The Comparative Analyses of Six Complete Chloroplast Genomes of Morphologically Diverse Chenopodium album L. (Amaranthaceae) Collected in Korea

Jongsun Park, Juhyeon Min, Yongsung Kim, and Youngjae Chung

1InfoBoss Inc., 301 Room, 670, Seolleung-ro, Gangnam-gu, Seoul, Republic of Korea
2InfoBoss Research Center, 301 Room, 670, Seolleung-ro, Gangnam-gu, Seoul, Republic of Korea
3Department of Biological Sciences, Sungkyunkwan University, Suwon, Republic of Korea
4Department of Biology, Shingyeong University, Hwaseong 18274, Republic of Korea

Correspondence should be addressed to Jongsun Park; starflr@infoboss.co.kr and Youngjae Chung; chenopodium@hanmail.net

Received 22 February 2021; Accepted 9 April 2021; Published 27 April 2021

Abstract

Chenopodium album sensu stricto belonging to C. album aggregate is an annual cosmopolitan weed displaying the diversity of morphologies. We completed the six chloroplast genomes of C. album s. str. collected in Korea to understand the relationship between the diversity of chloroplast genomes and their morphological variations. All six C. album chloroplast genomes have a typical quadripartite structure with length ranging from 151,906 bp to 152,199 bp, similar to the previously sequenced C. album chloroplast genome (NC_034950). In total, 56 single nucleotide polymorphisms (SNPs) and 26 insertion and deletion (INDEL) regions (308 bp in total) were identified from the six chloroplast genomes, presenting a low level of intraspecific variations in comparison to the other angiosperm species. 376 normal simple sequence repeats were identified in all seven C. album chloroplast genomes. The phylogenetic analysis based on all available complete Amaranthaceae chloroplast genomes presents phylogenetic positions of six C. album samples as well as correlation with one of C. album morphological features. Our results provide the way to investigate intraspecific features of C. album chloroplast genomes and also the insights of understanding various intraspecific characteristics including morphological features.

1. Introduction

Chenopodium album aggregate, also known as Chenopodium album sensu lato, is one of the challenging groups for delimitating species boundary by phenotypes [1]. The main possibility of its phenotypic diversity is hybridization and polyploidization [2] because this group shows a diploid-polyploid complex ranging from diploids to decaploids with individual species which are very tightly correlated with a specific ploidy level and genome size [3–5]. Genetic diversity of C. album aggregate has been analyzed displaying that at least eight groups were identified, and their evolutionary lineages were supported well using flowering locus T-like (FTL) intron and chloroplast DNA maker sequences [2]. In addition, Central Asia presents the most diverse haplotypes, indicating that this area can be an origin of C. album aggregate [2].

C. album sensu stricto belonging to C. album aggregate is an annual cosmopolitan weed native to Eurasia [6]. It is one of the notorious weeds which reduce crop yield by exploiting resources such as light and nutrients from soil [6]. In contrast, C. album s. str. has also been cultivated as a crop in some countries: it was especially considered better nutritional crop species than wheat, barley, maize, and rice in the Himalaya area [7–9]. Besides it, C. album s. str. was also cultivated as a leafy vegetable [9, 10].

C. album s. str. has been recognized as a morphologically diverse species with difficulty in species identification [1, 11, 12]. This hexaploid species exhibited wide phenotypic plasticity covering morphological variations of other Chenopo-
*diurn* species, which shows stable morphologies under the greenhouse environment [1]. This high-level phenotypic variation of *C. album* s. str. can be explained by its allopolyploid originated from maternal tetraploids and paternal diploid parents [13]. Even though a few studies to understand the origin of *C. album* s. str. in the evolutionary aspect have been conducted [13, 14], there is no research to unravel the relationship between its morphological plasticity and the genetic diversity of organelle genomes.

Due to low sequencing cost caused by rapid development of next-generation sequencing (NGS) and third-generation sequencing (TGS) technologies [15–17] as well as the conserved structure of chloroplast genome reducing difficulties of *de novo* assembly, a huge number of chloroplast genomes have been massively sequenced. One of the evidences which reflects this situation is that the number of angiosperm complete chloroplast genomes deposited in NCBI is 9,323 covering 6,727 species as of 11 Jan 2021. These chloroplast genome sequences have been analyzed for the identification of phylogenetic positions [18–20] and for developing useful molecular markers [21, 22]. It presents that sequence variations of complete chloroplast genomes are enough information to conduct the studies for those purposes that complete chloroplast genomes can be also used to reveal the relationship between morphological plasticity and genetic diversity of *C. album* s. str.

Here, we completed the six chloroplast genomes of *C. album* s. str. of which both morphological features and geographical locations in the Korean Peninsula are distinct from each other. The *C. album* chloroplast genomes display a relatively low level of intraspecific variations (56 single nucleotide polymorphisms (SNPs) and 26 insertion and deletion (INDEL) regions across the seven *C. album* chloroplast genomes) with low nucleotide diversity compared to those of other angiosperm chloroplast genomes. 376 normal simple sequence repeats (SSRs), including two individual-specific normal SSRs, are detected among the seven chloroplast genomes, and the numbers of SSRs and their distribution on each *C. album* chloroplast genome are similar. Maximum likelihood and Bayesian inference phylogenetic trees based on each chloroplast genome are similar. The six chloroplast genomes ranges from 151,906 bp (CCANG01) to 152,199 bp (CAGAP004; Table 2), presenting 293 bp differences. They are similar to that of one of the two previously sequenced *C. album* chloroplast genomes, of which GenBank accession is NC_034950 (152,167 bp) [23]. Interestingly, the other chloroplast genome of *C. album*, MF418659, displays much shorter (150,272 bp) than those of the remaining *C. album* chloroplast genomes [24]. Their overall GC contents are conserved as between 37.2% and 37.3%, which are slightly higher than that of MF418659 (37.0%; Table 2). The GC contents of LSC and SSC regions are identical in the seven *C. album* chloroplast genomes including the firstly sequenced chloroplast of *C. album* (NC_034950), which are 35.3% and 31.0%, respectively. Similarly, GC contents in the IR region of the seven chloroplast genomes are from 42.7% to 42.8% (that of MF418659 is 43.4%).

All seven *C. album* chloroplast genomes including NC_034950 contain 129 genes including 84 protein-coding genes (PCGs), 8 ribosomal RNAs (rRNAs), and 37 transfer RNAs (tRNAs; Table 3). Seventeen genes are duplicated in IR regions including 6 PCGs (rps12, rpl23, rpl2, ycf1, ycf68, and ycf5); 4 rRNAs (rrn4.5, rrn5.8, rrn16, and rrn21); and 7 tRNAs (trnL-CAA, trnM-UAC, trnG-GTG, trnC-CAU, trnT-UGC, trnK-ACC, and trnN-GUU). The number of PCGs from the seven *C. album* chloroplast genomes is the same except MF418659 (Table 3), presenting that MF418659 has a quite different gene configuration. It has six additional PCGs, rpl23, ycf15, and ycf68 in the IR region, and loses psal PCG. In addition, it also misses one of the two rps12, displaying the same sequences of three exons of rps12 to the rest seven *C. album* chloroplast genomes; however, this missed rps12 should be added in the annotation of MF418659, resulting in 89 PCGs. After reannotation of two Chenopodium quinoa chloroplast genomes (KY635884 and MF805727), one ycf1 and four additional PCGs, ycf1, ycf2, and two rps12, are added, respectively, resulting in all chloroplast genomes of Chenopodium species having the same number of genes except MF418659 (Table 3).

In the seven *C. album* chloroplast genomes, there are six ATP synthase subunit genes, 11 NADH dehydrogenase genes, four RNA polymerase genes, six genes encoding subunits of cytochrome b/f complex, and 15 photosystem

2. Results

### 2.1. Six Complete Chloroplast Genomes and Their Morphological Features of Chenopodium album Collected in Korea

Six samples of *C. album* collected in Korea named CAGAP004, CAGOH01, CAGEJG05, CCANG01, CSJUK01, and CVHUP01 were selected based on their distinct morphological features and collected in different geographical positions in Korean Peninsula (Table 1 and Figure 1). The leaf shape of the six *C. album* samples presents a wide range which is from lanceolate to ovate. Their leaf margin shows the two types: (i) serrate (CAGAP004, CAGOH01, and CCANG01) and (ii) entire (CAGEJG05, CSJUK01, and CVHUP01; Figure 1 and Table 1). In addition, the thickness of leaves of the six samples is also divided into two types: (i) thick type (CAGAP004, CAGEJG05, and CVHUP01) and (ii) thin type (CAGOH01, CCANG01, and CSJUK01).

To understand the relationship between morphological features with geographical distribution and its genetic background of chloroplast genome, we completed chloroplast genomes of the six *C. album* samples (Table 1). The six chloroplast genomes have a typical quadripartite structure which has one large single-copy (LSC), one small single-copy, and two inverted repeat (IR) regions (Figure 2). The length of the six chloroplast genomes ranges from 151,906 bp (CCANG01) to 152,199 bp (CAGAP004; Table 2), presenting 293 bp differences. They are similar to that of one of the two previously sequenced *C. album* chloroplast genomes, of which GenBank accession is NC_034950 (152,167 bp) [23]. Interestingly, the other chloroplast genome of *C. album*, MF418659, displays much shorter (150,272 bp) than those of the remaining *C. album* chloroplast genomes [24]. Their overall GC contents are conserved as between 37.2% and 37.3%, which are slightly higher than that of MF418659 (37.0%; Table 2). The GC contents of LSC and SSC regions are identical in the seven *C. album* chloroplast genomes including the firstly sequenced chloroplast of *C. album* (NC_034950), which are 35.3% and 31.0%, respectively. Similarly, GC contents in the IR region of the seven chloroplast genomes are from 42.7% to 42.8% (that of MF418659 is 43.4%).

All seven *C. album* chloroplast genomes including NC_034950 contain 129 genes including 84 protein-coding genes (PCGs), 8 ribosomal RNAs (rRNAs), and 37 transfer RNAs (tRNAs; Table 3). Seventeen genes are duplicated in IR regions including 6 PCGs (rpl23, ycf1, ndhB, rps12, rps7, and ycf1); 4 rRNAs (rrn16, rrn23, rrm4.5, and rrn5); and 7 tRNAs (trnL-CAA, trnM-UAC, trnG-GTG, trnC-CAU, trnT-UGC, trnK-ACC, and trnN-GUU). The number of PCGs from the seven *C. album* chloroplast genomes is the same except MF418659 (Table 3), presenting that MF418659 has a quite different gene configuration. It has six additional PCGs, rpl23, ycf15, and ycf68 in the IR region, and loses psal PCG. In addition, it also misses one of the two rps12, displaying the same sequences of three exons of rps12 to the rest seven *C. album* chloroplast genomes; however, this missed rps12 should be added in the annotation of MF418659, resulting in 89 PCGs. After reannotation of two Chenopodium quinoa chloroplast genomes (KY635884 and MF805727), one ycf1 and four additional PCGs, ycf1, ycf2, and two rps12, are added, respectively, resulting in all chloroplast genomes of Chenopodium species having the same number of genes except MF418659 (Table 3).

In the seven *C. album* chloroplast genomes, there are six ATP synthase subunit genes, 11 NADH dehydrogenase genes, four RNA polymerase genes, six genes encoding subunits of cytochrome b/f complex, and 15 photosystem
subunit II genes. Five genes encoding photosystem subunit I are found in the seven *C. album* chloroplast genomes (Table 4), while MF418659 has only four genes with losing *psaJ*. There are 20 ribosomal proteins presented in the seven *C. album* chloroplast genomes consisting of 12 ribosomal proteins encoding small subunit and eight genes for large subunits (Table 4). Interestingly, an additional *rpl23* gene is found only in the MF418659 chloroplast genome. Other remaining genes encode acetyl-CoA-carboxylase (*accD*), translation initiation factor (*infA*), protease (*clpP*), chloroplast envelope membrane protein (*cemA*), matrase K gene (*matK*), and cytochrome c biogenesis protein (*ccsA*; Table 4). The number of hypothetical proteins is four among the seven *C. album* chloroplast genomes, except for MF418659 which has six genes (Table 4), presenting the different gene configuration of MF418659.

In the seven *C. album* chloroplast genomes, nine PCGs contain one intron (*rps16, atpF, rpoC1, petB, petD, rpl16, rps12, ndhB, and ndhA*) and only *clpP* and *ycf3* have two introns, which are conserved across chloroplast genomes of the other *Chenopodium* species. MF418659 chloroplast also has the same intron structure as the remaining *Chenopodium* chloroplast genomes. Taken together with the different properties of the MF418659 chloroplast genome including its length, GC ratio, and a number of genes, we suspected that MF418659 chloroplast genome might not be *C. album*; hence, we will exclude MF418659 for further analyses conducted in this study.

### 2.2. Nucleotide Diversity and Intraspecific Variations Identified from the Seven *Chenopodium album* Chloroplast Genomes

To investigate nucleotide diversity (\(\pi\)) and intraspecific variations of *C. album* chloroplast genomes, the six *C. album* complete chloroplast genomes sequenced in this study are aligned against the previously sequenced *C. album* chloroplast genome (NC_034950). The average value of nucleotide diversity is 0.0000625 (Figure 3), and a total of 56 single nucleotide polymorphisms (SNPs) and 26 insertion and deletion (INDEL) regions (308 bp in total) are identified. The LSC region, where the average nucleotide diversity is the highest (\(\pi = 0.00102\)), contains 35 SNPs (62.5%) and 16 INDEL regions (47 bp in length; 59.3%). Fifteen SNPs
(26.8%) and 5 INDEL regions (193 bp in length; 8.93%) are found in the SSC region, displaying that the number of SNPs in the LSC region is larger more than twice that of the SSC region; however, the total length of INDEL regions in the SSC region is about 4 times greater than that of the LSC region. The main reason for this phenomenon is the presence of the 162 bp INDEL located between \textit{rpl32} and \textit{trnL-UAG} genes. An IR region covers three SNPs and three INDEL regions (68 bp in length), which corresponds to the lowest nucleotide diversity in the IR region ($\pi = 0.0000146$). The low level of sequence variations in the IR region is known as a general phenomenon in the chloroplast genomes [25–27].
26 INDEL regions (26.9%) are identified in the intronic regions of psbA, matK, ycf1, rpoC1, and atpB) contain only one SNP in their coding region, while four genes (matK, ccsA, ycf1, and ndhF) cover more than one SNP. In addition, three PCGs (clpP, petB, and ndhA) have one SNP in their intronic region, while rpl16 has two SNPs in their intronic region. In total, 15 nonsynonymous SNPs and 11 synonymous SNPs are identified in the 14 PCGs. Remarkably, a ratio of nonsynonymous to synonymous SNPs found on C. album chloroplast genomes is 15:11, which is different from the common phenomenon that the number of nonsynonymous SNPs is smaller than that of synonymous SNPs on the other plant chloroplast genomes [26, 28]. Each synonymous SNP is scattered in the ten genes (psbA, rps16, rpoC1, atpB, petA, clpP, rpoC1, rpl16, ndhF, ccsA, and ndhD), while 15 nonsynonymous SNPs are concentrated only in the five genes (matK, infA, ycf1, ndhF, and ccsA). Ycf1, which is duplicated in the IR region and extends between IR and SSC regions, contains three nonsynonymous SNPs in each pair of the IR region, and one additional nonsynonymous SNP in the SSC region. The PCG containing the largest SNPs is ndhF, which contains five SNPs and is located mainly in the SSC region and partially extended over the IR region. Four out of the five SNPs identified in ndhF are nonsynonymous, among which one SNP found in the IR region is commonly shared with ycf1. Three PCGs (psbA, matK, and ycf1) have one INDEL region (11.5%) of which lengths are 6 bp, 6 bp, and 27 bp, respectively: none of these INDEL regions cause any frameshift mutation. Seven INDEL regions (26.9%) are identified in the intronic regions of rps16, atpF, ycf3, clpP, and rpl16. Most of them are less than 5 bp long; however, 66 bp INDEL is found in the intron of trnL-GAU in each IR region. The remaining 16 INDEL regions (61.5%) are intergenic. The longest INDEL region is 162 bp long, detected between rpl32 and trnL-UAG.

There are eight sites that have relatively higher π values (>0.0008) including two PCGs matK (π = 0.00114) and ccsA (π = 0.00163) and six intergenic regions (trnH-psbA, petN-psbM, rpl11-rpl36, rpl36-infA, ycf1-ndhF, and rpl32-trnL; Figure 3). CcsA contains two nonsynonymous SNPs and one synonymous SNP, and matK has two nonsynonymous SNPs and one 6 bp INDEL, presenting that ccsA displays the highest SNP density among the PCGs. The highest π value of the intergenic region is observed between trnH and psbA (π = 0.00128; Figure 3).
and CSJK01 chloroplast genomes have additional monoSSR between ndhC and rbcL by insertion of 1 bp nucleotide "T." In contrast, the deletion of a single nucleotide of "A" between ndhF and rpl32 caused the removal of intergenic SSRs in CAGOH01, CCANG01, and CSJK01 chloroplast genomes. All seven C. album chloroplast genomes have seven common normal SSRs located in the intronic regions of five PCGs, rps16, atpF, ycf3, rpl16, and ndhA. In the case of CAGOH01 and CCANG01 chloroplast genomes, one additional monoSSR is identified in the intronic region of trnK because one SNP changing "A" to "T" occurred in both chloroplast genomes. Besides, CSJK01 has an extra unique

<table>
<thead>
<tr>
<th>Family</th>
<th>Species name</th>
<th>NCBI accession</th>
<th>Total length (bp)</th>
<th># of PCGs</th>
<th># of tRNAs</th>
<th># of rRNAs</th>
<th>GC ratio (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chenopodioidae</td>
<td>Chenopodium album</td>
<td>MW446241</td>
<td>151,906</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.3%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>MW446242</td>
<td>152,197</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.2%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>MW446243</td>
<td>152,183</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.2%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>MW446244</td>
<td>152,190</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.2%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>MW446245</td>
<td>152,196</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.3%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>MW446246</td>
<td>152,199</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.2%</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album</td>
<td>NC_034950</td>
<td>152,167</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.2%</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium album *</td>
<td>MF418659</td>
<td>150,272</td>
<td>89 (88)</td>
<td>33</td>
<td>8</td>
<td>37.0%</td>
<td>[24]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium ficifolium</td>
<td>NC_041200</td>
<td>151,923</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>37.3%</td>
<td>[45]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>KY635884</td>
<td>152,075</td>
<td>84 (83)</td>
<td>36</td>
<td>8</td>
<td>37.2%</td>
<td>[61]</td>
</tr>
<tr>
<td>Chenopodioidae</td>
<td>Chenopodium quinoa</td>
<td>MF805727</td>
<td>151,069</td>
<td>84 (80)</td>
<td>29</td>
<td>8</td>
<td>37.2%</td>
<td>[62]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>NC_034949</td>
<td>152,099</td>
<td>84</td>
<td>36</td>
<td>8</td>
<td>37.2%</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>NSDK10013185.1</td>
<td>152,282</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>37.2%</td>
<td>[63]</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Atriplex centralasiatica</td>
<td>NC_045304</td>
<td>152,237</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>37.3%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Dysphania ambrosioides</td>
<td>NC_041201</td>
<td>151,689</td>
<td>84</td>
<td>36</td>
<td>8</td>
<td>36.9%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Dysphania botrys</td>
<td>NC_042166</td>
<td>152,055</td>
<td>83</td>
<td>37</td>
<td>8</td>
<td>36.8%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Dysphania pumilio</td>
<td>MK541016</td>
<td>151,960</td>
<td>84</td>
<td>36</td>
<td>8</td>
<td>36.9%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Dysphania pumilio</td>
<td>NC_041159</td>
<td>151,962</td>
<td>84</td>
<td>36</td>
<td>8</td>
<td>36.9%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Oxybasis glauca</td>
<td>NC_047226</td>
<td>151,655</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>36.9%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa</td>
<td>Spinacia oleracea</td>
<td>NC_0022020</td>
<td>150,725</td>
<td>96</td>
<td>37</td>
<td>8</td>
<td>36.8%</td>
</tr>
<tr>
<td></td>
<td>Chenopodium quinoa cultivar Real Blanca</td>
<td>NSDK10013185.1</td>
<td>152,282</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>37.2%</td>
<td>[63]</td>
</tr>
<tr>
<td>Salicornioideae</td>
<td>Salicornia europaea</td>
<td>NC_027225</td>
<td>153,232</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>36.2%</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Salicornioideae</td>
<td>Salicornia bigelovii</td>
<td>NC_027226</td>
<td>153,076</td>
<td>83</td>
<td>37</td>
<td>8</td>
<td>36.3%</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Salicornioideae</td>
<td>Salicornia brachiata</td>
<td>NC_027224</td>
<td>153,324</td>
<td>84</td>
<td>37</td>
<td>8</td>
<td>36.2%</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Suaedoideae</td>
<td>Suaeda japonica</td>
<td>NC_042675</td>
<td>152,109</td>
<td>83</td>
<td>37</td>
<td>8</td>
<td>36.4%</td>
<td>[46]</td>
</tr>
<tr>
<td>Suaedoideae</td>
<td>Suaeda japonica</td>
<td>MK764271</td>
<td>152,112</td>
<td>80</td>
<td>37</td>
<td>8</td>
<td>36.4%</td>
<td>[38]</td>
</tr>
<tr>
<td>Suaedoideae</td>
<td>Suaeda salsa</td>
<td>NC_045302</td>
<td>151,642</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>36.4%</td>
<td>[69]</td>
</tr>
<tr>
<td>Suaedoideae</td>
<td>Suaeda glauca</td>
<td>NC_045303</td>
<td>149,807</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>36.5%</td>
<td>[70]</td>
</tr>
<tr>
<td>Suaedoideae</td>
<td>Suaeda malacopserma</td>
<td>NC_039180</td>
<td>151,989</td>
<td>83</td>
<td>37</td>
<td>8</td>
<td>36.4%</td>
<td>[71]</td>
</tr>
<tr>
<td></td>
<td>Bienertia sinuspersici</td>
<td>KU726550</td>
<td>153,472</td>
<td>86</td>
<td>36</td>
<td>8</td>
<td>37.8%</td>
<td>[72]</td>
</tr>
<tr>
<td>Salsoloideae</td>
<td>Haloxylon ammodendron</td>
<td>NC_027668</td>
<td>151,570</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>36.6%</td>
<td>[73]</td>
</tr>
<tr>
<td>Salsoloideae</td>
<td>Haloxylon persicum</td>
<td>NC_027669</td>
<td>151,586</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>36.6%</td>
<td>[73]</td>
</tr>
<tr>
<td>Betoideae</td>
<td>Beta vulgaris</td>
<td>EF534108</td>
<td>149,696</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>35.4%</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Betoideae</td>
<td>Beta vulgaris subsp. vulgaris</td>
<td>KJ081864</td>
<td>149,635</td>
<td>85</td>
<td>37</td>
<td>8</td>
<td>37.0%</td>
<td>[74]</td>
</tr>
<tr>
<td>Betoideae</td>
<td>Beta vulgaris subsp. vulgaris</td>
<td>KR230391</td>
<td>149,722</td>
<td>81</td>
<td>29</td>
<td>8</td>
<td>37.0%</td>
<td>[75]</td>
</tr>
<tr>
<td>Paronychieae</td>
<td>Gymnocarpos przewalskii</td>
<td>NC_036812</td>
<td>150,636</td>
<td>81</td>
<td>37</td>
<td>8</td>
<td>36.5%</td>
<td>[66]</td>
</tr>
</tbody>
</table>

* indicates that the species name should be reconsidered. ** Numbers in parenthesis are the original number of PCGs based on the annotation, and numbers outside of parenthesis indicate the number of PCGs based on our reannotation results.
tetraSSR "GGTT" in the intronic regions of $ycf3$ by 4 bp insertion. These differences of normal SSRs among the seven chloroplasts of $C. \ album$ can be utilized as molecular markers to distinguish their origins inside the Korean Peninsula once more chloroplast genomes of $C. \ album$ in Korea are available.

Due to the different lengths of the LSC and SSC regions, the density of SSRs per Kbp was calculated. Interestingly, three chloroplast genomes, CAGOH01, CCANG01, and CSJUK01, display similar density in both the LSC and SSC regions, while the remaining four chloroplast genomes present that the density in the SSC region is larger than that of the LSC region. Since in the SSC region of CAGAP004, CAJEJG05, CVHUP01, and NC_034950 there is an additional intergenic normal SSR located in between $rdhF$ and $rpl32$ (Figure 4(c)). The densities of normal SSRs in IR regions of all seven chloroplast genomes are lowest (0.159 to 0.198 normal SSRs/Kbp; Figure 4(c)).

To understand conserved normal SSRs across the seven chloroplast genomes, we calculated SSR groups which contain normal SSRs of which left and right flanking sequences are similar to each other (see Materials and Methods). In total, 58 SSR groups and two singleton SSRs were identified, and 50 of 58 SSR groups (86.2%) contain seven normal SSRs originating from all seven chloroplast genomes, called the common SSR group. Eleven out of the 50 common SSR groups (22.0%) are located in the coding region, and 36 common SSR groups (72.0%) are in the intergenic region (Figure 4(b)), which is congruent to the analysis result of normal SSRs mentioned in the previous section. Five intergenic loci contain two common SSR groups, $rpl32-trnL$, atpH-atpl, $ycf3-trnS$, trnQ-psbK, and trnK-rps16 in each, and 31 intergenic loci contain one common SSR group. Two singleton SSRs were found in CSJUK01 and CCANG01 chloroplast genomes. These intraspecific variations of normal SSRs will provide insights into changes of SSRs inside the species, which can also be utilized to develop molecular markers of $C. \ album$ efficiently.

### Table 4: List of genes encoded by $Chenopodium \ album$ chloroplast genomes.

<table>
<thead>
<tr>
<th>Category</th>
<th>Group of gene</th>
<th>Name of gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosomal RNAs</td>
<td></td>
<td>$rrn4.5$, $rrn5$, $rrn16$, $rrn23$</td>
</tr>
<tr>
<td>Transfer RNA genes</td>
<td></td>
<td>$psaA$, $psaB$, $psaC$, $psaI$, $psaJ$</td>
</tr>
<tr>
<td>Self-replication</td>
<td>Small subunit of ribosome</td>
<td>$rps2$, $rps3$, $rps4$, $rps7$, $rps8$, $rps11$, $rps12$, $rps14$, $rps15$, $rps16$, $rps18$, $rps19$</td>
</tr>
<tr>
<td></td>
<td>Large subunit of ribosome</td>
<td>$rpl2$, $rpl14$, $rpl16$, $rpl20$, $rpl22$, $rpl32$, $rpl33$, $rpl36$</td>
</tr>
<tr>
<td></td>
<td>RNA polymerase</td>
<td>$rpoA$, $rpoB$, $rpoC1$, $rpoC2$</td>
</tr>
<tr>
<td></td>
<td>Translation initiation factor</td>
<td>$infA$</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>ATP synthase</td>
<td>$atpA$, $atpB$, $atpE$, $atpF$, $atpH$, $atpI$</td>
</tr>
<tr>
<td></td>
<td>Cytochrome b/f complex subunit</td>
<td>$petA$, $petB$, $petD$, $petG$, $petL$, $petN$</td>
</tr>
<tr>
<td></td>
<td>Photosystem subunit I subunit</td>
<td>$psaA$, $psaB$, $psaC$, $psaI$, $psaJ$</td>
</tr>
<tr>
<td></td>
<td>Rubisco large subunit</td>
<td>$rbcL$</td>
</tr>
<tr>
<td></td>
<td>Maturase</td>
<td>$matK$</td>
</tr>
<tr>
<td></td>
<td>Protease</td>
<td>$clpP$</td>
</tr>
<tr>
<td></td>
<td>Envelope membrane protein</td>
<td>$cemA$</td>
</tr>
<tr>
<td></td>
<td>Subunit of acetyl-CoA-carboxylase</td>
<td>$accD$</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c-type biogenesis protein</td>
<td>$ccsA$</td>
</tr>
<tr>
<td>Other genes</td>
<td>Hypothetical reading frame</td>
<td>$ycf1$, $ycf2$, $ycf3$, $ycf4$</td>
</tr>
<tr>
<td>Genes of unknown function</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.4. Phylogenetic Analysis of Korean $C. \ album$ Chloroplast Genome Sequence. Bootstrapped maximum likelihood (ML)
and Bayesian inference (BI) phylogenetic trees of 34 Amaranthaceae chloroplast genomes including the six *C. album* chloroplasts sequenced in this study and one outgroup species, *Gymnocarpos przewalskii*, were constructed (see Materials and Methods). Phylogenetic trees present that six *C. album* chloroplast genomes are clustered with the previously sequenced *C. album* chloroplast genome (NC_034950) with high supportive values of ML and BI except the node containing NC_034950 in the ML tree (Figure 5(a)). The *C. album* s. str. clade is divided reciprocally into two clades in both trees. CAGAP004 and CAJEJG05 sharing the morphological feature of narrow leaves and collected in Jejudo island (Figure 1) only exhibit a correlation of geographical locations with high supportive values of ML and BI (Figure 5(b)). Leaf shape and margin of the six samples are not correlated to the two clades of *C. album* (Table 1 and Figure 5(b)); however, leaf thickness of the six *C. album* presents correlation to the clades: the clade containing CAGAP004, CAJEJG05, and CVHUP01 shows thick leaves, called as a thick-leaf clade, and the clade consisting of CAGOH01, CCANG01, and CSJUK01 displays thin leaves, called as a thin-leaf clade (Table 1 and Figure 5(b)). NC_034950 clustered in the thick-leaf clade was not possible to be confirmed whether its leaves are thick or not. Taken together, the phylogenetic relationship of the six *C. album* chloroplast genomes seems not to be highly correlated with their morphological features and geographical locations, supporting that their high plasticity of morphology links to other factors such as nuclear markers, polyploidy, or any regulatory factor of leaf morphologies. With the additional chloroplast genomes as well as nuclear marker sequences of *C. album* collected in Korea, these relationships of morphological features and geographical locations will be more explicit.

### 3. Discussions

#### 3.1. Species Incongruency of *C. album* Chloroplast Genomes

In this study, we sequenced six chloroplast genomes of *C. album* s. str. collected in Korea displaying various morphological features. One of the previously sequenced chloroplast genomes of *C. album*, MF418659, is quite different from the remaining *C. album* chloroplast genomes in the aspects of gene configuration of chloroplast genome (Table 3) as well as phylogenetic relationship (MF418659 chloroplast genome...
Figure 4: Number of normal SSRs identified on the seven Chenopodium album chloroplast genomes. (a) Displays the number of SSR types in each seven C. album samples, and the Y-axis indicates the number of normal SSRs. Five different colors indicate the five types of SSRs, monoSSRs, diSSRs, triSSRs, tetraSSRs, and pentaSSRs. The table below the X-axis presents numbers of SSRs along with samples and the five SSR types. (b) Shows the distribution of SSRs in noncoding, intron, intergenic, and coding regions along with the samples. The X-axis means the seven C. album samples, and the Y-axis indicates the number of normal SSRs. Four different colors correspond to the four different regions. The table below the X-axis shows the number of SSRs along with samples and regions. (c) Shows SSR density (# of SSRs/kb) of the LSC, SSC, and IR regions along with the samples. The X-axis means the seven C. album samples, and the Y-axis indicates the number of normal SSRs along with samples and the three regions. Three different colors in the bar graphs mean the three regions. The table below the X-axis shows SSR density along with samples and the three regions.
was located outside of the clade of Chenopodium and Atriplex in Figure 5). These differences indicate that MF418659 may neither be \textit{C. album} nor genus Chenopodium.

It is partially supported by the fact that the collection site of MF418659 is the Himalayan area in India [24] where \textit{C. album} has been mainly cultivated as crops [7–9]. Usually, species diversity of the Himalayan area is higher due to its wide variety of climates as well as various climatic perturbations that have been applied to different locations in the Himalayan area [29]. Considering the phylogenetic position of MF418659 (Figure 5), it is possible that MF418659 is misidentified or an unreported species which is very different from \textit{C. album}.

This kind of incongruence problem of species has sometimes been found during comparative analyses in plant species. For instance, two \textit{Magnolia} chloroplast genomes, \textit{Magnolia insignis} (NC_035657) and \textit{Magnolia alba} (NC_037005), were reported as examples of misidentification species based on the phylogenetic analysis based on complete chloroplast genomes [30]. This problem can occur due to not enough taxonomic coverage of whole chloroplast genomes or misidentification of the samples used in the studies because of difficulties in species identification based on morphologies. Therefore, the identification of MF418659 should be revised in some ways, such as species identification of the voucher used in the previous study or sequencing and analyzing more samples of \textit{C. album} collected in the Himalayan area.

3.2. Possible Causes of \textit{C. album} s. str. Morphological Variations at the Molecular Level. Based on cytogenic and nuclear molecular marker analysis of Chenopodium species, \textit{C. album} is distinct to \textit{C. ficifolium} (B genome diploid) and \textit{C. quinoa} (A genome tetraploid) [14], and two major groups of \textit{C. album} were identified based on the phylogenetic tree based on \textit{rrn5} and ITS sequences [14]. In comparison to the phylogenetic tree which displays that \textit{C. album} and \textit{C. ficifolium} were clustered in one clade with high supportive values (Figure 5), a maternal lineage of both species is nearer than that of biparental lineage. Several \textit{C. quinoa} chloroplast genomes were clustered in the distinct clade to that of \textit{C. album} and \textit{C. ficifolium} (Figure 5), reflecting the different types of their genomes [14]. This phylogenetic tree based on complete chloroplast genomes (Figure 5) also indirectly supports that various intraspecific evolutionary events in several Chenopodium species, including \textit{C. album} and \textit{C. quinoa}, may have occurred, such as hybridization and polyploidization [2, 14].

Polyploidization and hybridization events can usually cause morphological plasticity and diversity; e.g., \textit{Nicotiana} species display various flower colors based on events of polyploidization [31], and \textit{Centaura} stoeba, polyploidy species, shows that it causes various phenotypes to climate, resulting in boosting its invasion [32]. Similarly, morphological variations of \textit{C. album} are not related to maternal lineage (Figures 1 and 5). It can be inferred that \textit{C. album} presents various morphological differences because it is hexaploid.
species. It can also be interpreted that these morphological variations are not fully genetically fixed but may be caused by nuclear genes related to leaf development, such as Class I KNOX genes, homeobox transcription factors which can regulate leaf shapes in Arabidopsis thaliana [33]. In this study, we found that the leaf morphology of six C. album has a weak correlation with their phylogenetic relation (Figures 1 and 5). If diverse leaf shapes of C. album are caused by these key regulators, we can deduce that the general trend of evolutionary process inferred from organelle genomes including chloroplast cannot explain this diversity because these regulators can display different evolutionary speeds and patterns from those of organelle genomes.

Several studies tried to delimitate species presenting different morphological features using whole chloroplast genome sequences. For instance, a phylogenetic tree constructed based on chloroplast genomes of Anemopaegma acutifolium supported that two leaf morphological trait types of A. acutifolium were caused by different maternal origins [34]. In the case of Triprostegia glandulifera and T. grandi-flora, their chloroplast genomes were used for solving the boundary of the two species; however, they could not explain the high morphological plasticity of them [35]. Therefore, further analyses with more chloroplast genomes of C. album expressing various phenotypic characteristics will be necessary to understand the origin of its morphological plasticity.

3.3. Evaluation of Level of Intraspecific Variations on C. album Chloroplast Genomes. The intraspecific variations identified among the seven Chenopodium album chloroplast genomes (56 SNPs and 26 INDEL regions) are compared with the previous studies which investigated intraspecific variations on chloroplast genomes. Twenty cultivars and wild types of Ricinus communis (Castor bean) displayed 162 SNPs and 92 INDEL regions [28], which is three times more than those of C. album. Sixty-three chloroplast genomes of Macadamia integrifolia (Macadamia nut) are collected in eastern Australia, which is a smaller geographical range of C. album. DNA was extracted from fresh leaves of the six samples using a DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany).

4. Materials and Methods

4.1. DNA Extraction of Natural Collection of Korean C. album. Six samples of C. album were collected in various places in the Korean Peninsula (Table 1 and Figure 1). All vouchers of the six samples were deposited to the Sung Kyun Kwan University Herbarium (SKKU; Table 1). Their total DNA was extracted from fresh leaves of the six samples using a DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany).

4.2. Genome Sequencing and De Novo Assembly of the Natural Collection of Six C. album Chloroplast Genomes. Genome sequencing was performed using HiSeqX at Macrogen Inc., Korea, from the extracted DNA of the six C. album. De novo assembly with confirmation was accomplished with Velvet v1.2.10 [40] after filtering raw reads using Trimomatic v0.33 [41]. After obtaining the first draft of the chloroplast genome sequences, gaps were filled with GapCloser v1.12 [42], and all bases from the assembled sequences were confirmed by checking each base in the alignment (tvem mode in SAMtools v1.9 [43]) against the assembled chloroplast genome generated with BWA v0.7.17 [44]. All these processes were conducted under the environment of the Genome Information System (GeIS; http://geis.infoboss.co.kr; Park et al., in preparation) like other Amaranthaceae chloroplast genomes assembled [38, 39, 45–49].

4.3. Chloroplast Genome Annotation. Geneious Prime® 2020.2.4 (Biomatters Ltd, Auckland, New Zealand) was used for chloroplast genome annotation based on the C. album chloroplast genome (NC_034950) [23] by transferring annotations while correcting exceptional cases, including missing start or stop codons. tRNA was predicted and confirmed based on the prediction of tRNAscan-SE v20.6 [50]. A circular map of C. album chloroplast was drawn by using the OGDRAW v1.3.1 [51].

4.4. Identification of Sequence Variations on the Complete Chloroplast Genomes of C. album. Single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs) were identified from the pair-wise alignments of two selected chloroplast genomes conducted by MAFFT v7.450 [52]. When the number of INDELs was calculated, continuous INDEL bases were considered one INDEL. In addition, we denote the four regions: (i) coding region is exon that encodes a protein, (ii) intron regions indicate the region which does not translate inside protein-coding genes, (iii) intergenic regions are the sequence between two genes, and (iv) noncoding region means the sequence located in tRNAs or rRNAs.
4.5. Identification of Simple Sequence Repeats (SSRs). Simple sequence repeats (SSRs) were identified on the chloroplast genome sequence using the pipeline of the SSR database (SSRDB; http://ssrdb.infoboss.co.kr/; Park et al., in preparation). Based on the conventional definition of an SSR on the chloroplast genome, monoSSR (1 bp) to hexaSSR (6 bp), the total length of SSRs on the chloroplast genome exceeds 10 bp. Owing to the different criteria of SSRs on chloroplast genomes, we adopted the criteria used in chloroplast genomes of Diaphania [47] and Arabidopsis thaliana [53] and mitochondrial genome of Rosa rugosa [54] as follows: the monoSSR (unit sequence length of 1 bp) to hexaSSR (6 bp) are used as normal SSRs, and heptaSSR (7 bp) to decaSSR (10 bp) are defined as extended SSRs. Among the normal SSRs, pentaSSRs and hexaSSRs for which the repeat number of unit sequences is 2 are classified as potential SSRs. Classification of regions on chloroplast genome was conducted in the same way described in the above section.

4.6. Comparison of SSRs Identified from Seven C. album Chloroplast Genomes. SSRs identified from seven C. album chloroplast genomes were compared on their flanking sequences under the environment of the SSRDB (http://ssrdb.infoboss.co.kr/; Park et al., in preparation). The pipeline of the SSR comparison implemented in the SSRDB used in various organelle genome studies [53, 55] was used with the following conditions: a cut-off e value of 1e−10 and a maximum flanking sequence for the comparison of 60 bp.

4.7. Nucleotide Diversity Analysis. Nucleotide diversity was calculated using the method proposed by Nei and Li [56] based on the multiple sequence alignment of Chenopodium chloroplast genomes using a Perl script used in previous studies [47, 53, 57]. The window size was set to 500 bp, and the step size was 200 bp when using the sliding-window method. Genomic coordination of each window was compared to the gene annotation of the chloroplast genome under the GenomeArchive® (http://www.genomearchive.net/) [58] environment for further analyses.

4.8. Construction of Phylogenetic Trees. The whole 34 Amaranthaceae chloroplast genomes and one outgroup of Gymnocalycium przewalskii chloroplast genome were aligned by MAFFT v7.450 [52], and alignment quality was checked manually. The maximum likelihood (ML) tree was reconstructed in IQ-TREE v1.6.6 [59]. In the ML analysis, a heuristic search was used with nearest-neighbor interchange (NNI) branch swapping, TVM+F+R4 model, and uniform rates among sites. All other options used the default settings. Bootstrap analyses with 1,000 pseudoreplicates were conducted with the same options. The posterior probability of each node was estimated by Bayesian inference (BI) using the MrBayes v3.2.7a [60] plug-in implemented in Geneious Prime® 2020.2.4 (Biomatters Ltd, Auckland, New Zealand). The HKY85 model with gamma rates was used as a molecular model. A Markov chain Monte Carlo (MCMC) algorithm was employed for 1,100,000 generations, sampling trees every 200 generations, with four chains running simulta-

5. Conclusions

We completed the six chloroplast genomes of Chenopodium album showing various morphological features. The structure and gene order of chloroplast are conserved among seven C. album including the previously sequenced chloroplast genome (NC_034950). The average nucleotide diversity calculated from the seven C. album chloroplast genomes is 0.0000625, and a total of 56 SNPs and 26 INDEL regions are found. In comparison to the other cases of chloroplast intraspecific variations, C. album chloroplasts present a low level of sequence variation. The number of normal SSR identified from the seven C. album chloroplast genomes ranges from 33 to 35 displaying similar distribution and density of SSRs. Interestingly, specific SNPs and INDEL regions in intronic and intergenic regions make SSR variation among the seven chloroplasts. All seven C. album chloroplast genomes are clustered in high supportive values of ML and BI trees with a short length of branches. In addition, one of the morphological characters of C. album s. str., the thickness of leaves, presented correlation with the phylogenetic position. Taking together the results in this study, our six chloroplast genomes of C. album s. str. will provide the way to investigate intraspecific features of chloroplast genomes, also the insights of intraspecific variations to understand various characteristics of one species including morphological features.

Data Availability

Chloroplast genome sequences of C. album sequenced in this study can be accessed via accession numbers MW446241 to MW446246 in NCBI GenBank.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors’ Contributions

Jongsun Park are Juhyeon Min are co-first authors.

Acknowledgments

This work was supported by the InfoBoss Research Grant (IBG-0001) and Grant from Rural Development Administration (PJ01385505). We also thank Dr. Suhyeon Park for discussing about the misidentification problem of chloroplast genomes and Mr. Woochan Kwon for giving constructive ideas for a better manuscript.

Supplementary Materials

Supplementary Table 1: list of SSRs identified in CAGAP004 of C. album. Supplementary Table 2: list of SSRs identified in CAGOH01 of C. album. Supplementary Table 3: list of SSRs identified in CAJEJG05 of C. album. Supplementary Table 4:
list of SSRs identified in CCANG01 of C. album. Supplementary Table 5: list of SSRs identified in CSJUK01 of C. album. Supplementary Table 6: list of SSRs identified in CYHU01 of C. album. Supplementary Table 7: list of SSRs identified in NC_034950 of C. album. (Supplementary Materials)

References


