://doi.org/10.1016/j.gene.2007.07.019 PMID: 17728076

16. Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, et al. The complete sequence of the rice (Oryza sativa) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet. 1989; 217:185–194. PMID: 2770692

17. Maier RM, Neckermann K, Igloi GL, Kössel H. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol. 1995; 165:614–628.

18. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J. 1991; 165:3073–3078.

19. Nagano Y, Matsuno R, Sasaki Y. Sequence and transcriptional analysis of the gene cluster trnQ-zfpA-psal-ORF231-petA in pea chloroplasts. Curr Genet. 1991; 165:431–436.

20. Downie SR, Palmer JD. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular systematics of plants. New York: Chapman and Hall. 1992;14–35.

21. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS ONE. 2011; 6(5):E19254. https://doi.org/10.1371/journal.pone.0019254 PMID: 21637336

22. Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA. Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc Natl Acad Sci. 1995; 92:7759–7763. PMID: 7644491

23. Bock R, Khan MS. Taming plastids for a green future. Trends Biotechnol. 2004; 22:311–318. https://doi.org/10.1016/j.tibtech.2004.03.005 PMID: 15158061

24. Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. Focus 1990; 12:13–15.

25. Wyman SK, Boore JL, Jansen RK. Automatic annotation of organelar genomes with DOGMA. Bioinformatics 2004; 20:3252–3255. https://doi.org/10.1093/bioinformatics/bth352 PMID: 15180927

26. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snocan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33:W686–W689. https://doi.org/10.1093/nar/gki366 PMID: 15980563

27. Lohse M, Drechsel O, Bock R. Organelar genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. 2009; 25:1451–1452.

28. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004; 32:W273–W279. https://doi.org/10.1093/nar/gkh458 PMID: 15215394

29. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29:4633–4642 PMID: 11713313

30. Mayer C, Leese F, Tollian R. Genome-wide analysis of tandem repeats in Daphnia pulex–a comparative approach. BMC Genomics 2010; 11:277 https://doi.org/10.1186/1471-2164-11-277 PMID: 20433735
31. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25:1451–1452. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325
32. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Research 2012; 40(15):e115. https://doi.org/10.1093/nar/gks596 PMID: 22730293
33. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. https://doi.org/10.1093/molbev/mst101 PMID: 23329690
34. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404 PMID: 17846036
35. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–2690. https://doi.org/10.1093/bioinformatics/btl446 PMID: 16928733
36. Melodelima C, Lobréaús S. Complete Arabis alpina chloroplast genome sequence and insight into its polymorphism. Meta Gene 2013; 1:65–75. https://doi.org/10.1016/j.mgene.2013.10.004 PMID: 25606376
37. Hu Z, Hua W, Shunmou H, Wang H. Complete chloroplast genome sequence of rapeseed (Brassica napus L.) and its evolutionary implications. Genet Resour Crop Ev. 2011; 58(6): 875–887.
38. Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res. 1999; 29:283–290.
39. Curci PL, De Paola D, Danzi D, Vendramin GG, Sonnante G. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. PLoS ONE. 2015; 10(3): e0120589 https://doi.org/10.1371/journal.pone.0120589 PMID: 25774672
40. Tangpha tsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchaï N, et al. The chloroplast genome sequence of mungbean (Vigna radiata) determined by high-throughput pyrosequencing: Structural organization and phylogenetic relationships. DNA Res. 2010; 17:5554–5560. PMID: 20673851
41. Held WA, Nomura M. Escherichia coli 30 S ribosomal proteins uniquely required for assembly. J Biol Chem. 1975; 250:3179–3184. PMID: 804486
42. Jansen RK, Wojciechowski MF, Sanniayi E, Lee SB, Daniell H. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 2008; 48:1204–1217. https://doi.org/10.1016/j.ympev.2008.06.013 PMID: 18638561
43. Mrazek J. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol Biol and Evol. 2006; 23(7):1370–1385.
44. Roy S, Ueda M, Kadowaki KI, Tsutsui N. Different status of the gene for ribosomal protein S16 in the chloroplast genome during evolution of the genus Arabidopsis and closely related species. Genes and Genetic Systems 2010; 85:319–326. PMID: 21317544
45. Persson BC, Bylund GO, Berg DE, Wikstrom PM. Functional analysis of the ffh-trmD region of the Escherichia coli chromosome by using reverse genetics. J Bacteriol. 1995; 177:5554–5560. PMID: 7759342
46. Shi C, Liu Y, Huang H, Xia EH, Zhang HB, Gao LZ. Contradiction between plastid gene transcription and function due to complex post transcriptional splicing: an exemplary study of ycf15 function and evolution in angiosperms. PLoS One. 2013; 8:e59620. https://doi.org/10.1371/journal.pone.0059620 PMID: 23527231
52. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, et al. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 2002; 99:12246. https://doi.org/10.1073/pnas.182432999 PMID: 12218172

53. Matsuo M, Ito Y, Yamauchi R, Obokata J. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast—nuclear DNA flux. Plant Cell 2005; 17:665–675. https://doi.org/10.1105/tpc.104.027706 PMID: 15705954

54. Noutsos C, Richly E, Leister D. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res 2005; 15:616–628. https://doi.org/10.1101/gr.3788705 PMID: 15867426

55. Kimura M. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England. 1983.

56. Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci. 1998; 95:9407–9412. PMID: 9689093