ISSN 1479-2621

Plant Genetic Resources; 1–10
doi:10.1017/S1479262118000485

Comparative analysis of the complete chloroplast genome sequences of three *Amaranthus* species

Su-Young Hong[1]†, Kyeong-Sik Cheon[2]†, Ki-Oug Yoo[3], Hyun-Oh Lee[4], Manjulatha Mekapogu[5] and Kwang-Soo Cho[1]*

[1]*Highland Agriculture Research Institute, National Institute of Crop Science, Rural Development Administration, Pyeongchang, 25342, Republic of Korea,* [2]*Department of Biological Science, Sangji University, Wonju, Republic of Korea,* [3]*Department of Biology, Kangwon National University, Chuncheon, Republic of Korea,* [4]*Phygen Genomics Institute, Baekgoong Plaza 1, Bundang-gu, Seongnam, Republic of Korea and* [5]*Floriculture Research Division, National Institute of Horticultural & Herbal Science, Rural Development Administration, Wanju 55365, Republic of Korea*

Received 15 June 2018; Accepted 23 November 2018

Abstract

The complete chloroplast (cp) genome sequences of three *Amaranthus* species (*Amaranthus hypochondriacus, A. cruentus* and *A. caudatus*) were determined by next-generation sequencing. The cp genome sequences of *A. hypochondriacus*, *A. cruentus* and *A. caudatus* were 150,523, 150,757 and 150,523 bp in length, respectively, each containing 84 genes with identical contents and orders. Expansion or contraction of the inverted repeat region was not observed among the three *Amaranthus* species. The coding regions were highly conserved with 99.3% homology in nucleotide and amino acid sequences. Five genes – *matK*, *accD*, *ndhJ*, *ccsA* and *ndhF* – showed relatively high non-synonymous/synonymous values ($K_a/K_s > 0.1$). Sequence comparison identified two insertion/deletion (InDels) greater than 40 bp in length, and polymerase chain reaction markers that could amplify these InDel regions were applied to diverse Korean Genbank accessions, which could discriminate the three *Amaranthus* species. Phylogenetic analyses based on 62 protein-coding genes showed that the core Caryophyllales were monophyletic and Amaranthoideae formed a sister group with the Betoideae and Chenopodioideae clade. Comparing each homologous locus among the three *Amaranthus* species, identified eight regions with high Pi values (>0.03). Seven of these loci, except for *rps19-trnH* (*GUG*), were considered to be useful molecular markers for further phylogenetic studies.

Keywords: *Amaranthus*, chloroplast genome, InDel, phylogenetic analysis, tandem repeats

Introduction

Amaranthus is an annual crop believed to have originated in the South American Andes, although some scholars believe Central America or Central Asia as its true location of origin (Das, 2016). The genus *Amaranthus* consists of about 60–70 species and is widely distributed throughout the world's tropical, subtropical and temperate regions. *Amaranthus* is mainly used both as a seed and a leafy vegetable (Venskutonis and Kraujalis, 2013). The *Amaranthus* genus is divided into two sub-genera based on whether the species is identified as monoecious or dioecious (Sauer, 1967). The morphological and stem anatomy of *Amaranthaceae* were used to distinguish between the *Amaranthus* and *Albersia* subgenera, supporting the separation of the '*hybridus*' group from their presumed wild progenitors (Costea and DeMason, 2001). Aaron *et al.* (2001) studied pollen variations, Adhikary and Pratt

*Corresponding author. E-mail: kscholove@korea.kr
†These authors contributed equally to this work.

(2015) analysed pollen and floral variations within various species, and El-Ghamery *et al.* (2015) studied the root anatomies of 12 *Amaranthus* taxa.

Chloroplast (cp) is an organelle present in plant and algal cells which primarily functions in photosynthetic carbon fixation. Generally, cp genomes in angiosperms possess conserved quadripartite circular structures comprised of a large single copy (LSC) and a small single copy (SSC) region, as well as a pair of inverted repeats (IRs) (Jansen and Ruhlman, 2012; Cho *et al.*, 2016). Cp genomes are compact in size, present less recombination and possess fewer mutations genetically. The cp genome is therefore helpful for generating genetic markers for phylogenetic classification (Birky, 2001; Cho *et al.*, 2015; Hong *et al.*, 2017). Chaney *et al.* (2016) reported that the amaranth cp genome is highly conserved at the nucleotide level among the grain amaranth species. Nonetheless, single-nucleotide polymorphisms (SNPs), insertions/deletions (InDels) and polymorphic simple sequence repeat (SSRs) of cp genome in the *Amaranthus* taxa have not been identified. In terms of taxonomy, Xu and Sun (2001) inferred that the main grain amaranthus (*A. cruentus*, *A. caudatus*, *A. hypochondriacus*) were mixed among these related taxa and *A. caudatus* and *A. quitensis* were very closely related, whereas *A. powellii* was the most divergent taxon. Stetter and Schmid (2017) reported the phylogeny of the *Amaranthus* genus with 94 gene bank accessions by genotyping by sequencing (GBS) and found that it did not show differences among species assignment, but could be distinguished based on the region of origin (e.g. South or Central America). Nevertheless, only a few reports exist which comparatively analyse the complete cp genome sequences of *Amaranthus* species.

Amaranthus was introduced in Korea at end of the 1990s. However, *Amaranthus* as a crop is being cultivated after 2010. The three species, *A. hypochondriacus*, *A. cruentus* and *A. caudatus* are mainly cultivated for seeds, but they are grown in the field without species identification. Three *Amaranthus* species used in this study were identified to be belonging to the *Hybridus* clade in previous study (Waselkov *et al.*, 2018).

In this study, we report the high quality complete cp genome sequences for the three *Amaranthus* species and conducted comparative genomic analyses using tandem repeats (TRs), InDel polymorphisms and species-level identification using valuable markers and molecular phylogenetic analysis.

Materials and methods

Plant materials

The three *Amaranthus* species used in this study were acquired from the Germplasm Resources Information Network (GRIN, http://www.ars-grin.gov), USA and the National Agrodiversity Center of the Rural Development Administration (http://genebank.rda.go.kr), Korea (online Supplementary Table S1 and Fig. S1). For the cp genome assembly, the *A. hypocondriacus* accession (A6) was the same plant materials with Chaney *et al.* (2016) as cv. Plainsma and the *A. cruentus* and *A. caudatus* accessions was A7 (PI566897) and A15 (IT19999), respectively (online Supplementary Table S1). Four, six and one accession of *A. cruentus*, *A. caudatus* and *A. hypochondriacus* were used for InDel markers analysis, respectively. *Amaranthus* plants were grown in the Highland Agriculture Research Institute (HARI), Pyeongchang, Korea. Each sample constitutes approximately 100 mg of fresh leaves harvested from a single plant.

Chloroplast genome assembly and annotation

Genomic DNA was isolated from the leaves using a NucleoSpin Plant II kit (Macherey-Nagel, GmbH, Germany) following the manufacturer's instructions. For next-generation sequencing (NGS), pair-end (PE) libraries were constructed with an Illumina PE DNA library Kit and sequenced using an Illumina genome analyzer (Hiseq200) platform at Macrogen (http://www.macrogen.com/kor/). Reads with raw scores of 20 or less were removed from the total reads using the CLC-quality trim tool. About 2.7 and 2.0 gigabases (Gb) of total and trimmed reads respectively, were generated using NGS (online Supplementary Table S2). Cp genome assembly performed using a *de novo* assembly was implemented by CLC Genome Assembler (ver. 4.2.1, CLC Inc., Denmark). The parameters were set from 200 to 600 bp for the distance between forward start and reverse end reads (similarity = 0.8, and length fraction = 0.5). The putative cp contigs were selected from comparison with *Amaranthus hypochondriacus* (GenBank acc. KX279888) cp genome sequence as a reference by the nucmer tool in the MUMmer program (ver. 3.0). Selected contigs were merged into a single contig after which, the contig junctions and inner gaps were manually confirmed by read mapping using CLC read mapper applying the similar parameters as above. The average number of mapped bases and coverage of the three *Amaranthus* cp genomes were 91.7 Mb and 1474×, respectively (online Supplementary Table S2). Cp genome sequences were annotated using DOGMA (Wyman *et al.*, 2004) and via manual editing using comparisons with the reference species *A. hypochondriacus* (GenBank acc. KX279888). Circular maps of the cp genomes were obtained using OrganellarGenomeDRAW v1.2 (Lohse *et al.*, 2013).

Tandem repeat analysis

Nucleotide and amino acid diversity were analysed by BLASTN and BLASTP. TRs were analysed using Tandem

Repeat Finder (Benson, 1999) with advanced parameters. Alignment parameters match, mismatch and InDels were set to 2, 7, 7 and the minimum alignment score required to report a repeat was 50, the minimum length was 6 bp, and the motif identity percent was 100%.

InDel marker development and validation

Polymerase chain reaction (PCR) analysis was performed in 20 µl PCR mixtures containing 2× TOPsimple preMix-nTaq master mix (Enzynomics, Seoul, Korea) consisting of 0.2 U/µl n-taq DNA polymerase, 3 mM Mg^{2+} and a mixture containing 0.4 mM of each dNTP, with 10 pmol of each primer. The PCR reaction was performed in a thermocycler (Veriti, Applied Biosystems, CA, USA) using the following cycling parameters: 94°C (5 min); 35 cycles of 94°C (30 s), 58°C (30 s) and 72°C (1 min); and a final extension at 72°C (10 min). PCR products were analysed using 1.5% agarose gel electrophoresis and detected by DNA LoadingSTAR (DyneBio, Gyeonggi-do, Korea).

Phylogenetic analysis

In total, about 62 protein coding genes (online Supplementary Table S3) in 31 species (online Supplementary Table S4), including 30 core Caryophyllales species (one Aizoaceae, 14 Amaranthaceae, one Cactaceae and 14 Caryophyllaceae) and one outgroup (Polygonaceae; *Fagopyrum tataricum*) were compiled into a single file and was aligned with MAFFT v.7 (Katoh *et al*., 2002). Before maximum likelihood (ML) analysis, a search for the best fitting substitution model was performed using jModeltest v. 2.1.10 (Darriba *et al*., 2012). Based on the Akaike information criterion (AIC) and AIC with correction (AICc), GTR + I was the best model. ML analysis was performed using RAxML v7.4.2 (Stamatakis, 2006) with 1000 bootstrap replicates and the GTR + I model. Bayesian inference was performed using MrBayes v3.0b3 (Huelsenbeck and Ronquist, 2001).

Divergence hotspot identification in Amaranthus

The cp genome sequences of total four *Amaranthus* species, three of which (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*) discussed in this study and the other one (*A. tricolor*) published in a previous study (Viljoen *et al*., 2018), were analysed to identify the rapidly evolving molecular markers for the future phylogenetic studies of *Amaranthus*. The similarities among the three *Amaranthus* species were visualized using mVISTA (Frazer *et al*., 2004). Intergenic sequence (IGS) and gene regions were extracted separately from each plastid genome by applying the 'Extract' option in Geneious v.7.1.8 (Biomatters Ltd., Auckland, New Zealand). Homologous loci were then aligned individually using MAFFT v.7 (Katoh *et al*., 2002). To analyse nucleotide diversity (Pi), the total number of mutations (Eta), the average number of nucleotide differences (*K*) and parsimony informative characters were determined using DnaSP v.5.10 (Librado and Rozas, 2009). Non-synonymous (K_a) and synonymous (K_s) substitution values were calculated with the PAML 4 program (Yang, 2007). The procedure is as follows. First, we selected 62 protein coding genes and calculated K_a and K_s values for each gene for each species pairwise comparison (*A. caudatus*–*A. cruentus*, *A. caudatus*–*A. hypocondriacus* and *A. cruentus*–*A. hypocondriacus*). Second, we calculated the average K_a and K_s values for the three pairwise comparisons for each gene. Third, the ratio of K_a/K_s for each gene was based on the average K_a and K_s values from the three species comparison

Results

Complete chloroplast (cp) genome sequences

Each of the complete cp genomes of *A. hypochondriacus*, *A. cruentus* and *A. caudatus* consisted of a single circular molecule with a quadripartite structure (Fig. 1). The cp genome sizes of the *A. hypochondriacus*, *A. cruentus* and *A. caudatus* were 150,523, 150,757 and 150,523 bp, respectively. Its genome structures were comprised of a pair of IRs (IRa and IRb) separated by the LSC and one SSC region. The IRs for *A. hypochondriacus*, *A. cruentus* and *A. caudatus* totalled 24,352, 24,351 and 24,352 bp in length, respectively. The LSC sizes were 83,878, 84,101 and 83,878 bp, respectively, and the SSC sizes were 17,941, 17,954 and 17,941 bp, respectively (Table 1). The three genomes each contained 84 coding genes, accounting for 78,762 bp of the total genome length. Of these, 61, 6 and 12 genes were located in the LSC, IR and SSC regions, respectively (Fig. 1). The length of the coding sequences (CDS) of the three genomes was the same with 78,762 bp (average CDS length was 938 bp). The total number of RNA bases was 2704, and the overall GC-content was the same in all three genomes about approximately 36.6%. The complete cp genome sequences and Genbank files are available from the NCBI database (accession numbers MG83608, MG83606 and MG83605 for *A. caudatus*, *A. cruentus* and *A. hypochondriacus*, respectively). Chaney *et al*. (2016) have previously elucidated the whole genome sequence for *A. hypochondriacus* that was same accession, cv. Plainsman, using the PacBio RSII method, finding a total size of 150,518 bp (GenBank accession number KX279888), which was slightly (5 bp) shorter than our results indicated. And the two different sequencing platforms used which produced almost the same results. Looking at this discrepancy in further detail, we found that the 5 bp

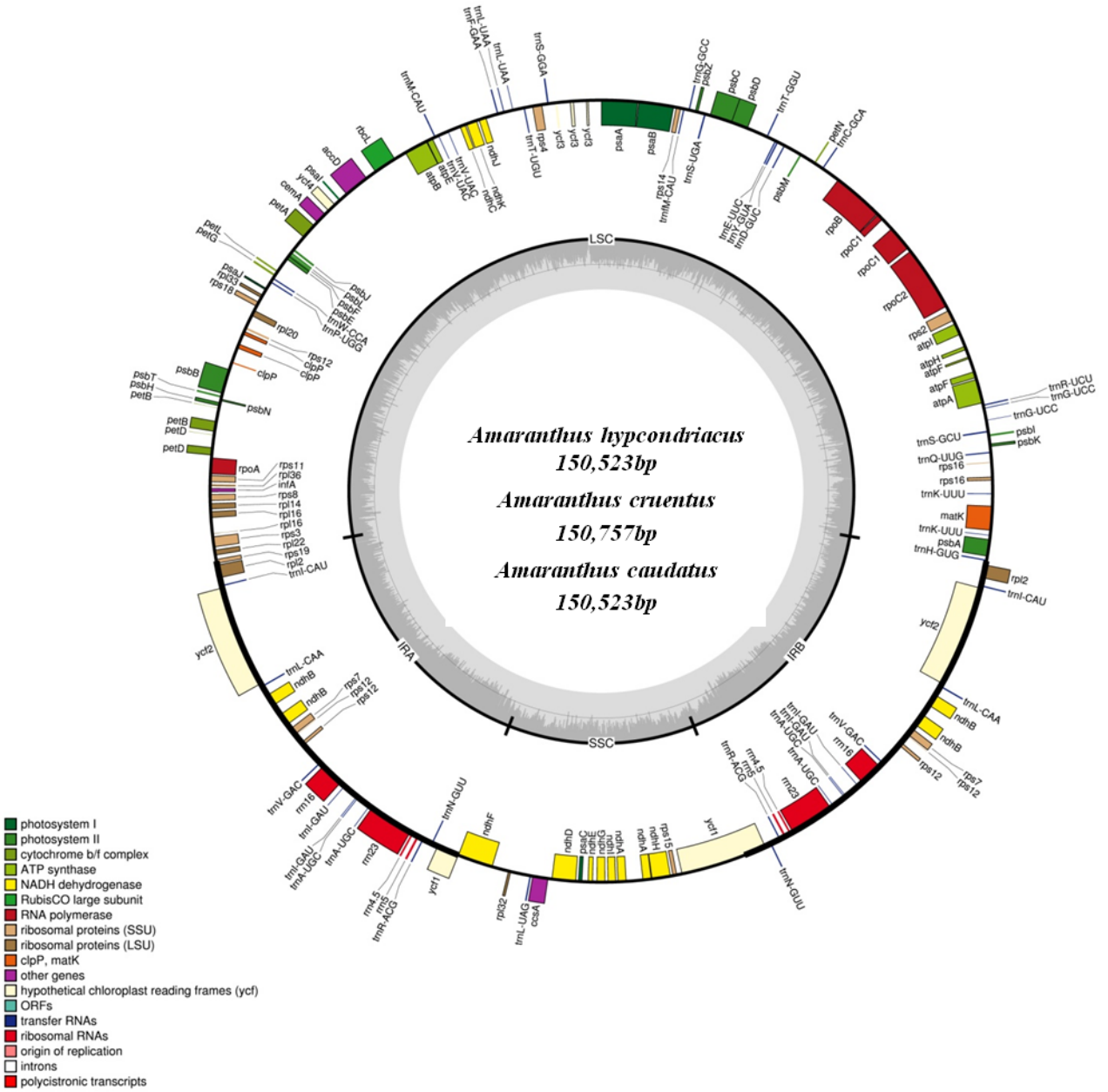

Fig. 1. The cp genome map for the three *Amaranthus* species. Genes shown inside the circle are transcribed clockwise and genes outside the circle are transcribed counterclockwise.

difference was located in IGSs between *petL* and *petG* in the T repeats in the LSC region (data not shown). There was a 20 bp poly T sequence identified in our study, while this same sequence was only 15 bp long in KX279888. The average sequencing coverage depth for the three genomes was 1747× (online Supplementary Table S2).

Cp genome gene content and hotspot regions

The complete cp genome genes of *A. hypochondriacus*, *A. cruentus* and *A. caudatus* were comparatively analysed. The gene content, order and orientation were similar in all the three species (Fig. 1). The IR/LSC and IR/SSC junction regions were compared with identify IR expansion or contraction. The *trnH*, *rpl2*, *ycf1*, *ndhF* and *rps19* genes were located in the LSC/IRa, IRa/SSC, SSC/IRb (both *ycf1* and *ndhF*) and IRb/LSC junction regions, respectively. The border position in the *A. hypochondriacus* cp genome was the same as that in the *A. cruentus* and *A. caudatus* cp genomes, implying the absence of IR expansion or contraction (online Supplementary Fig. S2). The coding regions of these three species were characterized to be highly

Table 1. Gene content comparison for the complete cp genomes of three *Amaranthus* species (*A. cruentus*, *A. caudatus and A. hypochondriacus*)

Features	*A. cruentus* (A7)	*A. caudatus* (A15)	*A. hypochondriacus* (A6)
Total sequence length (bp)	150,757	150,523	150,523
Large single copy length (bp)	84,101	83,878	83,878
Inverted repeat region length (bp)	48,702	49,064	49,064
Small single copy length (bp)	17,954	17,941	17,941
GC content (%)	36.6	36.6	36.6
No. of protein coding genes	84	84	84
Total CDS[a] bases (bp)	78,762	78,762	78,762
Average CDS length (bp)	938	938	938
No. of tRNA	37	37	37
Total tRNA bases (bp)	2704	2704	2704
Average tRNA length (bp)	75.1	75.1	75.1
Total tandem repeat length (bp)	1009	810	810
Average tandem repeat length (bp)	48	45	45
Average sequencing coverage depth (X)	1012	1931	1479

[a]Coding sequences.

conserved, presenting 99.78% homology at the amino acid level except the *matK* gene (Fig. 2). The overall similarity of nucleotide and amino acid sequences in the coding genes was 99.98 and 99.97%, respectively, with the IR region having the lowest identity. In general, the IR region, taken as a whole, was highly conserved compared with the LSC and SSC regions, also when each was taken as a whole. Owing to the presence of highly conserved coding regions, the K_a/K_s ratio was very low, approaching zero. Although the coding region is highly conserved, we observed a slight variation in the divergence of the coding genes, *matK*, *ndhJ*, *accD*, *ndhF* and *ccsA*, above 0.1 (Fig. 2). Among them three genes, *matK*, *ndhJ* and *accD*, and two genes, *ndhF* and *ccsA* were located in LSC and SSC regions, respectively.

Tandem repeats analysis

The TR sequences in the cp genome were characterized and compared among the three *Amaranthus* species. A total of 18 TRs with various sizes and repeat unit copy numbers were identified for *A. hypochondriacus* and *A. caudatus*, whereas around 21 TRs were observed for *A. cruentus*. The majority of the TRs (11 in *A. hypochondriacus* and *A. caudatus* and around 14 in *A. cruentus*) were found in the LSC region, while six were found in the IR region for all the three species. Only one TR was observed in the SSC region (online Supplementary Table S5). The total TR length of the three genomes was 1009, 810 and 810 bp in *A. cruentus*, *A. caudatus* and *A. hypocondriacus*, respectively (Table 1). The average TR length was 48 bp in *A. cruentus* and it was 45 bp in *A. caudatus* and *A. hypocondriacus* (Table 1). Most of the InDels were located in the IR region; two InDels (both longer than 40 bp) were in the IGSs for *trnK*, *rps16*, *psbM* and *trnD* for all three species, and were 74 and 45 bp long, respectively (online Supplementary Table S6).

InDel marker analysis

We designed InDel specific primers to confirm the InDel presence in intron sequences using PCR amplification (Table 2). ClustalW analysis of the InDel region for the three species also showed that both *A. caudatus* and *A. hypochondriacus* were identical, whereas *A. cruentus* and *A. tricolor* showed a variation in both InDel regions (online Supplementary Table S6). The presence of these two InDels in the three species was further confirmed by PCR amplification. We observed no variation between *A. hypochondriacus* and *A. caudatus*, which both yielded the amplicons of 493 and 302 bp for InDel_1 and InDel_2, respectively. However, *A. cruentus* showed amplicons of 567 and 347 bp for InDel_1 and InDel_2, respectively. Additionally, two InDels were found in the same regions with different size in *A. tricolor*. The expected amplicon sizes were 458 and 253 bp in the InDel_1 and InDel_2, respectively with *in silico* analysis (online Supplementary Table S6). Two of the *A. caudatus* accessions (A14 and A18) showed variable amplicon sizes compared with the other *A. caudatus* accessions for both InDel_1 and InDel_2 (Fig. 3). This may be possibly due to the misgrouping of A14 and A18 as the member of *A. caudatus*

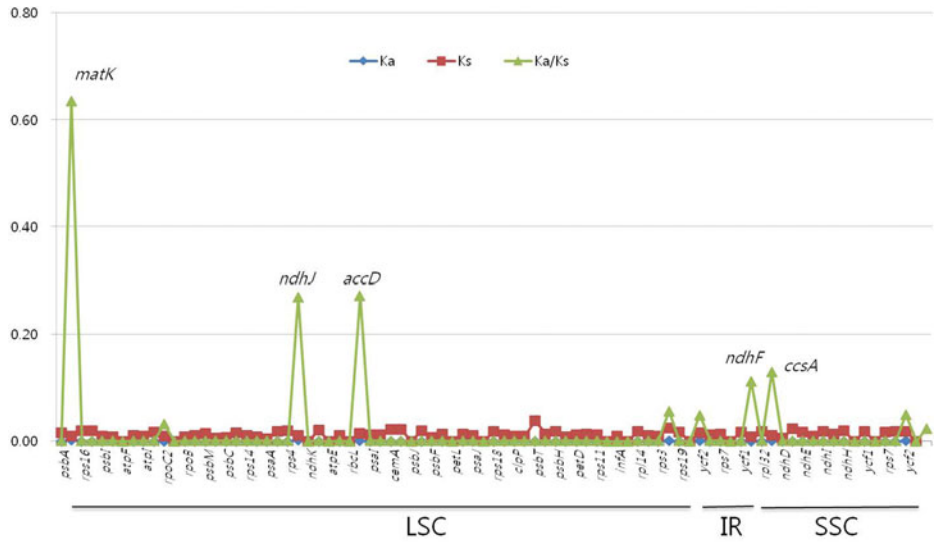

Fig. 2. Non-synonymous (K_a) and synonymous (K_s) substitution rates and K_a/K_s values among cp genomes for three *Amaranthus* species. Five genes (*matK*, *accD*, *ndhJ*, *ccsA* and *ndhF*) showed K_a/K_s values greater than 0.1. LSC, large single copy; IR, inverted repeat; SSC, small single copy.

as the plants from which the DNA was extracted was mis-identified.

Comparison of phylogenetic relationships with previous studies

A ML analysis was carried out based on 62 protein-coding genes from 31 taxa (Fig. 4). The length of the aligned sequences was 53,661 bp. The resultant phylogenetic tree generated using the aligned sequences resulted in a well-resolved topology supporting the monophyly of the tested families and subfamilies. The ML tree generated in this study was the same as the APG IV system (The Angiosperm Phylogeny, 2009), but it showed a slight difference from what was reported in a previous study. Specifically, Cactaceae and Aizoaceae formed a clade which was a sister group to Amaranthaceae and Caryophyllaceae. Also, Amaranthoideae formed a clade with Betoideae and Chenopodioideae. In contrast, the previous study had presented Aizoaceae forming the most basal branch, followed by Cactaceae, and Amaranthoideae forming a sister group to all the other subfamilies of Amaranthaceae (Hong *et al.*, 2017). Additionally, this study found that Chenopodioideae branched relatively late in Amaranthaceae, and *A. hypochondriacus* were the closest to *A. caudatus*. However, previous studies presented Chenopodioideae as the earliest diverging and forming a sister group to all the other subfamilies (Ogundipe and Chase, 2009), and also placed *A. hypochondriacus* and *A. cruentus* as having the closest relationship (Stetter and Schmid, 2017). Although Stetter and Schmid (2017) reported that *A. hypocondriacus* and *A. cruentus* could be grouped together based on the GBS and morphology, Viljoen *et al.* (2018) suggested that *A. hypocondriacus* and *A. caudatus* could be grouped together as leafy and grain amaranth by DNA barcoding of cp and internal transcribed sequences gene sequences similar to our result. We believe that these discrepancies would be due to the increased sequencing resolution resulting from the addition of more gene regions and using various species. However, our study analysed families, subfamilies, and species that are not relatively diverse. Therefore, in order to solve the various taxonomic problems currently unresolved, more extensive studies featuring more species would be required.

Selection of useful molecular marker regions for further phylogenetic studies

About 222 loci containing 113 genes and 109 IGS were compared among the three *Amaranthus* species. Sequence divergence ranged from 0 to 0.06701 (online Supplementary Fig. S3). The IR region was much more conserved compared with the LSC and SSC regions. Despite the multitude of phylogenetic studies already performed (Xu and Sun, 2001; Ogundipe and Chase, 2009; Park 2015; Stetter and Schmid, 2017), the phylogenetic relationships between *Amaranthus* species remain unclear. We believe that this is due to a paucity of nucleotide variations in the molecular markers used in previous studies. Indeed, the nucleotide variation in cpDNA molecular markers used in previous studies, including *matK*, *atpB-rbcL*, *rpoC1* and *trnL-F*, was only 0.0016–0.0074. The results of this study showed that

Table 2. Primer list for InDel validation for the cp genomes of four *Amaranthus* species

Primer	Sequence (5′–3′)		Expected size (bp)			
	Forward	Reverse	*A. hypochondriacus* and *A. caudatus*	*A. cruentus*	*A. tricolor*	Position
Indel_1	TATTGGGCGGAAGACAGAAA	GTGCCAATCCAACACAAAAA	493	567	458	*trnK*~*rps16*
Indel_2	TGGTTACAAAACCCCCAAAA	CACGAAAAGGAGGAGCGATA	302	347	253	*psbM*~*trnD*

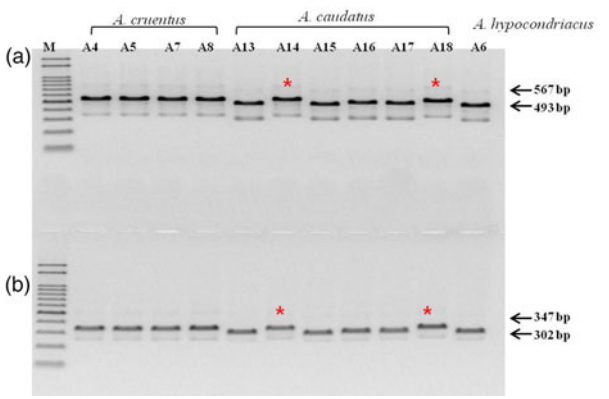

Fig. 3. PCR-amplification of InDel regions present in the cp genomes of *Amaranthus hypochondriacus* (A6), *A. cruentus* (A4, A5, A7, A8) and *A. caudatus* (A13, A14, A15, A16, A17, A18) with InDel specific markers. (a) InDel_1 marker. (b) InDel_2 marker. M: 100 bp DNA ladder. InDel marker information is shown in Table 2. Two of the *A. caudatus* accessions with asterisks (A14 and A18) showed variable amplicon sizes compared with the other *A. caudatus* accessions for both InDel_1 and InDel_2.

the Pi values in eight IGS regions, *rps19-trnH* (*GUG*), *rps16-trnQ* (*UUG*), *psbK-psbI*, *psbI-trnS* (*GCU*), *trnS* (*GCU*)-*trnG* (*UCC*), *atpI-rps2*, *rps22-rps19* and *ndhF-rpl32*, had relatively high calculated values (>0.03). However, *rps19-trnH* (GUG) is very short, with a length of 12–23 bp, and is therefore not considered suitable as a molecular phylogenetic marker. Thus, we presume that the remaining seven regions can potentially be useful for resolving many unclear phylogenetic relationships within the genus *Amaranthus*.

Discussion

Complete cp genome sequences are widely used in constructing plant evolutionary relationships (Xu and Sun, 2001). Here, we obtained the complete cp genome sequences of three *Amaranthus* species: *A. hypochondriacus*, *A. caudatus* and *A. cruentus*, and applied this information for their comparative analysis. Generally, repeat sequences are useful for studying genome rearrangement and play an important role in phylogenetic analysis (Qian *et al.*, 2013; Cheon *et al.*, 2017; Hong *et al.*, 2017). As reported by Huang *et al.* (2013), repeat occurrence is more prevalent in IGSs compared with genic sequences, a finding which was also confirmed in the current study. TRs and SSRs may possibly be related to cp genome size variation and divergence because of recombination (Hong *et al.*, 2017). In this study, TRs were prevalent in the LSC region and contributed to the *A. cruentus* cp genome being 199 bp longer than those of *A. caudatus* and *A. hypochondriacus* (Table 1 and online Supplementary Table S5). In a

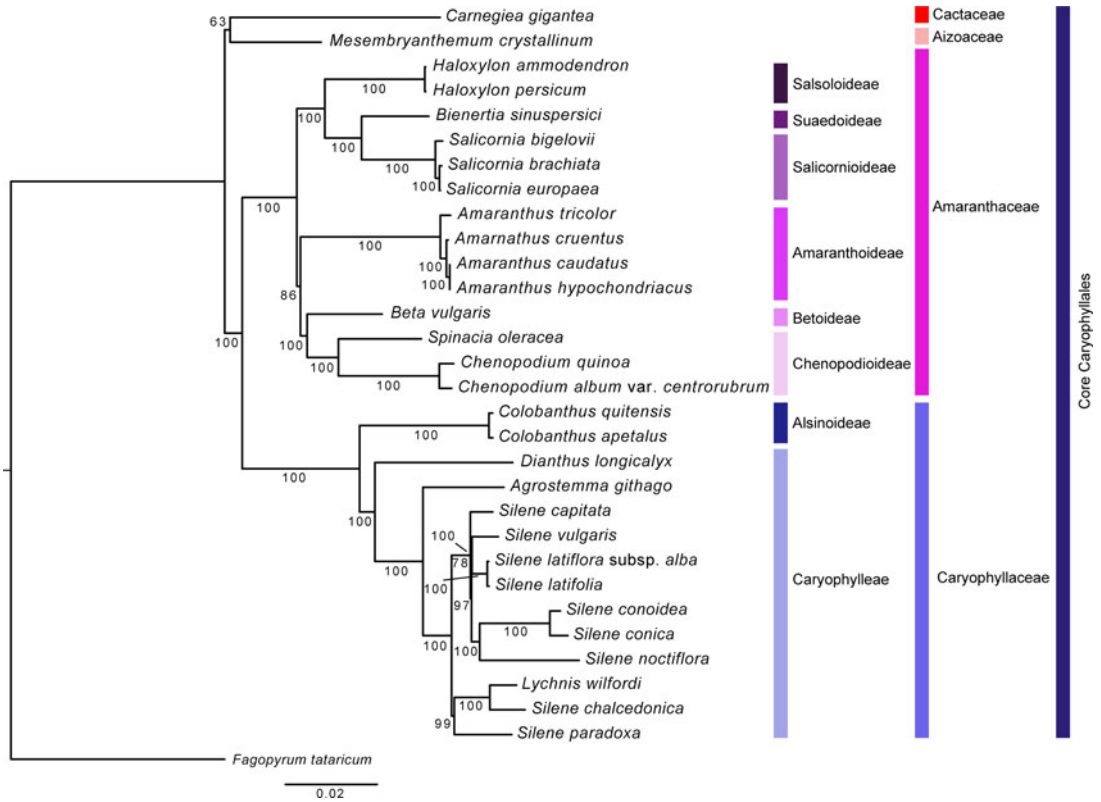

Fig. 4. Phylogenetic analysis of 31 species using 62 protein coding genes found in complete cp genome sequences. Out-group was set as *Fagopyrum tataricum*. The phylogenetic tree was generated using the ML method and bootstrap 1000-time iteration with MrBayes v3.0b3.

previous molecular phylogenetic study, Chaney *et al.* (2016) reported the assembly and annotation of the first reference-quality cp genome for the genus *Amaranthus*. In our study, the cp genome retained the quadripartite structure and nucleotide level was highly conserved among the grain amaranths. SNPs, InDels and polymorphic SSRs were identified and can serve as genetic markers in future studies. An earlier study showed that the *Chenopodium* species of Amaranthaceae formed a polyphylogenetic group (Hong *et al.*, 2017). In the current study, cp region nucleotide diversity was relatively low (*trnL* (*UAA*)-*trnF* (*GAA*), 0.00162; *matK*, 0.00737; *trnK-UUU*, 0.00534; *atpB*, 0.00256; *atpB-rbcL*, 0.00347; *rbcL*, 0.00187). The level of nucleotide diversity was similar to that of *Chenopodium* species (*trnL-trnF*, 0.01918; *matK*, 0.00982; *trnK-UUU* intron, 0.01359; *atpB*, 0.00601; *atpB-rbcL*, 0.00689; *rbcL*, 0.00493) (Hong *et al.*, 2017). Although the coding regions were highly conserved, three genes (*matK*, *accD*, *ndhJ*) from the LSC, *ccsA* and *ndhF* from the SSC returned non-synonymous to synonymous (K_a/K_s) values greater than 0.1 relative to other genes. The K_a/K_s value for *matK* was higher than that of the *accD* gene in our study, which agreed with the earlier report of Cuenoud *et al.* (2002). Comparing the nucleotide diversity amongst the three regions, the IR region was highly conserved relative to the LSC and SSC regions. Earlier reports have also shown that the IR region diverged at a slower rate compared with the LSC and SSC regions (Huang *et al.*, 2013; Wang *et al.*, 2017).

Our study identified nine high sequence variations in intergenic regions as follows: *rps19-trnH* (*GUG*), *rps16-trnQ* (*UUG*), *psbK-psbI*, *psbI-trnS* (*GCU*), *trnS* (*GCU*)-*trnG* (*UCC*), *atpI-rps2*, *trnG* (*GCC*)-*trnfM* (*CAU*), *rpl22-rps19*, *ndhF-rpl32* (online Supplementary Fig. S3). These regions were considered as useful markers for elucidating phylogenetic relationships amongst *Amaranthus* species. However, when selecting suitable molecular markers, the length of the amplified regions must also be considered. The five aforementioned intergenic regions, *rps19-trnH* (*GUG*), *psbI-trnS* (*GCU*), *atpI-rps2*, *trnG* (*GCC*)-*trnfM* (*CAU*) and *rpl22-rps19*, have relatively short InDels (19, 2, 6, 27 and 19 bp, respectively). Therefore, it is insufficient for reproducing nucleotide variation in various taxa (Dong *et al.*, 2012; Cho *et al.*, 2017). The other four regions (*rps16-trnQ* (*UUG*), *psbK-psbI*, *trnS* (*GCU*)-*trnG* (*UCC*) and *ndhF-rpl32*) were suitable for use in phylogenetic analysis of *Amaranthus* species and helpful for evaluating unresolved phylogenetic relationships.

We applied InDel markers to validate our cp genome sequencing results against *Amaranthus* accessions for *A. hypochondriacus* (A6), *A. cruentus* (A4, A5, A7, A8) and *A. caudatus* (A13, A14, A15, A16, A17, A18). *A. cruentus* showed a variation in amplicon size compared with *A. caudatus* and *A. hypochondriacus* for both InDel markers. Among the *A. caudatus* accessions, A14 and A18 showed a complete variation in amplicon size, showing the results similar to amplicon sizes for *A. cruentus*. Although A14 and A18 were initially classified as *A. caudatus* accessions, based on InDel marker amplification results, they may belong to *A. cruentus* instead. Alternatively, inter-species hybridization may also be the possible reason for this observed size similarity. In nature, a number of spontaneous inter-specific hybrids have been reported among grain species (Sauer, 1967). Most *Amaranthus* hybrids exhibit relatively high levels of sterility (Gupta and Gudu, 1991). However, it is possible that the parental species could hybridize relatively easily (Lanta *et al.*, 2003). Hence, it is expected that these markers could be present due to natural or artificial inter-specific hybridization occurring in *Amaranthus* taxa. The InDel markers can be identified by *in silico* analysis to also distinguish *A. tricolor* from other *Amaranthus* species.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S1479262118000485.

Acknowledgements

This work was carried out with the support of 'Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ01135402),' Rural Development Administration, Republic of Korea.

Author contribution

S-Y Hong conceived the design of the study, analysed the data and drafted the manuscript. K-S Cheon and H-O Lee performed the bioinformatics work. K-O Yoo collected and identified samples. M Mekapogu grew and collected samples of *Amaranthus* germplasm in HARI. K-S Cho was responsible for data analysis and writing of the manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical statements

This study does not contain any studies with human participants or animals performed by any of the authors and hence a formal consent is not required.

References

Aaron SF, Daniel ZS, Al-Khatib K and Michael JH (2001) Pollen morphological differences in *Amaranthus* Species and interspecific hybrids. *Weed Science* 49: 732–737.

Adhikary D and Pratt DB (2015) Morphologic and taxonomic analysis of the weedy and cultivated *Amaranthus* hybridus Species Complex. *Systematic Botany* 40: 604–610.

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

Birky CW Jr (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics* 35: 125–148.

Chaney L, Mangelson R, Ramaraj T, Jellen EN and Maughan PJ (2016) The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). *Applications in Plant Sciences* 4: apps.1600063.

Cheon K-S, Kim K-A and Yoo K-O (2017) The complete chloroplast genome sequences of three Adenophora species and comparative analysis with Campanuloid species (Campanulaceae). *PLoS ONE* 12: e0183652.

Cho K-S, Cheon K-S, Hong S-Y, Cho J-H, Im J-S, Mekapogu M, Yu Y-S and Park T-H (2016) Complete chloroplast genome sequences of *Solanum commersonii* and its application to chloroplast genotype in somatic hybrids with *Solanum tuberosum*. *Plant Cell Reports* 35: 2113–2123.

Cho K-S, Hong S-Y, Yun B-K, Won H-S, Yoon Y-H, Kwon K-B and Mekapogu M (2017) Application of InDel markers based on the chloroplast genome sequence for authentication and traceability of Tartary and common buckwheat. *Czech Journal of Food Sciences* 35: 122–130.

Cho K-S, Yun B-K, Yoon Y-H, Hong S-Y, Mekapogu M, Kim K-H and Yang T-J (2015) Complete chloroplast genome sequence of Tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (*F. esculentum*). *PLoS ONE* 10: e0125332.

Costea M and DeMason DA (2001) Stem morphology and anatomy in *Amaranthus* l. (Amaranthaceae), taxonomic significance. *The Journal of the Torrey Botanical Society* 128: 254–281.

Cuenoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ and Chase MW (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *American Journal of Botany* 89: 132–144.

Darriba D, Taboada GL, Doallo R and Posada D (2012) Jmodeltest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772.

Das S (2016) Taxonomy and phylogeny of grain amaranths. In: *Amaranthus: A Promising Crop of Future*. Singapore: Springer Singapore, pp. 57–94. doi: 10.1007/978-981-10-1469-7_5.

Dong W, Liu J, Yu J, Wang L and Zhou S (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* 7: e35071.

El-Ghamery AA, Sadek AM and Abd Elbar OH (2015) Root anatomy of some species of *Amaranthus* (amaranthaceae) and

formation of successive cambia. *Annals of Agricultural Sciences* 60: 53–60.

Frazer KA, Pachter L, Poliakov A, Rubin EM and Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32: W273–W279.

Gupta V and Gudu S (1991) Interspecific hybrids and possible phylogenetic relations in grain amaranths. *Euphytica* 52: 33–38.

Hong S-Y, Cheon K-S, Yoo K-O, Lee H-O, Cho K-S, Sih J-T, Kim S-J, Nam J-H, Sohm J-B and Kim Y-H (2017) Complete chloroplast genome sequences and comparative analysis of *Chenopodium quinoa* and *C. album*. *Frontiers in Plant Science* 8: 1696.

Huang Y-Y, Matzke AJM and Matzke M (2013) Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). *PLoS ONE* 8: e74736.

Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *(Oxford, England)* 17: 754–755.

Jansen RK and Ruhlman TA (2012) Plastid genomes of seed plants. In: *Genomics of Chloroplasts and Mitochondria*. Springer, pp. 103–126.

Katoh K, Misawa K, Kuma K and Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

Lanta V, Havranek P and Ondrej V (2003) Morphometry analysis and seed germination of *Amaranthus cruentus*, *A. retroflexus* and their hybrid (*A. x turicensis*). *Plant Soil and Environment* 49: 364–369.

Librado P and Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *(Oxford, England)* 25: 1451–1452.

Lohse M, Drechsel O, Kahlau S and Bock R (2013) OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* 41: W575–W581.

Ogundipe OT and Chase M (2009) Phylogenetic analyses of Amaranthaceae based on matK DNA sequence data with emphasis on West African species. *Turkish Journal of Botany* 33: 153–161.

Park Y-H (2015) *A taxonomic study of genus* Amaranthus *in Korea*. Master Thesis, Graduate school of Kangwon National University.

Qian J *et al.* (2013) The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS ONE* 8: e57607.

Sauer JD (1967) The grain amaranths and their relatives: a revised taxonomic and geographic survey. *Annals of the Missouri Botanical Garden* 54: 103–137.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *(Oxford, England)* 22: 2688–2690.

Stetter MG and Schmid KJ (2017) Analysis of phylogenetic relationships and genome size evolution of the *Amaranthus* genus using GBS indicates the ancestors of an ancient crop. *Molecular Phylogenetics and Evolution* 109: 80–92.

The Angiosperm Phylogeny G (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. *APG III Botanical Journal of the Linnean Society* 161: 105–121.

Venskutonis PR and Kraujalis P (2013) Nutritional components of amaranth seeds and vegetables: a review on composition, properties, and uses. *Comprehensive Reviews in Food Science and Food Safety* 12: 381–412.

Viljoen E, Odeny DA, Coetzee MPA, Berger DK and Rees DJG (2018) Application of chloroplast phylogenomics to resolve species relationships within the plant genus Amaranthus. *Journal of Molecular Evolution* 86: 216–239.

Wang W, Yu H, Wang J, Lei W, Gao J, Qiu X and Wang J (2017) The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *International Journal of Molecular Sciences* 18: 2288.

Waselkov JE, Boleda AS and Olsen KM (2018) A phylogeny of the genus *Amaranthus* (amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. *Systematic Botany* 43: 439–458.

Wyman SK, Jansen RK and Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* (Oxford, England) 20: 3252–3255.

Xu F and Sun M (2001) Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (*Amaranthus*; Amaranthaceae) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence repeat markers. *Molecular Phylogenetics and Evolution* 21: 372–387.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.