Characteristics of the Completed Chloroplast Genome Sequence of *Xanthium Spinosum:* Comparative Analyses, Identification of Mutational Hotspots and Phylogenetic Implications

Gurusamy Raman
Yeungnam University
KyuTae Park
Yeungnam University
Joo Hwan Kim
Gachon University
SeonJoo Park (sjpark01@ynu.ac.kr)
Yeungnam University

Research article

Keywords: Xanthium spinosum, hotspot regions, divergence, Ambrosiinae, chloroplast genome, SSR markers, accD

Posted Date: August 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-60093/v1

Version of Record: A version of this preprint was published at BMC Genomics on December 2nd, 2020. See the published version at https://doi.org/10.1186/s12864-020-07219-0.

Abstract

Background: The invasive alien species, *Xanthium spinosum* has been used as a traditional Chinese medicine for many years. Unfortunately, there are no extensive molecular studies for this plant.

Results: Here, the complete chloroplast genome sequence of *X. spinosum* was assembled and analyzed. The cp genome of *X. spinosum* was 152,422 bp and possessed quadripartite circular structure. The cp genome contained 115 unique genes, including 80 protein-coding genes, 31 tRNA genes and 4 rRNA genes. Comparative analysis revealed that *X. spinosum* encoded a higher number of repeats (999 repeats) and 701 SSRs in their cp genome. Also, fourteen divergences (Pi > 0.03) were found in the intergenic regions. The *accD* gene underwent positive selection within Heliantheae, which contributes to further investigation of the adaptive plant evolution in the ecosystem. Additionally, the phylogenetic analysis revealed that *Parthenium* is a sister clade to both *Xanthium* and *Ambrosia* and it is an early-diverging lineage of subtribe Ambrosiinae though it supports with very weak bootstrap value.

Conclusion: The identified hotspot regions were thought to be useful molecular markers for resolving phylogenetic relationships and species validation of *Xanthium*.

Background

The invasive alien species, *Xanthium spinosum* belongs to the Asteraceae family and it is one of the genera of subtribe Ambrosiinae (Heliantheae) that includes annual and perennial herbaceous plants [1]. The plant distributed across the world including Canada, the US, Central, and South America, parts of Africa, the Middle East, Russia, China, Australia and the Korean peninsula [2]. The genus, *Xanthium* has been widely used for various traditional medicinal treatments in different countries [3]. The plant parts of the *X. spinosum* are used for the treatment of cancer and diarrhea [4, 5] intermittent fever hydrophobia and rabies [6] and rheumatoid arthritis [7] and also has antibacterial [6] and antiviral properties [6, 8–10]. Though several antimicrobial and their functional studies of *X. spinosum* have been carried out for the past five decades, there are no exclusive genetic or genomic studies so far.

Universal molecular markers such as plastid genes *rbcL* and *psbA* and nuclear *ITS* have been widely used for the identification of some plant species quickly and precisely but unsuccessful to differentiate between the very closely related species [11–13]. The genus, *Xanthium* is commonly known as cock-labur, which is close relative to the *Ambrosia* genus. Also, the number of species in the *Xanthium* genus is still debate and considered as this genus comprises five to more than twenty species [14–17]. Recently, Tomasello and Heubl (2007) used plastid (*psbA-trnH*, *trnQ-rps16* and *trnL*-F) markers and nuclear (ITS and D35) markers to resolve the phylogenetic position of *Xanthium* and identified that the genus *Parthenium* as the early diverging lineage in the subtribe Ambrosiinae. In contrast, Somaratne et al., (2019) used 46 cp protein-coding genes to resolve the phylogenetic position of *Xanthium* and *Parthenium* and revealed that *Parthenium* is not an early-diverging lineage of subtribe Ambrosiinae. However, most plant chloroplast genomes are highly conserved structures and considered as efficient molecular markers for the identification of plant species in the genome-wide evolutionary studies which deliver significant quantities of genetic information, resolving taxonomic and phylogenetic relationships [18, 19].

In the present study, the complete chloroplast genome of *X. spinosum* was assembled and analyzed. Comparative analysis revealed that the IR regions and protein-coding sequence regions are highly conserved, and several highly variable regions were in the intergenic regions. Also, the phylogenetic position of *Xanthium* and early-diverging lineage of subtribe Ambrosiinae is resolved using 79 protein-coding genes. Further, fourteen hotspots were identified to distinguish the two species. Furthermore, the two sites of protein-coding gene *accD* undergoing positive selection that contributes to further investigation on the adaptive evolution in the plant ecosystem. Overall, the sequencing of the *X. spinosum* chloroplast genome will be useful for enhancing the species identification and provide new insight for understanding the phylogenetic position of *Xanthium* and early-diverging lineage of subtribe Ambrosiinae, hotspot mutations and adaptive evolution in ecosystems.

Results

General features of the chloroplast genome and organization

The complete chloroplast genome of *X. spinosum* was 152,422 bp in length. The cp genome shows a typical quadripartite structure and contained two short inverted repeats (IRa and IRb) regions (25,075 bp) which were separated by a small single-copy (SSC)

region (18,083 bp) and a large single-copy (LSC) region (84,189 bp) (Fig. 1). The cp genome encodes 115 unique genes, including 80 protein-coding genes, 31 transfer RNA (tRNA) genes and 4 ribosomal RNA (rRNA) genes. Six protein-coding, six tRNA and four rRNA genes were duplicated in their IR regions. The overall GC content of the cp genome was 37.4% while that of LSC, SSC and IR regions was 35.4%, 31.2% and 43%, respectively (Table 1).

Comparative analysis of the *Xanthium* sp.

The cp genome border LSC-IRb and SSC-IRa of *X. spinosum* are compared with three other closely related species of Heliantheae such as *X. sibiricum*, *Ambrosia artemisiifolia* and *Parthenium argentatum* [20, 21] (Fig. 2). The intact copy of the gene *rps19* is present in the LSC/IRb border of the *X. spinosum*, *A. artemisiifolia* and *P. argentatum* and shares 95 bp to 119 bp in the IRb region that leads to *rpl2* gene is located in the IRb region. In contrast, the *X. sibiricumrps19* gene is completely shifted to the LSC region and 71 bp away from the IRb region and although the *rpl2* gene is present in the LSC/IRb border. Besides, 154 – 175 bp partially fragmented *rps19* gene of all these four species is present in the IRa border. The pseudogene, *ycf1* is present in the IRa/SSC border of X. *spinosum* on the other hand ψ*ycf1* is either located in the IRb region (*X. sibiricum* and *A. artemisiifolia*) or in the SSC region (*P. argentatum*) of the respective cp genomes. But the *ndhF* gene is entirely present in the SSC region of all the four cp genomes. Comparably, the intact *ycf1* gene in all the cp genomes except *P. argentatum* crosses SSC/IRA region with a 565-583 bp length fragment of *ycf1* located in the IRa region. However, *P. argentatum* encoded two copies of ψ*ycf1* in their genome. The *trnH* gene sequences are found in the LSC region and it is ~0-118 bp away from the IRA/LSC border of all the cp genomes.

The genomic sequences of four Heliantheae species were analyzed by the mVISTA software, detecting the variations of the sequences (Fig. 3). The sequence divergence is not similar to each other sequences. The data plot revealed that the non-coding region was more divergent than its coding counterparts. As compared with LSC and SSC regions, IR regions were less divergent in all the cp genomes.

Repeat structure and SSRs analysis

The existence of repeat sequences in the *X. spinosum* and *X. sibiricum* cp genomes were analyzed and compared. The repeats of the *X. spinosum* cp genome consist of 264 forward, 256 palindromic, 251 reverse and 228 complement. By contrast, a major variant number of repeats was found in *X. sibiricum*, which contained 18 forward, 15 palindromic, six reverse and two complement (Fig. 4a). In total, *X. spinosum* and *X. sibiricum* contains 999 repeats and 41 repeats, respectively. Among 999 repeats in *X. spinosum*, 30-39 bp length repeats (983) predominantly present in their genome and the longest repeat length is 115 bp, which is palindrome sequence. Similarly, in *X. sibiricum* 34 repeats are 30-39 bp and the longest is 177 bp that represents the also palindromic sequence (Fig. 4b).

A total of 701 and 705 simple sequence repeats (SSR) are identified in the *X. spinosum* and *X. sibiricum* cp genomes, respectively. Of 701 SSRs in the *X. spinosum*, 247 (35.24%) were mono-nucleotide repeats, 30 (4.3%) di-nucleotide repeats, 58 (8.3%) tri-nucleotide repeats, 67 (9.6%) tetra-nucleotide repeats, 80 (11.4%) penta-nucleotide repeats, 112 (15.98%) hexa-nucleotide repeats, 37 (4.6%) and 31 (4.42%) 7-nucleotide repeats and other repeat length from 8-nucleotide to 27-nucleotide repeats occupies 10.84% (76 repeats) (Fig. 5a). Similarly, in *X. sibiricum*, 250 (35.46%) were mono-nucleotide repeats, 28 (3.97%) di-nucleotide repeats, 63 (8.94%) tri-nucleotide repeats, 74 (10.5%) tetra-nucleotide repeats, 81 (11.49%) penta-nucleotide repeats, 114 (16.18%) hexa-nucleotide repeats, and 32 (4.54%) 7-nucleotide repeats and other repeat length from 8-nucleotide to 21-nucleotide repeats occupies 8.94% (63 repeats). Furthermore, the distribution of SSRs in the LSC, IR and SSC regions of *X. spinosum* and *X. sibiricum* were evaluated and discovered that the corresponding genome contains 483 and 481 in the LSC, 91 and 93 in the IR and 127 and 131 in the SSC regions (Fig. 5b). Likewise, SSRs also analyzed in the protein-coding and intron and intergenic regions (IGS) of *X. spinosum* and *X. sibiricum* and identified that the respective genome comprises 244 and 252 in the CDs, 69 and 69 in the intron and 388 and 384 in the IGS regions (Fig. 5c).

Nucleotide diversity analysis

The nucleotide diversity of 208 regions was analyzed using DnaSP software, including 79 protein coding genes and 129 intergenic and intron regions among two *Xanthium* cp genomes namely *X. spinosum* and *X. sibiricum*. The results showed that the highest variable region was *infA* (0.034188) among protein-coding genes (Fig. 6a), as were the *trnH-psbA* (0.047739), *psbA-trnK* (0.057143), *trnK* exon2-*matK* (0.092857), *psbI-trnS* (0.046667), *ycf3-trnS* (0.0683376), *trnF-ndhJ* (0.209402), *ndhC-trnV* (0.12551), *trnV* intron

(0.073604), *petD-rpoA* (0.051813), *infA-rps8* (0.181818), *rpl14-rpl16* (0.046729), *rpl16-rps3* (0.032258), *psaC-ndhD* (0.086207) and *trnL-rpl32* (0.080882) among intron and intergenic regions (Fig. 6b; Table 2).

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

Synonymous and nonsynonymous substitution rates were evaluated for 79 protein-coding genes of *X. spinosum* and *X. sibiricum* cp genomes. The K_A/K_S ratio of nearly all the genes is less than 1, except for the protein-coding gene, *accD* (1.56) (Fig. 7).

Positive selection analysis of the accD gene

The positive selection of the *accD* protein-coding gene of Heliantheae cp genome species was investigated. The ω_2 values of the *accD* gene are 3.70208 in the M2a model. Also, Bayes empirical Bayes (BEB) analysis is used to analyze the location of consistent selective sites in the *accD* protein-coding gene using M7 vs. M8 model and identified that one site under potentially positive selection with posterior probabilities more than 0.95 and one sites greater than 0.99 (Table 3) and the 2ΔLnL value is 25.91159 (Table 4).

Phylogenetic analysis

A total of 79 protein-coding genes of 20 cp genome sequences were selected to infer the phylogenetic relationships among the closely related species of Heliantheae and *Achyrachaena mollis* (NC_036504) was selected as an outgroup. The maximum likelihood tree was constructed using the concatenated 79 cp protein-coding genes. The topology of the phylogenetic tree showed that *X. spinosum* has a close relationship with the species of *Ambrosia* (Fig. 8). Although, the analysis showed that *Parthenium* is the sister clade to both *Xanthium* and *Ambrosia* and it is an early-diverging lineage of subtribe Ambrosiinae with weak bootstrap value (54%).

Discussion

The chloroplast genome structure of *X. spinosum* is similar to *X. sibiricum*, which displays a single circular molecule with a typical quadripartite structure. Comparative analysis revealed that these two species exhibited the same gene content and arrangement in the chloroplast genome. Also, these two species contained 37.45% GC content, which was distributed unevenly across the whole chloroplast genome. When compared with LSC and SSC regions, the GC contents in the IR region displayed a higher value across the whole chloroplast genome, 43% in both *X. spinosum* and *X. sibiricum*, possibly resulting from the presence of extremely conserved four rRNA genes with high GC content in the IR regions. Also, the expansion and contraction of the IR regions taken into consideration for the common evolutionary events and are the main cause of variations in the chloroplast genome size, and assessing them could shed some light on the evolution of some taxa [22, 23]. The chloroplast boundary regions of *X. spinosum* were compared with its closely related species and showed that little difference found in their boundary regions. The presence of ψ*ycf1* in the *X. spinosum* crosses IRb/SSC boundary region, whereas in *X. sibiricum* and *A. artemisiifolia*, the ψ*ycf1* is in the IRb region. Though the ψ*ycf1* is shifted to the IR region, the length of the IR regions in the two cp genomes *X. sibiricum* and *A. artemisiifolia* does not affect the length of the IR or SSC regions. In contrast, the ψ*ycf1* gene in the *P. argentatum* moved to SSC a region that leads to the contraction of IR and expansion of SSC regions in their cp genome. But overall, the size of the four cp genomes (*X. spinosum*, *X. sibiricum*, *A. artemisiifolia* and *P. argentatum*) were not affected. Also, the length of each region and the genome size is similar to those of most plant chloroplast genomes reported previously [24].

The repeat units, that are dispersed in chloroplast genomes with high frequency, play a significant role in genome evolution [25–28]. The *X. spinosum* chloroplast consists of 264 direct, 256 palindromic, 251 reverse and 228 complement repeats. In contrast, the *X. sibiricum* encoded the least number of repeats are present in the genome that accounts for a total of 41 repeats such as 18 direct, 15 palindromic, six reverse and two complement repeats. Also, the length of the repeats analyzed. In this analysis, > 30 bp repeat regions were calculated. The *X. spinosum* occupied 983 (98.40%) repeats, which are 30–39 bp length. Similarly, 34 (82.93%) repeats are 30–39 bp length which is present in the *X. sibiricum*. An earlier study reported that the variation in number and variety of repeats play a major role in the plastome organization, but we could not found any correlation between these large repeat regions and rearrangement endpoints [29]. Microsatellite repeats are predominantly present in the chloroplast genome, which displays a high level of polymorphism and used as a genetic marker in genetic investigations [30, 31]. A total of 701 and 705 simple sequence repeats (SSR) were discovered in the *X. spinosum* and *X. sibiricum* cp genomes, respectively. The majority were mononucleotides (~ 35.3%), followed by hexa-nucleotide repeats (~ 16%) and penta-nucleotide repeats (~ 11.45%). Similarly, the intergenic regions

(55.35%) contained a high amount of SSRs than protein-coding (34.8%) and intron (9.85%) regions. Moreover, the content of different types of SSRs and distribution on various chloroplast regions were almost similar in both *X. spinosum* and *X. sibiricum* species.

Though the chloroplast genome of *X. spinosum* occupies similar quadripartite structure and gene contents as *X. sibiricum*, we identified few variations in the non-coding regions than its coding counterparts. The LSC region exhibited higher divergence levels than the IR and SSC regions (Fig. 6c). Furthermore, the two IR regions were less divergent than the LSC and SSC regions, which might be due to the presence of highly conserved four ribosomal RNAs located in their regions. In addition, the nucleotide diversity of 79 protein-coding genes and 129 intergenic and intron regions of two *Xanthium* cp genomes were analyzed. The results revealed that intergenic regions were more divergent than protein-coding regions. The average nucleotide diversity (π) in the intergenic regions was 0.0170, almost four times as that in protein-coding genes (Pi = 0.004195). The results also support that intergenic regions showed greater divergence than protein-coding regions (Fig. 6D).

Not all protein-coding genes are phylogenetically helpful in determining taxonomic discrepancies [32]. Tomasello and Heubel (2017) used chloroplast genome sequences of plastid *psbA-trnH* and *trnQ-rps16* and nuclear ITS and D35 to resolve the phylogenetics of *Xanthium* species. Similarly, the ITS2 sequence also used as a DNA barcode to reveal seven *Xanthium* species [33]. Although, previous studies suggested that the use of the additional markers and a broader taxon sampling are required to increase the phylogenetic resolution at low taxonomic levels. So, in the present study, we suggest a set of fourteen divergence regions by comparing nucleotide diversity among regions between *X. spinosum* and *X. sibiricum* to solve taxonomic discrepancies and provision of barcode for the genus *Xanthium*. All the regions belong to intergenic regions, and these might be useful for the development of molecular markers for phylogenetic and phylogeographic studies. Although, the fourteen sequences identified in the present study had nucleotide diversity 0.181818 to 0.032258 from highest to lowest. So, the sequences discovered in the present study are extremely polymorphic when compared to the sequences that were used in the earlier studies [1, 3, 33]. Consequently, based on data described in the present study, the robust and authentic molecular markers can be developed for these intergenic regions and can be used for the phylogenetic, phylogeographic and barcoding of *Xanthium*. Moreover, this is the first report to develop authentic markers using these regions and could be used for the deep divergence in the genus of *Xanthium*.

In the present study, we also analyzed the ratio of K_A/K_S (ω) value of 79 protein-coding genes of *X. spinosum* and *X. sibiricum* species. The analysis revealed that the ω value of *accD* is 1.56 and suggested that this gene could be under positive selection. So, we further investigated a selective analysis of the *accD* protein-coding gene using site-specific models with four comparison models (M0 vs. M3, M1 vs. M2a, M7 vs. M8, M8a vs. M8, likelihood ratio test (LRT) threshold $p \leq 0.05$ in EasyCodeML software. Among these models, M2a is the positive selective model and p (p_0, p_1 and p_2) are the proportions of negative or purifying, neutral and positive selection respectively. The ω_2 value of the *accD* gene is 3.70208 in the M2a model (Table 2). Also, to determine which sites are subject to positive selection, Bayes empirical Bayes (BEB) analysis is used to analyze the location of consistent selective sites in the *accD* gene of *X. spinosum* and other Heliantheae cp genomes using M7 vs. M8 model. The analysis of BEB revealed that one site under potentially positive selection in the *accD* gene with posterior probabilities more than 0.95 and one sites higher than 0.99 (Table 3). The 2ΔLnL value of the *accD* gene is 25.91159 with a *p*-value of LRT is 0.0000024. The positively selected sites detected in the present study may drive the *accD* protein-coding gene allowing the occupation of diverse habitats [34, 35].

In the past few years, more plastid genome database offers an important foundation for the resolve of the evolutionary, taxonomic and phylogenetic studies of plants [36–42]. The phylogenetic analysis in the present study showed that the genus *Xanthium* was most closely related to the genus *Ambrosia*. Several studies have been carried out with different methods such as cladistic analysis [43, 44], cp restriction site variation [45], sequence analysis [3, 46] to understand the position of *Xanthium* and identified that it is more closely related with *Ambrosia* species. Also, the phylogenetic study showed that the genus *Parthenium* presented the early-diverging lineage in the subtribe Ambrosiinae with very weak bootstrap value (54%). Though. to support this analysis, Kumar et al., (2009) and Tomasello and Heubl (2017) also identified the same results using three plastid and two nuclear marker regions. In contrast to the most recent study, Somaratne et al., (2019) suggest that *Parthenium* is not as the early-diverging lineage of subtribe Ambrosiinae and it clustered with Helianthus. Moreover, Somaratne et al., (2019) used only 46 cp protein-coding genes for phylogenetic analyses, whereas we used 79 protein-coding genes in this study. Due to the limited cp protein-coding genes used for the phylogenetic analysis by Somaratne et al., (2019), they obtained inconsistent phylogenetic results. Hence, we suggest that *Parthenium* is an early-diverging lineage of subtribe Ambrosiinae based on 79 protein-coding analysis. Therefore, extend sampling

and additional nuclear marker genes will provide more precise information for the phylogenetic position of *Parthenium* than a few cp protein-coding genes.

Conclusion

In the present study, the complete chloroplast genome sequence of the *Xanthium spinosum* was assembled and analyzed. Overall, the gene contents and gene arrangements are similar and highly conserved in the *Xanthium* species. Also, the comparative analysis such as repeat sequence analysis and nucleotide substitution patterns were carried out to understand their genomic patterns and their evolution in their cp genomes. Additionally, a comparison of *X. spinosum* with its closely related species *X. sibiricum*, several highly divergent noncoding regions were identified that would be useful for the developing high-resolution molecular markers. Furthermore, the *accD* gene was found to be undergone positive selection, which might be the result of adaptation to the environment. Finally, phylogenetic relationship analysis revealed the genus *Xanthium* and *Ambrosia* are the sister clade to each other, consistent with earlier studies and *Parthenium* is an early-diverging lineage of subtribe Ambrosiinae with weak bootstrap value.

Methods

DNA extraction and sequencing of *Xanthiumspinosum*

The *Xanthium spinosum* plant leaf material was obtained from the George A Yatskievych, Curator, Plant Resources Center, University of Texas Herbarium (19-056), Austin, USA. Total genomic DNA was extracted using a modified cetyltrimethylammonium bromide method [47]. The Illumina sequencing was carried out by LabGenomics, Seongnam, South Korea using the Illumina HiSeq 2500 sequencing system. The pair-end library (150x2) was constructed with an insert size of 350 base pairs (bp). Read quality was analyzed with FastQC [48] and low-quality reads were removed with Trimmomatic 0.39 [49]. The resultant clean reads were filtered with GetOrganelle pipe-line (https://github.com/Kinggerm/GetOrganelle) to get plastid-like reads, then the filtered reads were assembled by *de nova* approach using SPAdes version 3.12.0 [50]. The complete chloroplast genome sequence of *X. spinosum* and their gene annotation were submitted to GenBank (MT668935).

Annotation of *X. spinosum* chloroplast genome

The online program Dual Organeller GenoMe Annotator (DOGMA) was performed to annotate the chloroplast genome sequence of *X. spinosum* [51]. The initial annotation, putative starts, stops, and intron positions were changed by comparing with closely related species *X. sibiricum* homologous genes [1]. Transfer RNA genes were validated using tRNAscan-SE version1.21 with default settings [52]. OGDRAW program was employed to draw a circular map of the *X. spinosum* cp genome [53].

Comparative chloroplast genome analysis

The mVISTA program in the Shuffle-LAGAN model was employed to evaluate the cp genome of *X. spinosum* with closely related three other cp genomes such as *X. sibiricum*, *Ambrosia artemisiifolia* and *Parthenium argentatum* using *X. spinosum* annotation as a reference [54]. The boundaries between IR and SC regions of these species were also compared and investigated.

Repeat sequences and single sequence repeats (SSR) analysis

The program REPuter was used to predict the presence of repeat sequences in the *X. spinosum* and *X. sibiricum* cp genomes, comprising forward, reverse, palindromic, and complementary repeats [55]. The following parameters 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. Also, Phobos software v1.0.6 was applied to discover SSRs in the *X. spinosum* and *X. sibiricum* cp genomes; parameters for the match, mismatch, gap, and N positions were set at 1, -5, -5 and 0, respectively [56]. In the repeat and SSR marker analysis, only one IR region was used.

Analysis of the genetic divergence

To analyze genetic divergence, the protein-coding genes, intergenic and intron-containing regions of *X. spinosum* and *X. sibiricum* cp genome was extracted and aligned independently using Geneious Prime (Biomatters, New Zealand). Genetic divergence between

Xanthium species was calculated using nucleotide diversity (π) and the total number of polymorphic sites by DnaSP. In this analysis, gaps and missing data were excluded.

Characterization of substitution rates

To calculate synonymous (K_S) and nonsynonymous (K_A) substitution rates, the cp genome of *X. spinosum* was compared with the *X. sibiricum* cp genome. The related single functional protein-coding gene exons of these genomes were extracted and aligned independently using Geneious Prime (Biomatters, New Zealand). The aligned sequences were translated into protein sequences and analyzed using DnaSP for K_A and K_S substitution rates without stop codon [57].

Positive selection analysis

The positive selection model (M2a and M8) and the control model (M1a, M7 and M8a) provided by EasyCodeML software were used to determine the existence of positive selection (ω > 1) in the *accD* locus of Heliantheae cp genomes. The sequence of the *accD* gene was aligned using the MAFFT program [58] and the maximum likelihood phylogenetic tree was constructed using RAxML v. 7.2.6 [59]. The site-specific model was performed to calculate nonsynonymous (K_A) and synonymous substitution (K_S) ratio using EasyCodeML [60]. The codon substitution models M0, M1a, M2a, M3, M7, M8 and M8a were studied. The likelihood ratio test was conducted to identify the positively selected sites: M0 (one-ratio) vs. M3 (discrete), M1a (neutral) vs. M2a (positive selection) and M7 (β) vs. M8 (β and ω > 1) and M8a ((β and ω = 1) vs. M8, which were compared using a site-specific model [60]. The likelihood ratio test (LRT) of the above comparison has been conducted respectively to evaluate the selection strength and the *p*-values of Chi-square (x^2) lesser than 0.05 were considered as significant. If the LRT *p*-values were significant (<0.05), Bayes Empirical Bayes (BEB) method was implemented to identify codons under positive selection. The BEB values higher than 0.95 and 0.99 indicate the sites potentially under positive selection and highly positive selection, which is indicated by asterisks and double asterisks, respectively.

Phylogenetic tree analysis

A phylogenetic tree was formed using 71 protein-coding genes of 20 Heliantheae cp genomes and *Achyrachaena mollis* used as the outgroup. The 19 completed cp genome sequences were downloaded from the NCBI Organelle Genome Resource database. The aligned protein-coding gene sequences were saved in PHYLIP format using Clustal X v2.1 [61] and phylogenetic analysis was constructed based on maximum likelihood (ML) analysis using the GTRI model by RAxML v. 7.2.6 with 1000 bootstrap replications [59].

Abbreviations

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

Declarations

Ethics approval and consent to participate

Our study does not involve ethics approval and consent to participate.

Consent to publish

All authors read and approved the final manuscript.

Availability of data and materials

All the data involved in this article is true and reliable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by a funded project (PQ20180B009) entitled “Development of molecular markers for putative invasive alien” by Research of Animal and Plant Quarantine Agency, South Korea.

Authors contributions

SJP, JHK and GR conceived and designed the experiments. KP performed the experiments. GR analyzed the data and prepared a draft of the manuscript and figures. SJP, JHK and GR modified the manuscript. All authors read and approved the final manuscript.

Acknowledgment

Not applicable

References

1. Somaratne Y, Guan D-L, Wang W-Q, Zhao L, Xu S-Q: **Complete chloroplast genome sequence of *Xanthium sibiricum* provides useful DNA barcodes for future species identification and phylogeny**. *Plant Systematics and Evolution* 2019, **305**(10):949-960.
2. Amin S, Khan H: **Revival of Natural Products: Utilization of Modern Technologies**. *Current Bioactive Compounds* 2016, **12**(2):103-106.
3. Tomasello S, Heubl G: **Phylogenetic Analysis and Molecular Characterization of *Xanthium sibiricum* Using DNA Barcoding, PCR-RFLP, and Specific Primers**. *Planta Med* 2017, **83**(11):946-953.
4. Romero M, Zanuy M, Rosell E, Cascante M, Piulats J, Font-Bardia M, Balzarini J, De Clerq E, Pujol MD: **Optimization of xanthatin extraction from *Xanthium spinosum* L. and its cytotoxic, anti-angiogenesis and antiviral properties**. *Eur J Med Chem* 2015, **90**:491-496.
5. Salinas A, Ruiz RELd, Ruiz SO: **Sterols, flavonoids, and sesquiterpenic lactones from *Xanthium spinosum* (Asteraceae)**. *Acta Farm Bonaerense* 1998, **17**(4):297-300.
6. Ginesta-Peris E, Garcia-Breijo FJ, Primo-Yufera E: **Antimicrobial activity of xanthatin from *Xanthium spinosum* L.***Letters in Applied Microbiology* 1993, **18**:206-208.
7. Yoon JH, Lim HJ, Lee HJ, Kim HD, Jeon R, Ryu JH: **Inhibition of lipopolysaccharide-induced inducible nitric oxide synthase and cyclooxygenase-2 expression by xanthanolides isolated from *Xanthium strumarium***. *Bioorg Med Chem Lett* 2008, **18**(6):2179-2182.
8. Willians RH, Martin FB, Henley ED, Swanson HE: **Inhibitors of insulin degradation, metabolism**. *Clinical and Experimental Rheumatology* 1959, **8**:99-113.
9. Naidenova E, Kolarova-Pallova I, Popov D, Dimitrova-Konaklieva S, Dryanovska-Noninska L: **Isolation and obtaining of sesquiterpene lactones with antitumor properties – xanthinin, stizolicin, and solstitialin**. *National Oncology Center Medical Academy* 1988, **41**:105-106.
10. Cunat P, Primo E, Sanz I, Garcera MD, March MC, Bowers WS, Martinez-Pardo R: **Biocidal activity of some Spanish Mediterranean plants**. *Journal of Agricultural and Food Chemistry* 1990, **38**(2):497-500.
11. Turnne CY, Sanche SE, Hoban DJ, Karylowsky JA, Kabani AM: **Rapid Identification of Fungi by Using the ITS2 Genetic Region and an Automated Fluorescent Capillary Electrophoresis System**. *Journal of Clinical Microbiology* 1999, **37**(6):1846-1851.
12. Trobajo R, Mann DG, Clavero E, Evans KM, Vanormelingen P, McGregor RC: **The use of partialcox1, *rbcL* and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae)**. *European Journal of Phycology* 2010, **45**(4):413-425.
13. Liu C, Liang D, Gao T, Pang X, Song J, Yao H, Han J, Liu Z, Guan X, Jiang K *et al*: **PTIGS-IdIt, a system for species identification by DNA sequences of the *psbA-trnH* intergenic spacer region**. *BMC Bioinformatics* 2011, **12**:S4.
14. Millspaugh C, Sherff E: **Revision of the North American species of *Xanthium*.***Field Museum of Natural History Zoological series* 1919, **4**:9-49.
15. Widder F: **Die Arten der Gattung *Xanthium*. Beiträge zu einer Monographie.***Repert Spec Nov Regn Veget* 1923, **20**:1-223.

16. Löve D, Dansereau L: **Biosystematic studies on *Xanthium*: taxonomic appraisal and ecological status**. *Canadian Journal of Botany* 1959, **37**:173-208.
17. Strother J: ***Xanthium*. In: Flora of North America**, vol. 21. New York: Oxford University; 2006.
18. Hollingsworth PM, Graham SW, Little DP: **Choosing and Using a Plant DNA Barcode**. *Plos One* 2011, **6**(5).
19. Sanita Lima M, Woods LC, Cartwright MW, Smith DR: **The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes**. *Mol Ecol Resour* 2016, **16**(6):1279-1286.
20. Kumar S, Hahn FM, McMahan CM, Cornish K, Whalen MC: **Comparative analysis of the complete sequence of the plastid genome of *Partheniumargentatum* and identification of DNA barcodes to differentiate *Parthenium* species and lines**. *Bmc Plant Biol* 2009, **9**.
21. Amiryousefi A, Hyvonen J, Poczai P: **The plastid genome sequence of the invasive plant common Ragweed (*Ambrosia artemisiifolia*, Asteraceae)**. *Mitochondrial DNA B* 2017, **2**(2):753-754.
22. Boudreau E, Turmel M: **Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat**. *Plant Mol Biol* 1995, **27**(2):351-364.
23. Nazareno AG, Carlsen M, Lohmann LG: **Complete Chloroplast Genome of *Tanaecium tetragonolobum*: The First Bignoniaceae Plastome**. *Plos One* 2015, **10**(6).
24. Zhang X, Zhou T, Kanwal N, Zhao YM, Bai GQ, Zhao GF: **Completion of Eight *Gynostemma* BL. (Cucurbitaceae) Chloroplast Genomes: Characterization, Comparative Analysis, and Phylogenetic Relationships**. *Front Plant Sci* 2017, **8**.
25. Dong WP, Xu C, Cheng T, Lin K, Zhou SL: **Sequencing Angiosperm Plastid Genomes Made Easy: A Complete Set of Universal Primers and a Case Study on the Phylogeny of Saxifragales**. *Genome Biol Evol* 2013, **5**(5):989-997.
26. Liu WZ, Kong HH, Zhou J, Fritsch PW, Hao G, Gong W: **Complete Chloroplast Genome of *Cercis chuniana* (Fabaceae) with Structural and Genetic Comparison to Six Species in Caesalpinioideae**. *Int J Mol Sci* 2018, **19**(5).
27. Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD, He XJ: **Comparative Analysis of the Chloroplast Genomes of the Chinese Endemic Genus *Urophysa* and Their Contribution to Chloroplast Phylogeny and Adaptive Evolution**. *Int J Mol Sci* 2018, **19**(7).
28. Shi HW, Yang M, Mo CM, Xie WJ, Liu C, Wu B, Ma XJ: **Complete chloroplast genomes of two *Siraitia Merrill* species: Comparative analysis, positive selection and novel molecular marker development**. *Plos One* 2019, **14**(12).
29. Liu Y, Huo NX, Dong LL, Wang Y, Zhang SX, Young HA, Feng XX, Gu YQ: **Complete Chloroplast Genome Sequences of Mongolia Medicine *Artemisia frigida* and Phylogenetic Relationships with Other Plants**. *Plos One* 2013, **8**(2).
30. Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J: **An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree**. *Theoritical applied Genetics* 1997, **95**:163-173.
31. Kawabe A, Nukii H, Furihata HY: **Exploring the History of Chloroplast Capture in *Arabis* Using Whole Chloroplast Genome Sequencing**. *Int J Mol Sci* 2018, **19**(2).
32. Pfeil BE, Brubaker CL, Craven LA, Crisp MD: **Phylogeny of *Hibiscus* and the tribe Hibisceae (Malvaceae) using chloroplast DNA sequences of *ndhF* and the *rpl16* intron**. *Syst Bot* 2002, **27**(2):333-350.
33. Wallace LJ, Boilard SMAL, Eagle SHC, Spall JL, Shokralla S, Hajibabaei M: **DNA barcodes for everyday life: Routine authentication of Natural Health Products**. *Food Res Int* 2012, **49**(1):446-452.
34. Hu QM, Zhu Y, Liu Y, Wang N, Chen SL: **Cloning and characterization of *wnt4a* gene and evidence for positive selection in half-smooth tongue sole (*Cynoglossussemilaevis*)**. *Sci Rep-Uk* 2014, **4**.
35. Dhar D, Dey D, Basu S, Fortunato H: **Understanding the adaptive evolution of mitochondrial genomes in intertidal chitons**. *bioRxiv* 2020:2020.2003.2006.980664.
36. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD: **The chloroplast *ycf3* and *ycf4* open reading frames of *Chlamydomonasreinhardtii* are required for the accumulation of the photosystem I complex**. *EMBO J* 1997, **16**(20):6095-6104.
37. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ: **Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses**. *Bmc Evol Biol* 2013, **13**.
38. Raman G, Park S: **The Complete Chloroplast Genome Sequence of *Ampelopsis*: Gene Organization, Comparative Analysis, and Phylogenetic Relationships to Other Angiosperms**. *Front Plant Sci* 2016, **7**.

39. Zhang YJ, Du LW, Liu A, Chen JJ, Wu L, Hu WM, Zhang W, Kim K, Lee SC, Yang TJ *et al*: **The Complete Chloroplast Genome Sequences of Five *Epimedium* Species: Lights into Phylogenetic and Taxonomic Analyses**. *Front Plant Sci* 2016, **7**.
40. Kahraman K, Lucas SJ: **Comparison of different annotation tools for characterization of the complete chloroplast genome of *Corylusavellana* cv *Tombul***. *Bmc Genomics* 2019, **20**(1).
41. Li XQ, Zuo YJ, Zhu XX, Liao S, Ma JS: **Complete Chloroplast Genomes and Comparative Analysis of Sequences Evolution among Seven *Aristolochia* (Aristolochiaceae) Medicinal Species**. *Int J Mol Sci* 2019, **20**(5).
42. Li CJ, Wang RN, Li DZ: **Comparative analysis of plastid genomes within the Campanulaceae and phylogenetic implications**. *Plos One* 2020, **15**(5).
43. Karis P: **Heliantheae sensu lato (Asteraceae), clades and classification**. *Plant Systematics and Evolution* 1993, **188**:139-195.
44. Karis P: **Cladistics of the subtribe Ambrosiinae (Asteraceae: Heliantheae)**. *Syst Bot* 1995, **20**:40-54.
45. Miao B, Turner B, Mabry T: **Systematic implications of chloroplast DNA variation in the subtribe Ambrosiinae (Asteraceae: Heliantheae)**. *American Journal of Botany* 1995, **82**:924-932.
46. Martin MD, Quiroz-Claros E, Brush GS, Zimmer EA: **Herbarium collection-based phylogenetics of the ragweeds (*Ambrosia*, Asteraceae)**. *Mol Phylogenet Evol* 2018, **120**:335-341.
47. Doyle J: **Isolation of plant DNA from fresh tissue**. *Focus* 1990, **12**:13-15.
48. Andrews S: **FASTQC. A quality control tool for high throughput sequence data**. 2010.
49. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
50. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y *et al*: **Assembling single-cell genomes and mini-metagenomes from chimeric MDA products**. *J Comput Biol* 2013, **20**(10):714-737.
51. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA**. *Bioinformatics* 2004, **20**(17):3252-3255.
52. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W686-689.
53. Lohse M, Drechsel O, Bock R: **OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes**. *Curr Genet* 2007, **52**(5-6):267-274.
54. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W273-279.
55. 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**(22):4633-4642.
56. Mayer C, Leese F, Tollrian R: **Genome-wide analysis of tandem repeats in Daphnia pulex–a comparative approach**. *BMC Genomics* 2010, **11**:277.
57. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data**. *Bioinformatics* 2009, **25**(11):1451-1452.
58. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.
59. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML Web servers**. *Syst Biol* 2008, **57**(5):758-771.
60. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW: **EasyCodeML: A visual tool for analysis of selection using CodeML**. *Ecology and Evolution* 2019, **9**:3891-3898.
61. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

Tables

Table 1
Comparative analysis of chloroplast genome of *X.spinosum* and *X. sibiricum*

Characteristics		***X. spinosum***	***X. sibiricum***
Size (bp)		152,422	151,897
LSC length (bp)		84,189	83,847
SSC length (bp)		18,083	17,900
IR length (bp)		25,075	25,070
Total number of genes		132	132
Protein-coding genes		87	87
tRNA genes		37	37
rRNA genes		8	8
Duplicate genes		17	17
GC content	Total (%)	37.4	37.5
	LSC (%)	35.4	35.5
	SSC (%)	31.2	31.4
	IR (%)	43	43
	CDS (%)	37.9	37.9
	rRNA (%)	55.2	55.2
	tRNA (%)	53	52.9
	All genes (%)	39.5	39.5
Protein-coding genes (%bp)		51.65	51.74
All genes (%bp)		72.5	72.89
Non-coding regions (%)		27.5	27.11

Table 2
Mutational hotspots between *X. spinosum* and *X. sibiricum*

S. No.	Region	Nucleotide diversity (Pi)	Total number of mutations	Region length (bp)
1	*trnH - psbA*	0.04774	19	398
2	*psbA - trnK*	0.05714	12	210
3	*trnK* exon 2 - *matK*	0.09286	26	280
4	*psbI - trnS*	0.04667	7	150
5	*ycf3 - trnS*	0.06838	56	819
6	*trnF - ndhJ*	0.20940	49	234
7	*ndhC - trnV*	0.12551	123	980
8	*trnV* intron	0.07360	29	394
9	*petD - rpoA*	0.05181	10	193
10	*infA - rps8*	0.18189	22	121
11	*rpl14 - rpl16*	0.04673	5	107
12	*rpl16 - rps3*	0.03226	5	155
13	*psaC - ndhD*	0.08621	10	116
14	*trnL - rpl32*	0.08088	44	544

Table 3
Comparison of site models, positive selective amino acid loci and estimation of parameters for *accD* protein-coding genes in the Heliantheae species.

Protein-coding gene	Model	np	Ln L	Estimates of parameters				Model compared	LRT P-value	Positive sites
accD	M3	45	-1567.812	p:	0.05791	0.85997	0.08211	M0 vs. M3	0	[]
				ω:	0.22724	0.22724	3.70209			
	M0	41	-1597.602	ω0:	0.36846					Not Allowed
	M2a	44	-1567.812	p:	0.91789	0	0.08211	M1a vs. M2a	0.000196255	[]
				ω:	0.22724	1	3.70208			
	M1a	42	-1576.349	p:	0.82957	0.17043				Not Allowed
				ω:	0.11911	1				
	M8	44	-1567.829	p0 = 0.91839	p = 29.41052	q = 99.00000		M7 vs.M8	0.000002362	35 S 0.936, 42 R 0.963*, 98 H 0.701, 129 Q 0.685, 177 S 0.616, 181 S 0.765, 184 N 0.519, 185 A 1.000**, 187 A 0.905
				(p1 = 0.08161)	ω = 3.72034					
	M7	42	-1580.784	p=	0.07226	q=	0.08717			Not Allowed
	M8a	43	-1576.468	p0 = 0.82999	p = 13.67839	q = 99.00000		M8a vs.M8	0.000032257	Not Allowed
				(p1 = 0.17001)	ω = 1.00000					

Note:

[] – No data available

np represents the degree of freedom

Positively selected sites (* $p > 95\%$; ** $p > 99\%$)

Table 4
Comparison of likelihood ratio test (LRT) statistics of positive selection models against their null models (2ΔLnL) for the *accD* gene.

Protein-coding genes	Comparison between models	2ΔLnL	*d.f.*	*p*-value
accD	M0 vs M3	59.579136	4	0
	M1 vs M2A	17.07219	2	0.0001963
	M7 vs M8	25.91159	2	0.0000024
	M8a vs M8	17.27995	1	0.0000323

Figures

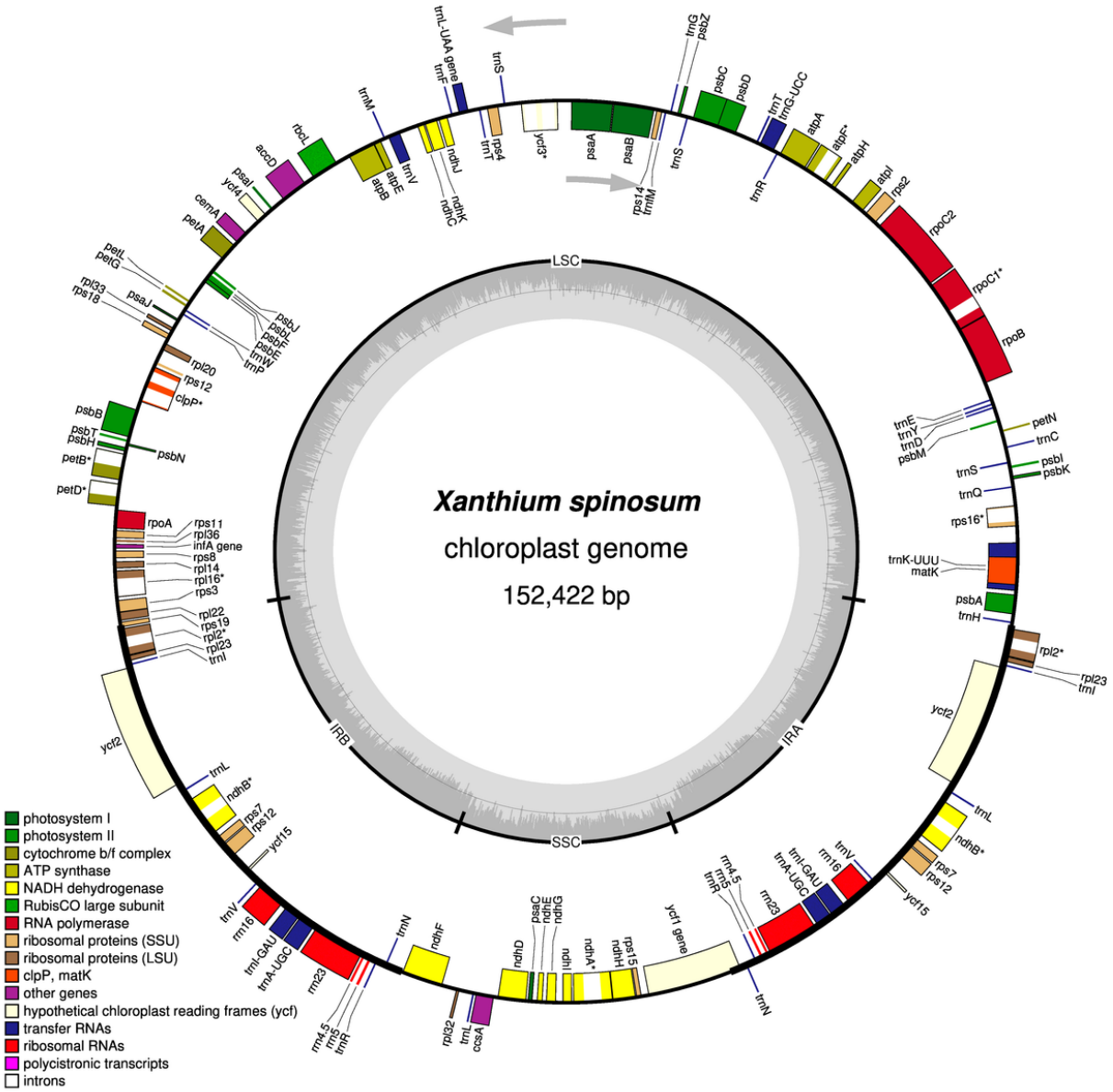

Fig. 1

Figure 1

Gene map of Xanthium spinosum. Genes lying outside the outer circle are transcribed in a counter-clockwise direction, and genes inside this circle are transcribed in a clockwise direction. The coloured bars indicate known protein-coding genes, transfer RNA

genes, and ribosomal RNA genes. The dashed, dark grey area in the inner circle denotes GC content, and the light grey area indicates genome AT content. LSC, large single-copy; SSC, small single-copy; IR, inverted repeat.

Fig. 2

Figure 2

Comparison of the large single-copy (LSC), small single-copy (SSC) and inverted repeat (IR) border regions of four Heliantheae (Xanthium spinosum, X. sibiricum, Ambrosia artemisiifolia and Parthenium argentatum) chloroplast genomes. Ψ indicates a pseudogene. The figure is not drawn to scale.

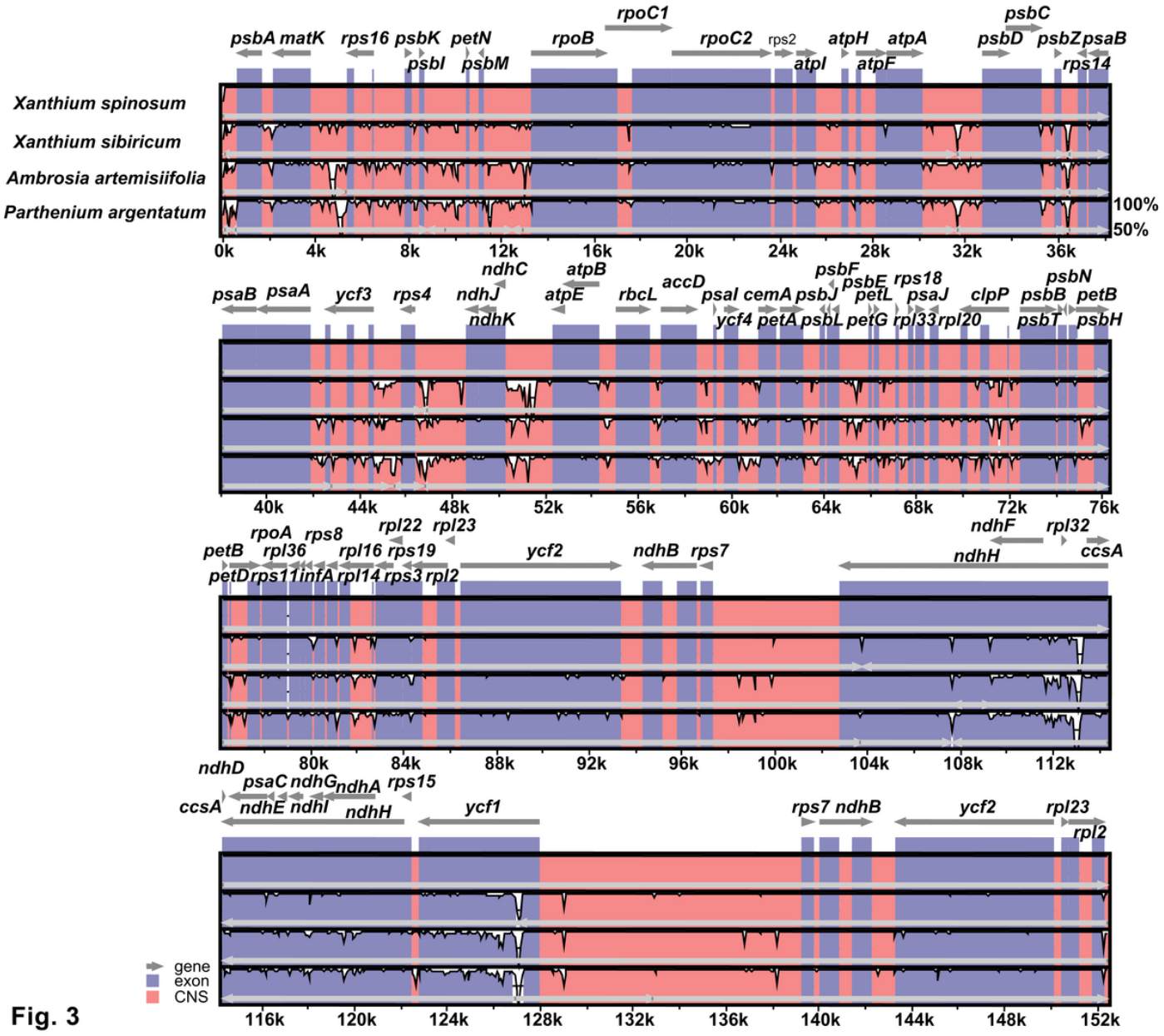

Fig. 3

Figure 3

Sequence alignment of four Heliantheae chloroplast genomes performed using the mVISTA program with Xanthium spinosum as a reference. The top grey arrow shows genes in order (transcriptional direction) and the position of each gene. A 70% cut-off was used for the plots. The y-axis indicates a percent identity of between 50 and 100%, and the red and blue areas indicate intergenic and genic regions, respectively.

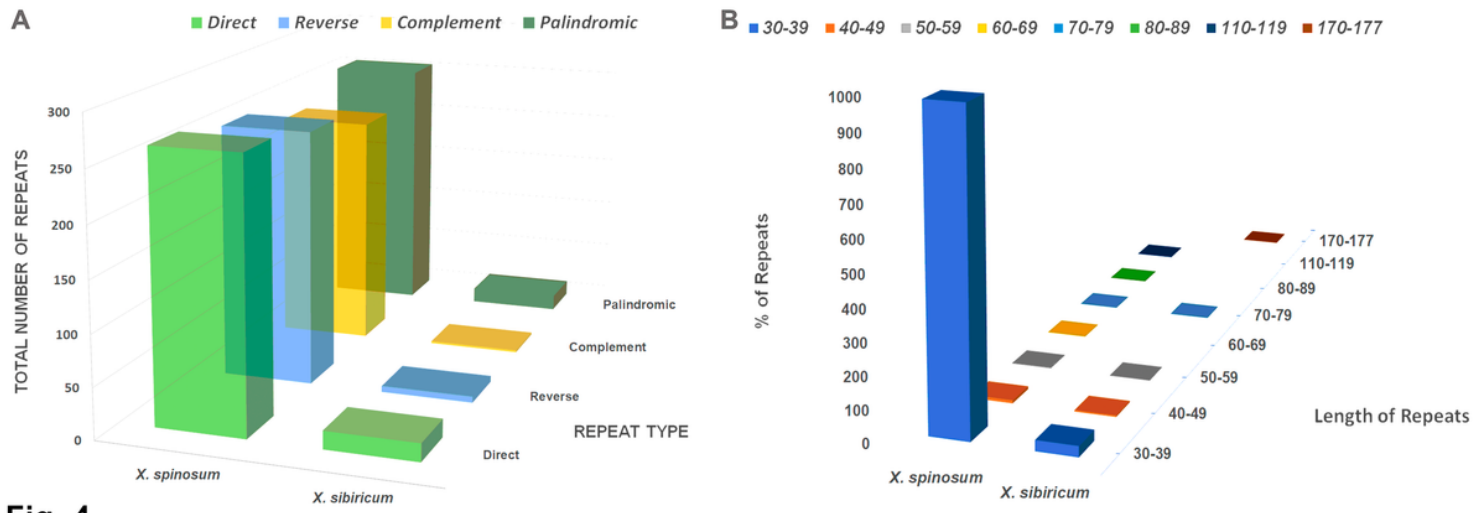

Fig. 4

Figure 4

Comparison he distribution of different repeat types in the Xanthium spinosum vs. X. sibiricum cp genomes. (A) The number of different types of repeats. F – forward repeats; R – Reverse repeats; P – palindromic repeats; C – complement repeats. (B) The length and the total number of repeat sequences present in their respective cp genomes.

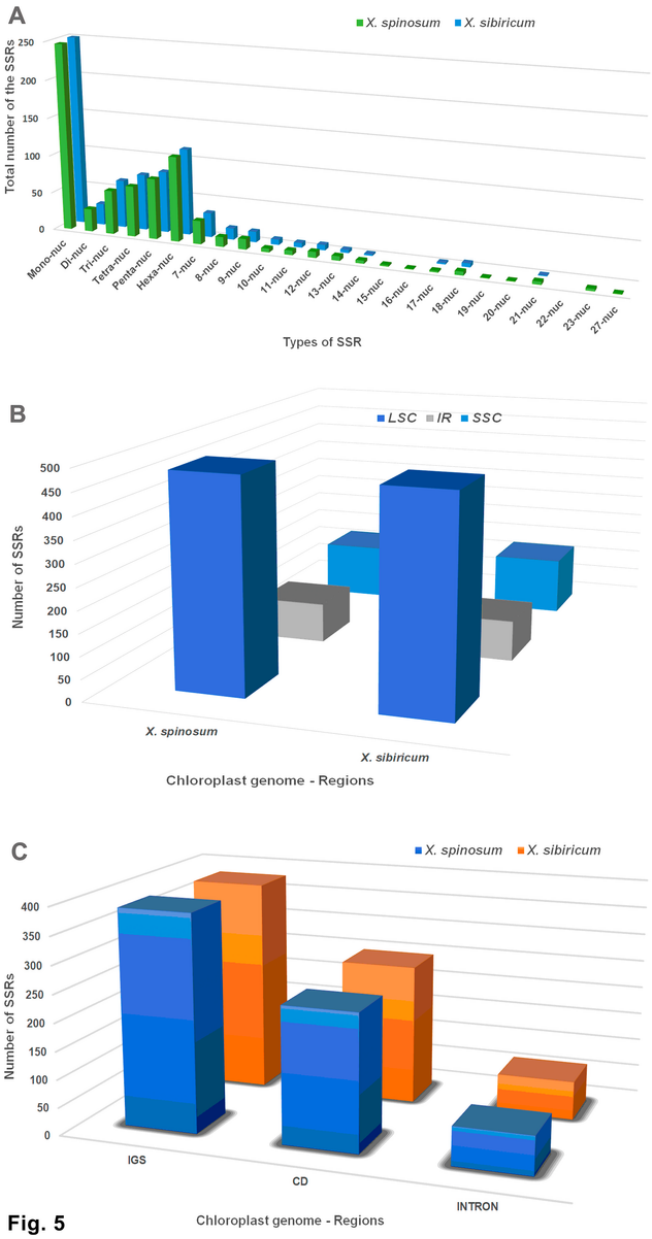

Fig. 5

Figure 5

Comparison the presence of simple sequence repeats (SSRs) in the Xanthium spinosum vs. X. sibiricum cp genomes. (A) Distribution of different types of SSRs. (B) Presence of SSRs in the LSC, SSC, and IR regions. (C) Presence of SSRs in intergenic spacers, protein-coding regions, and intron regions.

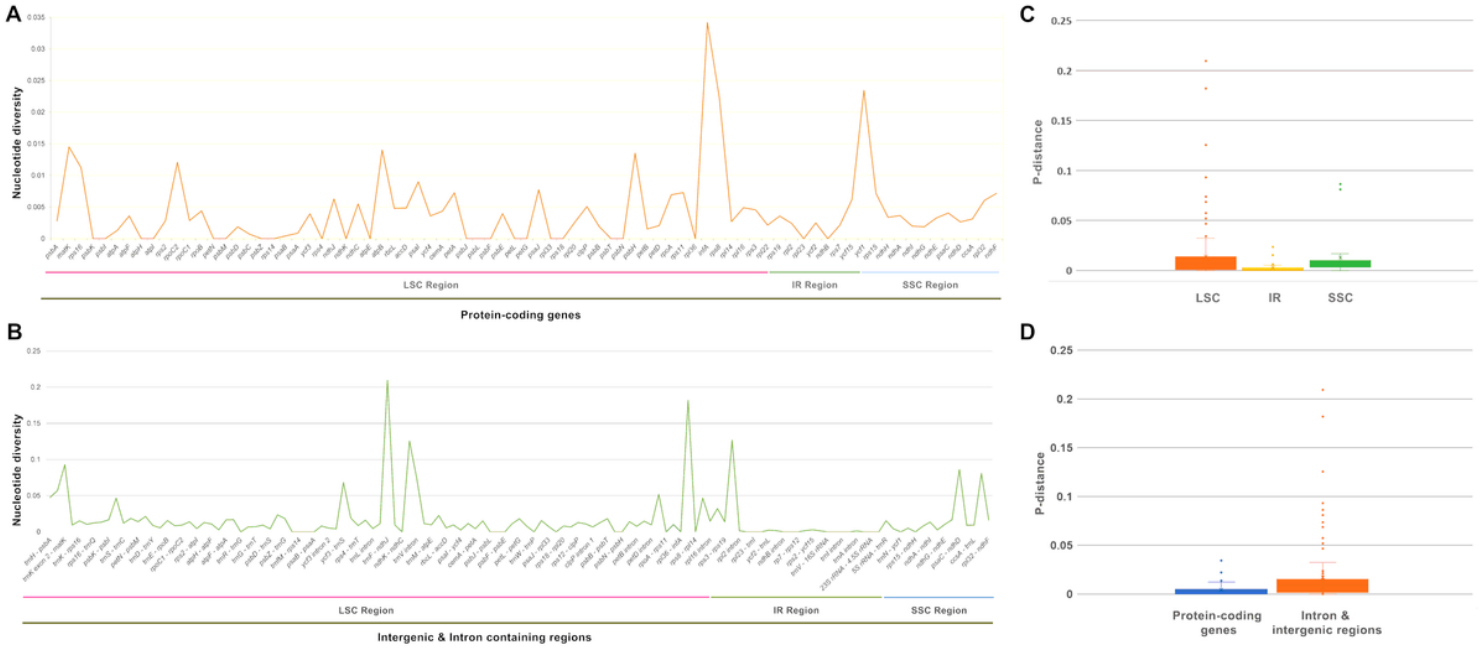

Fig. 6

Figure 6

The genetic diversity based on Kimura's two-parameter model. (A) The P-distance value of protein-coding genes (B) the P-distance value of intron and intergenic regions (C) Boxplots of P-distance value difference among LSC, IR and SSC regions (D) Boxplots of P-distance value differences between protein-coding genes and intron and intergenic regions.

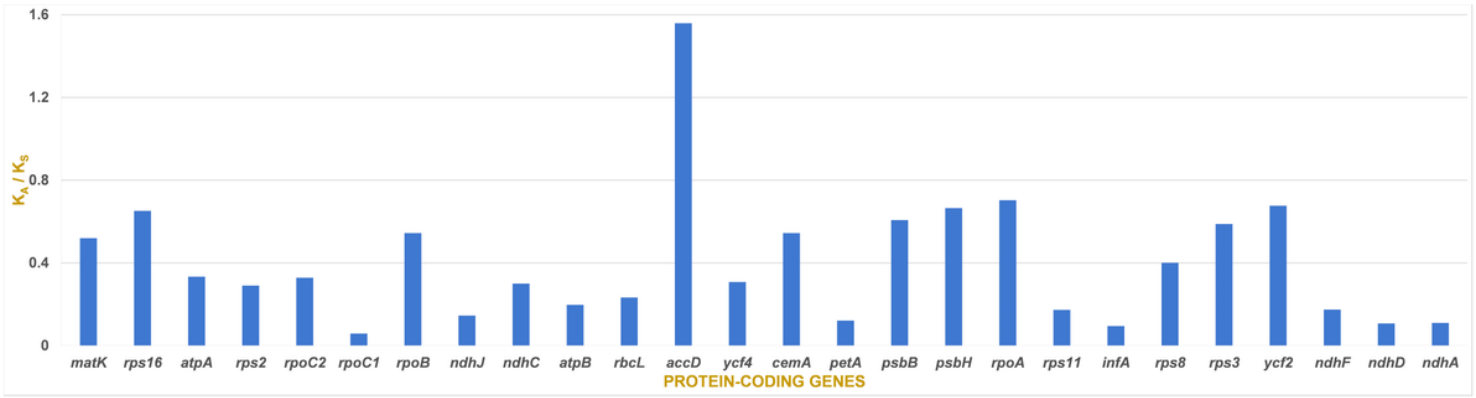

Fig. 7

Figure 7

Comparison the ratio of non-synonymous (KA) to synonymous (KS) substitutions of 79 protein-coding genes of Xanthium spinosum vs. X. sibiricum cp genomes

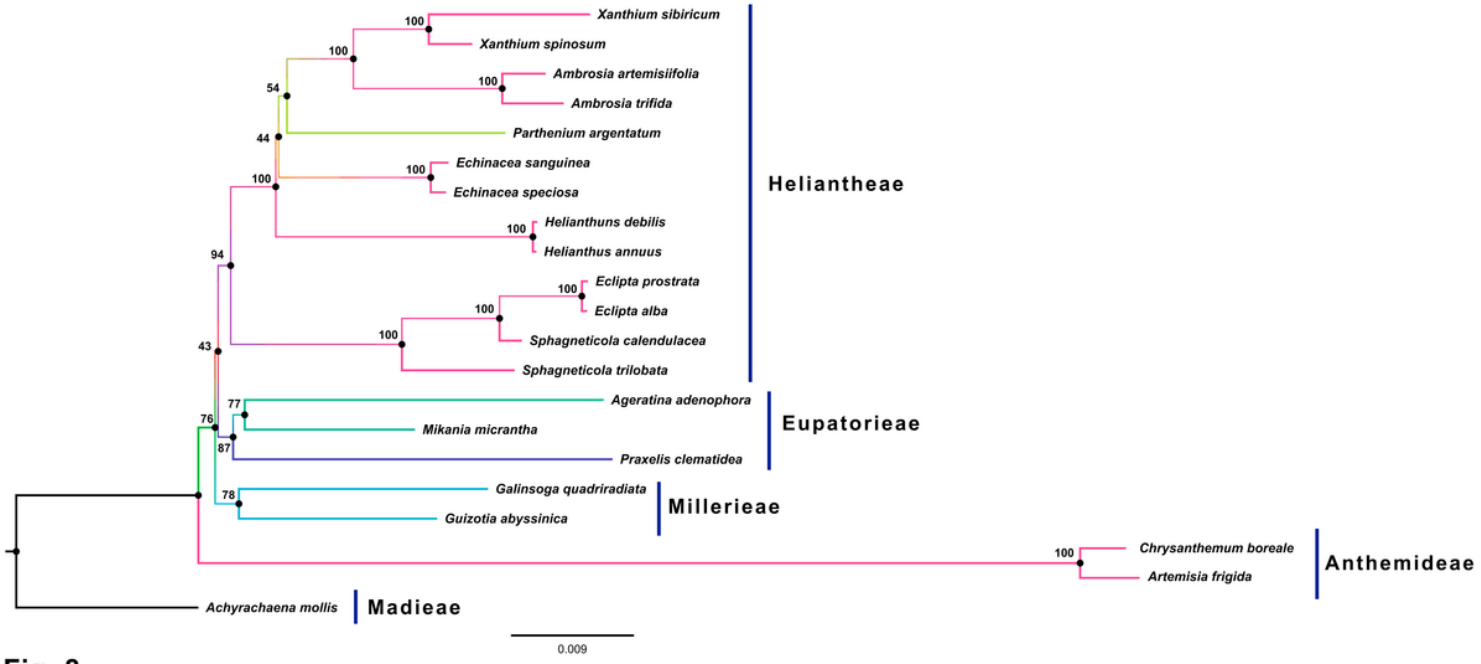

Fig. 8

Figure 8

Molecular phylogenetic tree based on 79 protein-coding genes of 21 Asteraceae chloroplast genomes. Achyrachaena mollis set as the outgroup. The tree was constructed by maximum likelihood analysis of the conserved regions using the RAxML program and the GTRI nucleotide model. The stability of each tree node was tested by bootstrap analysis with 1,000 replicates. Bootstrap values are indicated on the branches, and the branch length reflects the estimated number of substitutions per 1,000 sites.