

The complete chloroplast genome sequence of the medicinal plant *Ardisia crispa* (Myrsinaceae)

Juan Ye

Guizhou University of Traditional Chinese Medicine

Qin Luo

Guizhou University of Traditional Chinese Medicine

Yun-hu Lang

Guizhou University of Traditional Chinese Medicine

Ning Ding

Guizhou University of Traditional Chinese Medicine

Ying-quan Jian

Guizhou Hanfang Pharmaceutical Co., LTD

Zhi-kun Wu

Guizhou University of Traditional Chinese Medicine

Sheng-hua Wei

Guizhou University of Traditional Chinese Medicine

Fu-lin Yan

yanfulin743@gzy.edu.cn

Guizhou University of Traditional Chinese Medicine

Article

Keywords:

Posted Date: March 22nd, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4013297/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on August 16th, 2024. See the published version at <https://doi.org/10.1038/s41598-024-66563-3>.

Abstract

Ardisia crispa, a member of the Myrsinaceae family, possesses significant horticultural and medicinal properties as an ethnomedicine. The study aimed to analyze the chloroplast genome of *A. crispa* and compare it with other *Ardisia* species, revealing a length of 156,785 bp with a quadripartite structure and 131 genes, including 86 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. Furthermore, 59 simple sequence repeat (SSR) sites were identified in the genome. Examination of codon usage within the chloroplast genome indicated a greater inclination towards A/U nucleotides over G/C nucleotides, with leucine displaying the highest frequency among amino acids. The chloroplast genomes of the nine *Ardisia* species demonstrate conserved gene content and quantity, presenting more consistent boundaries and decreased variability. In the phylogenetic tree, *A. crispa* is clustered with *A. crispa* var *dielsii*, suggesting a close relationship with *A. mamillata* and *A. pedalis*. This study involved the construction and analysis of the chloroplast genome structure of *A. crispa*, as well as phylogenetic analysis using extensive chloroplast genome sequence data from *Ardisia* plants. This research is crucial for understanding the genetic basis of *A. crispa* and the adaptive evolution within the *Ardisia* genus.

Introduction

Ardisia crispa (Thunb.) A. DC., a member of the Myrsinaceae family, is predominantly found in the provinces south of the Yangtze River Basin in China. The root of this plant is utilized in traditional medicine and is a key component of the renowned Hmong medicinal specialty, Radix Ardisia, in Guizhou^{1,3}. Modern research has identified *A. crispa* as containing triterpene glycosides, flavonoids, isocoumarins, and other bioactive compounds, known for their therapeutic properties such as relieving heat, soothing the throat, and activating tendons. This plant is commonly used to treat conditions like sore throat, tonsillitis, nephritis, and edema, earning it the nickname "aryngeal medicine"². Two subspecies of *A. crispa* have been identified based on morphological differences, specifically *A. crispa* var *dielsii* (characterized by shorter stature and narrow lanceolate leaves) and *A. crispa* var *amplifolia* (recognized for its stout stature and broad leaves)⁴. However, the lack of a scientific foundation for this classification hinders a comprehensive understanding of the plant's genetic background, conservation of germplasm resources, and phylogenetic evolution. Furthermore, the structural features of the plastid genome, functional classification, codon preference analysis, and differences in the chloroplast genome between the two variants have not been thoroughly explored. This knowledge gap not only limits our understanding of *A. crispa* and its variants but also complicates the accurate identification of mixed pseudo-products involving *A. crispa* in medicinal herb preparations. Chloroplasts are semi-autonomous organelles found in plants, characterized by having the second largest genome in plant cells⁵. The chloroplast genome typically consists of a large single copy (LSC) region, a small single copy (SSC) region, and two inverted repeats (IR)⁶. Due to its uniparental inheritance, moderate mutation rate, and ease of sequencing, the chloroplast genome is often regarded as a more efficient resource compared to nuclear and mitochondrial genomes. Consequently, it is frequently utilized for investigating the origins and evolution of plants, elucidating phylogenetic relationships among different taxonomic classes, and

for species identification purposes⁷. In recent years, advancements in high-throughput sequencing technologies have facilitated the successful assembly, annotation, and analysis of a substantial number of subspecies chloroplast genomes. The application of chloroplast genome analysis has yielded promising results in the investigation of identification, genetic relationships, and phylogenetics of various species, including *Sabia*, *Phoebe*, *Vaccinium*, *Hibiscus rosa-sinensis*, *Dalbergia hainanensis*, *Litsea*, and *Zingiber*^{8–14}. Therefore, this research aimed to utilize high-throughput sequencing technology to sequence, assemble, and annotate the complete chloroplast genome of *A. crispa*. Additionally, comparative analysis of structural features and phylogenetic relationships among *A. crispa* and its varieties *A. crispa* var *dielsii* and *A. crispa* var *amplifolia*, as well as other *Ardisia* species, was conducted using bioinformatics tools. The findings of this study contribute to the understanding of species identification, phylogeny, and conservation efforts related to *A. crispa* and other medicinal herbs.

Results

Chloroplast genome assembly and annotation.

The chloroplast genome of *A. crispa* exhibits a structural composition typical of most angiosperms, characterized by a cyclic double-stranded molecule with a quadratic configuration (Fig. 1). The genome spans a total length of 156,785 bp, with a GC content of 37.0%. Specifically, the large single copy (LSC) region measures 86,342 bp, the small single copy (SSC) region measures 18,417 bp, and the inverted repeat (IR) region measures 26,014 bp. The GC contents of the LSC, IR, and SSC regions are 35.0%, 43.0%, and 30.1%, respectively. A total of 131 genes were annotated, including 86 protein-coding genes, 37 tRNA genes, 8 rRNA genes, 15 genes with 2 copies (*rrn4*, *rrn5*, *rrn16*, *rrn23*, *trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, *trnV-GAC*, *rps12*, *rps7*, *rpl2*, *ndhB*), and 21 genes had 1 intron (*trnA-UGC* (×2), *trnG-UCC*, *trnI-GAU* (×2), *trnK-UUU*, *trnL-UAA*, *trnV-UAC*, *rps12* (×2), *rps16*, *rpl16*, *rpl2* (×2), *rpoC1*, *petB*, *petD*, *atpF*, *ndhA*, *ndhB* (×2)), and 2 genes with 2 introns (*ycf3*, *clpP*) (Table 1.).

Table 1

Gene composition of *A. crisper* chloroplast genome. Genes marked with the sign are the genewith a single (*) or double (**) introns and duplicated genes (x2).

Gene function	Gene type	Gene names	Number
Self-replication	Ribosomal RNAs	<i>rrn4.5(x2)rrn5(x2)rrn16(x2)rrn23(x2)</i>	8
	Transfer RNAs	<i>trnA-UGC*(x2)trnC-GCAtrnD-GUCtrnE-UUCtrnF-GAAtrnG-GCCtrnG-UCC*trnH-GUGtrnI-CAU(x2)trnI-GAU*(x2)trnK-UUU*trnL-CAA(x2)trnL-UAA*trnL-UAGtrnM-CAUtrnN-GUU(x2)trnP-UGGtrnQ-UUGtrnR-ACG(x2)trnR-UCUtrnS-GCUtrnS-GGAtrnS-UGAtrnT-GGUtrnT-UGUtrnV-GAC(x2)trnV-UAC*trnW-CCAtrnY-GUAtrnfM-CAU</i>	37
	Proteins of small ribosomal subunit	<i>rps11rps12**(x2)rps14rps15rps16*rps18rps19rps2rps3rps4rps7(x2)rps8</i>	14
	Proteins of large ribosomal subunit	<i>rpl14rpl16*rpl2*(x2)rpl20rpl22rpl32rpl33rpl36</i>	9
	Subunits of RNA polymerase	<i>rpoArpoBrpoC1*rpoC2</i>	4
Photosynthesis	Subunits of photosystem I	<i>psaApsaBpsaCpsalpsaJ</i>	5
	Subunits of photosystem II	<i>psbApsbBpsbCpsbDpsbEpsbFpsbHpsbIpsbJpsbKpsbLpsbMpsbNpsbTpsbZ</i>	15
	Subunits of cytochrome b/f complex	<i>petApetB*petD*petGpetLpetN</i>	6
	Subunits of ATP synthase	<i>atpAatpBatpEatpF*atpHatpI</i>	6
	Protease	<i>clpP**</i>	1
	Large subunit of rubisco	<i>rbcL</i>	1
	NADH dehydrogenase subunit	<i>ndhA*ndhB*(x2)ndhCndhDndhEndhFndhGndhHndhIndhJndhK</i>	12
Other genes	Maturase	<i>matK</i>	1
	Envelope membrane protein	<i>cemA</i>	1
	Acetyl-CoA carboxylase	<i>accD</i>	1
	c-type cytochrome	<i>ccsA</i>	1

Gene function	Gene type	Gene names	Number
	synthesis gene		
	Translation initiation factor	<i>InfA</i>	1
Genes of unknown function	Conserved hypothetical chloroplast ORF	<i>ycf1</i> – <i>ycf15</i> (×2)– <i>ycf2</i> (×2)– <i>ycf3</i> **– <i>ycf4</i> – <i>ycf68</i>	8

Repeat sequences analysis.

The chloroplast genome of *A. crispa* contained 59 SSRs, consisting of 44 single-nucleotide repeats, 5 dinucleotide repeats, 8 tetranucleotide repeats, and 2 pentanucleotide repeats, with no trinucleotide or hexanucleotide repeats identified. The predominant types of SSRs were A/T repeats (41) (Fig. 2A). Additionally, a total of 49 long repetitive sequences were identified, comprising 22 forward repeats, 26 palindromic repeats, and 1 inverted repeat. No complementary repetitive sequences were observed, with the lengths of forward and palindromic repetitive sequences primarily falling within the range of 30 to 49 bp (Fig. 2B). A total of 38 tandem repeats were identified, ranging in length from 10 to 30 base pairs.

Codon analysis.

The analysis of chloroplast codon statistics in *A. crispa* revealed that the 86 protein-coding genes encompassed 61 distinct codon species, totaling 52,261 codons that encode 20 different amino acids, including 3 termination codons (Fig. 3). Among the amino acid codons, leucine (Leu), serine (Ser), and arginine (Arg) were each encoded by six codons, with frequencies of 5,218 (9.98%), 4,779 (9.14%), and 3,216 (6.15%), respectively. Leucine was found to be the most commonly utilized codon, followed by serine, while tryptophan (Trp) was the least utilized with only 688 codons (1.32%). Among all the codons observed in the chloroplast genome of *A. crispa*, the codon AAA exhibited the highest frequency of 2,182 occurrences with a relative synonymous codon usage (RSCU) value of 1.35, while the codon GCG had the lowest frequency of 221 occurrences with an RSCU of 1.22. A total of 35 codons in the genome had an RSCU value greater than or equal to 1, with 8 ending in G or C and 27 ending in A or U. Additionally, the *A. crispa* chloroplast genome displayed an effective number of codons (ENc) value of 55.71, a codon adaptation index (CAI) value of 0.652, and GC, GC1, GC2, and GC3 contents of 37.04%, 36.89%, 36.61%, and 37.62%, respectively.

IR contraction and expansion in the chloroplast genome.

The study compared the boundaries between the inverted repeat (IR) and large single copy (LSC) regions in the chloroplast genomes of 9 species of *Ardisia*. Results indicated the presence of 4 distinct boundaries in the chloroplast genome of *Ardisia*, with variations in the genes located at these boundaries and their respective lengths (Fig. 4). Specifically, the *A. crispa* and the remaining 8 species exhibited the presence of *rpl22* gene on the left side of the genome-wide LSC/IRb boundary (JLB), *rpl2* gene on the right side, and *rps19* gene spanning the JLB. Notably, an expansion of the *rps19* gene into

the LSC region was observed in *A. crispa* and *A. crispa* var. *dielsii*, with a 240 bp expansion, while *A. crispa* var. *amplifolia*, *A. crenata*, *A. crenata* var. *Bicolor*, *A. japonica*, *A. polysticta* showed a 232 bp expansion, and a 69 bp expansion in *A. gigantifolia* and *A. bullata*. The analysis of the IRb/SSC (JSB) boundary expansion revealed that, with the exception of *A. japonica*, whose JSB boundary was positioned to the left of the *ndhF* gene, the *ndhF* gene of the remaining 8 species encompassed the JSB boundary. However, the degree of expansion exhibited slight variation, with *A. crispa*, *A. crispa* var. *dielsii*, *A. gigantifolia*, and *A. bullat*, expanding by 5 bp into the IRb region, while the remaining 4 species expanded by 3 bp. The analysis of the SSC/IRa (JSA) and LSC/IRa (JLA) boundaries in nine *Ardisia* species revealed variations in the expansion of *ycf1* and *trnH* genes, respectively. Specifically, the *ycf1* gene exhibited expansion ranging from 4,600 bp to 4,614 bp towards the SSC region, while the *trnH* gene spanned the JLA boundary. Additionally, the *rps1* gene of *A. crenata* and *A. polysticta* were found to be located at the JLA boundary.

Comparative chloroplast genomic and nucleotide diversity analyses.

In this study, the chloroplast genome sequences of 9 species of *Ardisia* were compared and analyzed to evaluate the extent of differences. *A. crispa* (OP626693) was used as the reference genome, and the mVISTA online tool was employed for the analysis (Fig. 5). The nucleotide polymorphism analysis revealed a mean nucleotide diversity (P_i) value of 0.00459 among the 9 species. Six highly variable regions (*trnT-psbD*, *ndhB-trnL*, *rpl32-trnL*, *trnL-ccsA*, *trnL-ndhB*) were identified when $P_i > 0.02$, with one region in IRa, IRb, and LSC, and two regions in SSC showing significant variability (Fig. 6).

Phylogenetic analyses.

In order to ascertain the phylogenetic placement of *A. crispa*, the Bayesian inference (BI) phylogenetic tree was generated using 29 chloroplast genome sequences from 22 *Ardisia* species (Fig. 7). The analysis revealed that *A. crispa* and *A. crispa* var. *dielsii* formed a sister relationship on a single branch, showing closer affinity to *A. mamillata* and *A. pedalis*. Conversely, *A. crispa* var. *amplifolia* clustered with *A. crenata* var. *bicolor* on a separate branch, indicating a closer relationship to *A. crenata*. Furthermore, the findings of this study provide support for the taxonomic separation of the Primulaceae from the Myrsinaceae.

Discussion

In this study, the chloroplast genome of *A. crispa* was successfully sequenced, assembled, and annotated. Consistent with the characteristics of most angiosperms, the chloroplast genome of *Paris mairei* exhibits a typical tetrameric structure, with the GC content in the sequences of each region following the pattern of IRs > LSC > SSC. This distribution may be attributed to the presence of high GC content rRNA genes in the IR region, which aligns with findings from previous studies on *Paris mairei*, *A. crenata*, *A. crenata* var. *bicolor*, *A. crispa* var. *dielsii*, and *A. crispa* var. *amplifolia*^{15–17}. Repetitive sequences are prevalent in chloroplast genomes, with their type, number, and distribution varying among

species or populations. Interspersed repeats within the chloroplast genome of *A. crispa* have been identified as valuable tools in genetic variation, structure analysis, and species identification^{18–21}. Among the three types detected (forward, reverse, and palindromic), forward and palindromic repeats were found to be the most prevalent, with repeat sequence lengths predominantly falling within the range of 30–49 base pairs. These findings align with previous studies on chloroplast genome repeats in various plant species, including A, B, and C, among others^{17,22}.

In this study, 59 SSRs were identified, with the predominant occurrence of single-nucleotide repeats composed of A or T. This observation suggests a high prevalence of A or T nucleotides in the base composition of the *A. crispa* chloroplast genome. The SSRs identified in the *A. crispa* chloroplast genome may serve as valuable resources for the development of molecular markers and species identification within the *Ardisia*.

Codon preference is the unequal use of synonymous codons encoding the same amino acids by species. Codon usage bias is an important feature of genome evolution and is important for the study of molecular evolution and gene ectopic expression²³. The Relative Synonymous Codon Usage (RSCU) metric quantifies the ratio between the observed frequency of a codon and its expected frequency based on theoretical calculations. A value of RSCU = 1 indicates equal usage frequency among synonymous codons, suggesting no preference in codon usage. Values of RSCU > 1 suggest a strong preference in codon usage, while values of RSCU < 1 indicate a weak preference. Mutations in the third position of codons are subject to less selective pressure compared to mutations in the first and second positions, and are often associated with changes in amino acid species. The codon's 3rd base composition and content is one of the most important indicators of genomic preference, and higher plants tend to use codons ending in A/U²⁴. In the analysis of codon preference in the chloroplast genome of *A. crispa*, 35 codons with RSCU $\geq$ 1 were identified, of which 27 codons terminated in A/U. This observation suggests that synonymous codons in the chloroplast genome of *A. crispa* exhibit a preference for ending in A/U, consistent with findings in other genomes such as *Phyllanthaceae*²⁵, *Notopterygium*²⁶, *Cinnamomum camphora*²⁷, among others. Furthermore, the ENc value of the *A. crispa* chloroplast genome was calculated to be 55.71, indicating a mild preference. The GC and GC3 contents of the *A. crispa* chloroplast genome were both below 50%, suggesting a bias towards the use of A and T bases, a pattern that is similar to the results of *Dendrobium devonianum*²⁸ and *Glycyrrhiza eurycarpa*²⁹.

The chloroplast genome's inverted repeat (IR) region is a prevalent feature in the genomes of many higher plants, and its dynamic contraction and expansion are widely recognized as a key evolutionary process contributing to variations in chloroplast genome size³⁰. Expansion of the IR region can facilitate the incorporation of genes located at the genome's periphery. Furthermore, the presence of reverse repeat sequences within the IR region can result in the formation of intact genes or partial gene fragments on the opposite side of the region.

This study examined the chloroplast genome boundaries of nine *Ardisia* species, revealing significant differences between the JLB and JLA boundaries, with the JLB and JSA boundaries showing more

conservation. The absence of the *ycf1* gene in *A. crispa*, *A. crispa* var. *dielsii*, and *A. crispa* var. *amplifolia*, as well as the presence of a pseudogene *ycf1* at the JSB border, has been documented³¹. Previous research suggests that variations in selective pressure on the *ycf1* gene contribute to differences in evolutionary rates^{32–33}.

The presence of the *ndhF* gene was limited to the SSC region in *A. japonia*, while the *rps1* gene was found in the IRa region in *A. polysticta*, *A. gigantifolia* and *A. crenata*. The expansion of the *rps19* gene into the IRb region was observed in *A. bullata* and *A. gigantifolia*. The phenomenon of boundary expansion and contraction was evident in the analysis of chloroplast gene boundaries in other plants within the same genus. Similar results were found in the chloroplast genome boundary analysis of species such as *Rubia cordifolia*³⁴, *Polygala sibirica*³⁵, and *Triticum*³⁶, indicating that changes in chloroplast genome boundaries do not follow a consistent pattern.

Nucleotide diversity can be calculated to quantify differences in cp genomes at the sequence level³⁷. These regions may undergo accelerated nucleotide substitution at the species level, suggesting their potential for use as molecular markers in plant identification and phylogenetic analysis³⁸. The results of nucleotide polymorphism analysis in this study showed that the non-coding regions of the cp genome sequences of the 9 species of *Ardisia* were highly variable. There were obvious differences in the spacer regions of the *trnT-psbD*, *rpl32-trnL*, *trnL-ccsA*, *trnL-ndhB*, and *ndhB-trnL* genes, and these regions of variability can provide the basis for the development of molecular markers, species identification, and DNA barcode screening of *Ardisia*.

The chloroplast genome has demonstrated efficacy in elucidating phylogenetic relationships among plant taxa^{39–40}. In this investigation, a Bayesian inference phylogenetic tree was constructed utilizing 29 chloroplast whole genomes of *Lysimachia christinae* (Primulaceae) as an outgroup. The analysis revealed a close relationship between *A. crispa* and *A. crispa* var. *dielsii*, with both taxa forming a distinct clade. However, *A. crispa* var. *amplifolia* did not cluster with these two taxa, indicating significant intraspecific variation within Hypericum, potentially influenced by geographical factors. The clustering of *A. crispa* var. *amplifolia* with *A. crenata* and *A. crenata* var. *bicolor* at 100% support aligns with the ITS and ITS2 sequence identifications documented in previous literature^{41–42}. Consistent with earlier research, our findings suggest a closer relationship between *A. crenata* and *A. crenata* var. *bicolor*, as well as *A. polysticta*^{17,43}. Our study utilized a phylogenetic analysis of chloroplasts to distinguish Myrsinaceae and Primulaceae into separate branches with strong support, advocating for the continued classification of Myrsinaceae as a distinct taxonomic group.

Conclusion

This study involved sequencing the chloroplast genome of *A. crispa*, conducting a comprehensive analysis of its sequence, structure, and characteristics, identifying differential sequences in the chloroplast genome of *A. crispa* and its two variants, and investigating the phylogenetic relationships of

Ardisia. The findings not only enhanced the knowledge of the chloroplast genome of *A. crispa*, but also served as a valuable resource for taxonomic identification and phylogenetic analysis of the genus *Ardisia*. This has significant implications for the conservation of *Ardisia* germplasm resources and the identification of valuable genetic resources.

Material and methods

Sample collection.

The fresh leaves of sample (Fig. 8) were collected from the Ceheng county, Guizhou Province, China, (coordinates: E105°47'29.73", N24°59'59.51"; altitude: 933 m). It was identified as *Ardisia crispa* (Thunb.) A. DC.) by Associate Professor Yan Fulin of Guizhou University of Traditional Chinese Medicine. The voucher specimen (with collection numbers of YFL_2021040307) has been deposited in the Herbarium of Guizhou University of Traditional Chinese Medicine (GZYGH), Guizhou, China. The collection of plant materials complies with the wild plant protection regulations of the People's Republic of China, and we obtained the permission of local authorities on forestry and the grassland bureau in Guizhou province in China.

DNA extraction and chloroplast genome sequencing.

Total DNA was extracted from the fresh leaves of *A. crispa* by the modified CTAB method⁴². Sequencing was carried out on the Illumina HiSeq XTen to generate approximately 3 GB 150 bp reads at Beijing Genomics Institute (BGI, Wuhan, China).

Genome assembly and annotation.

The filtered reads were assembled into a complete chloroplast genome by the program GetOrganelle v1.5⁴⁵. In this pipeline, the complete chloroplast genome reads were extracted from total genomic reads and were subsequently assembled using SPAdes version 3.10⁴⁶. The genes were annotated using PGA⁴⁷ and Geneious 11.0.3⁴⁸ with the published complete chloroplast genome of *A. crenata* (GenBank accession number: NC_059021) as the reference. Transfer RNAs (tRNAs) were confirmed by their specific structure predicted by tRNAscan-SE 2.0⁴⁹. The OGDRAW (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>)⁵⁰ was used to draw a detailed physical map of the *A. crispa* chloroplast genome.

Chloroplast genome structural analysis.

The online software REPuter (<https://bibiserv.cebitec.unibielefeld.de/reputer>) was employed to analyze forward (F), palindromic (P), reverse (R), and complement (C) repeats, with the following settings: Minimum repeat size of 3 bp and hamming Distance of 30 bp⁵¹. We used the default parameters in the online Tandem Repeats Finder (<http://tandem.bu.edu/trf/trf.html>)⁵² to search for tandem repeats in DNA sequences. The online software MISA⁵³ was applied to predict SSRs with parameter thresholds set at 1,

2, 3, 4, 5, and 6, and nucleotide parameters of 10, 5, 4, 3, 3, 3, and 3, and the distance between two SSRs was not less than 100 bp. Relative synonymous codon usage (RSCU) analysis of the chloroplast genome of *A. crispa* was performed using condon W (<https://galaxy.pasteur.fr/?form=codonw>) online software. By using CUSP online software (<http://imed.med.ucm.es/EMBOSS/>)⁵⁴, we calculated the effective codon count (ENc), codon adaptation index (CAI), and counted the total codon GC content (GCall), the GC contents of positions 1, 2, and 3 (GC1, GC2, and GC3), and the GC content of position 3 of the synonymous codons (GC3s).

Genome comparison.

The genes of the boundaries in the chloroplast genomes of the *Ardisia* species were compared and visually represented using IRscope⁵⁵(<https://irscope.shinyapps.io/irapp/>) to reveal contraction and expansion of the IR regions. The comparative analysis of the whole sequence identity of the chloroplast genomes was performed using mVISTA⁵⁶ with the chloroplast genome of *A.crispa* (OP626693) as the reference sequence.

Phylogenetic analysis.

A phylogenetic analysis was conducted based on chloroplast genomes from 29 species, including those of the one *A. crispa* sequenced and assembled in this study and another 28 downloaded from GenBank. Twenty-nine complete plastid sequences were aligned using MAFFT v7.017⁵⁷, and a BI phylogenetic tree was constructed using MrBayes 3.2.7⁵⁸ with *Lysimachia christinae* (Primulaceae) as the outgroup.

Declarations

Data availability

The complete chloroplast genomes and annotations are available at the NCBI database (*Ardisia crispa*: OP762693).

Competing interests

The authors declare no competing interests.

Author Contribution

F.L.Yan designed the research. Z.K.Wu, Y.Q.J., S.H.Wei, F.L.Yan and L.D. collected the samples, J.Ye, F.L.Yan, Q.Luo, and Y.H.Lang conceived the experiments, J.Ye, F.L.Yan, Q.Luo, and Y.H.Lang did computational analysis and deposited sequences. J.Ye, F.L.Yan, and Y.H.Lang wrote the manuscript. All authors have read and approved the manuscript.

Acknowledgements

This study was supported by the National Key Research and Development Program Project (2018YFC1708101), Guizhou Provincial Science and Technology Program Project (No. Qiankezhongyindi [2022] 4016), Guizhou Modern Industrial Technology System Construction of Chinese Herbal Medicines (No. GZCYTX2019-2024), and State Administration of Traditional Chinese Medicine Seedling Breeding of Chinese Herbal Medicines Seeds Required for National Essential Drugs (Guizhou) Base Construction Project (2014–2017).

Additional information

Correspondence and requests for materials should be addressed to J.Ye or F.L.Yan.

References

1. Guizhou Medical Products Administration. Quality standards for traditional Chinese medicinal materials and ethnic medicinal materials in Guizhou Province. Guiyang: *Guizhou Science and Technology Press*. 164, (2003).
2. Zhang, N. L. *et al.* Chemical Constituents of *Ardisia crispa* (Thunb.) A. DC. *Natural Product Research and Development* **22**, 587-589. <http://doi.org/10.16333/j.1001-6880.2010.04.039> (2010).
3. Li, M. *et al.* Investigation and Application Evaluation of *Ardisia* Resources in Guizhou. *Guizhou Agricultural Sciences*. **47**, 140-144(2019).
4. Editorial Committee of Flora of China, Chinese Academy of Sciences. Flora of China. Beijing: Science Press. **34** (1979).
5. Xue, S. *et al.* Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Hort. Res.* **6**, 89. <https://doi.org/10.1038/s41438-019-0171-1> (2019).
6. Li, X. *et al.* Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. *Sci China Life Sci.* **56**, 189-198. <http://doi.org/10.1007/s11427-012-4430-8> (2013).
7. Chen, Q., Hu, H. & Zhang, D. DNA Barcoding and phylogenomic analysis of the genus *Fritillaria* in China based on complete chloroplast genomes. *Front. Plant Sci.* **13**, 764255. <http://doi.org/10.3389/fpls.2022.764255> (2022).
8. Chen, Q. *et al.* Complete chloroplast genomes of 11 *Sabia* samples: Genomic features, comparative analysis, and phylogenetic relationship. *Front. Plant Sci.* **13**, 1052920. <http://doi.org/10.3389/fpls.2022.1052920> (2022).
9. Shi, W. *et al.* Comparative chloroplast genome analyses of diverse *Phoebe* (Lauraceae) species endemic to China provide insight into their phylogeographical origin. *PeerJ.* **11**, e14573. <http://doi.org/10.7717/peerj.14573> (2023).

10. Annette, M. *et al.* Chloroplast genome assemblies and comparative analyses of commercially important *Vaccinium* berry crops. *Sci. Rep.* **12**, 21600. <https://doi.org/10.1038/s41598-022-25434-5> (2022).
11. Abdullah, Mehmood. F. *et al.* Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. *Genomics.* **112**, 581–591. <https://doi.org/10.1016/j.ygeno.2019.04.010> (2020).
12. Deng, C. Y. *et al.* Characterization of the complete chloroplast genome of *Dalbergia hainanensis* (Leguminosae), a vulnerably endangered legume endemic to China. *Conserv Genet Resour.* **11**, 105–108. <https://doi.org/10.1007/s12686-017-0967-y> (2019).
13. Song, W. *et al.* Comparative Analysis of Complete Chloroplast Genomes of Nine Species of *Litsea* (Lauraceae): Hypervariable Regions, Positive Selection, and Phylogenetic Relationships. *Genes.* **13**, 1550. <https://doi.org/10.3390/genes13091550> (2022).
14. Jiang, D. Z. *et al.* Complete chloroplast genomes provide insights into evolution and phylogeny of *Zingiber* (Zingiberaceae). *BMC Genom.* **24**, 30. <https://doi.org/10.1186/s12864-023-09115-9> (2023).
15. Jiang, R. *et al.* Complete chloroplast genome of *Paris mairei*: characterization and phylogeny. *Chinese Herb. Med.* **52**, 4014-4022. <https://doi.org/10.7501/j.issn.0253-2670.2021.13.024> (2021).
16. Zeng, X. F. *et al.* Chloroplast genome resolution and phylogenetic analysis of *Ardisia crispa* var. *amplifolia* and *Ardisia crispa* var. *dielsii*. *Acta Pharmaceutica Sinica.* **58**, 217-228. <https://doi.org/10.16438/j.0513-4870.2022-0874> (2023).
17. Liu, X. W. *et al.* Comparative and Phylogenetic Analyses of Complete Chloroplast Genomes in *Ardisia crenata*. *Biotechnology Bulletin.* **39**, 232-242. <https://doi.org/10.13560/j.cnki.biotech.bull.1985.2022-0471> (2023).
18. Yuan, Q. *et al.* Impacts of recent cultivation on genetic diversity pattern of a medicinal plant, *Scutellaria baicalensis* (Lamiaceae). *BMC Genet.* **11**, 1-13. <https://doi.org/10.1186/1471-2156-11-29> (2010).
19. Chmielewski, M. *et al.* Chloroplast microsatellites as a tool for phylogeographic studies: the case of white oaks in Poland. *IFOREST.* **8**, 765. <https://doi.org/10.3832/ifor1597-008> (2015).
20. Asaf, S. *et al.* Complete chloroplast genome of *Nicotiana otophora* and its comparison with related species. *Front Plant Sci.* **7**, 843. <https://doi.org/10.3389/fpls.2016.00843> (2016).
21. Zhuo, L. *et al.* Advances in the application of SSR markers in the identification of plant germplasm resources. *Contemporary Horticulture.* **44**, 9-11. <https://doi.org/10.14051/j.cnki.xdy.2021.15.005> (2021).
22. Shang, M. Y. *et al.* Analysis of chloroplast genome structure and phylogeny of endangered *Dendrobium devonianum*. *Chinese Traditional and Herbal Drugs.* **54**, 6424-6433. <https://doi.org/10.7501/j.issn.0253-2670.2023.19.023> (2023).
23. Wang, Y. Z. *et al.* Comparative analysis of codon usage patterns in chloroplast genomes of ten *Epimedium* species[J]. *BMC Genom Data.* **24**, 3. <https://doi.org/10.1186/s12863-023-01104-x> (2023).

24. Wang, Z. J. *et al.* Comparative analysis of codon usage patterns in chloroplast genomes of six *Euphorbiaceae* species. *Peer J.* **8**: e8251. <https://doi.org/10.7717/peerj.8251> (2020).
25. Gao, C. *et al.* Codon bias analysis of chloroplast genome of *Artocarpus heterophyllus*. *Journal of Fujian Agriculture and Forestry University (Natural Science Edition)*. **52**, 776-784. [https://doi.org/10.13323/j.cnki.j.fafu\(nat.sci.\).2023.06.008](https://doi.org/10.13323/j.cnki.j.fafu(nat.sci.).2023.06.008) (2023).
26. Long, T. *et al.* Codon Usage Bias Analysis in the *Acer amplum* subsp. *catalpifolium* Genome. *Journal of Northwest Forestry University*. **38**, 61-66+80. <https://doi.org/10.3969/j.issn.1001-7461.2023.06.08> (2023).
27. Hong, S. R. *et al.* Analysis of the Complete Chloroplast Genome Sequence Characteristics and Its Code Usage Bias of *Sorghum bicolor*. *Acta Agrestia Sinica*. <https://link.cnki.net/urlid/11.3362.S.20231026.1009.002> (2023).
28. Shang, M. Y. *et al.* Complete chloroplast genome of endangered *Dendrobium devonianum* Paxt.: characterization and phylogeny. *Chinese Traditional and Herbal Drugs*. <http://kns.cnki.net/kcms/detail/12.1108.R.20230828.1757.004>. (2023).
29. Zhang, J. *et al.* Characteristics of the chloroplast genome of *Glycyrrhiza eurycarpa* P.C.Li from Xinjiang with comparison and phylogenetic analysis of the chloroplast genomes of the medicinal plants of *Glycyrrhiza*. *Acta Pharmaceutica Sinica*. **57**, 1516-1525. <https://doi.org/10.16438/j.0513-4870.2021-1661> (2022).
30. Wang, R. J. *et al.* Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol Biol.* **8**, 1-14. <https://doi.org/10.1186/1471-2148-8-36> (2008).
31. Jiang, M. *et al.* Assembly and sequence analysis of *Tetrastigma hemsleyanum* chloroplast genome. *Chinese Traditional and Herbal Drugs*. **51**, 461-468. <https://doi.org/10.7501/j.issn.0253-2670.2020.02.024> (2020).
32. Yang, X. *et al.* PBR1 selectively controls biogenesis of photosynthetic complexes by modulating translation of the large chloroplast gene Ycf1 in *Arabidopsis*. *Cell Discov.* **2**, 1-19. <https://doi.org/10.1038/celldisc.2016.3> (2016).
33. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu Rev Genet.* **47**, 97-120. <https://doi.org/10.1146/annurev-genet-111212-133526> (2013).
34. Chen, X. Y. *et al.* Complete Chloroplast Genome and Phylogenetic Analysis of *Rubia cordifolia*. *Acta Botanica Boreali-Occidentalia Sinica*. **43**, 1855-1865. <https://doi.org/10.7606/j.issn.1000-4025.2023.11.1855> (2023).
35. Luo, Y. *et al.* Chloroplast genome sequence characteristics and phylogenetic analysis of *Polygala sibirica*. *Chinese Traditional and Herbal Drugs*. **54**, 6065-6073. <https://doi.org/10.7501/j.issn.0253-2670.2023.18.024> (2023).
36. Li, Y. H. Bioinformatics Analysis of *Triticum* species Chloroplast Genomes. *Shanxi University*. <https://doi.org/10.27284/d.cnki.gsxiu.2021.000053> (2021).

37. Li, H. *et al.* Chloroplast genomic comparison of two sister species *Allium macranthum* and *A. fasciculatum* provides valuable insights into adaptive evolution. *Genes and Genomics*. **42**, 507–517. <https://doi.org/10.1007/s13258-020-00920-0> (2020).
38. Suo, Z. *et al.* A new nuclear dna marker revealing both microsatellite variations and single nucleotide polymorphic loci: a case study on classification of cultivars in *Lagerstroemia indica* L. *Journal of Microbial & Biochemical Technology*. **8**, 266–271. <https://doi.org/10.4172/1948-5948.1000296> (2016).
39. Li, E. Z. *et al.* Insights into the phylogeny and chloroplast genome evolution of *Eriocaulon* (Eriocaulaceae). *BMC Plant Biol.* **23**, 32. <https://doi.org/10.1186/s12870-023-04034-z> (2023).
40. Yang, L. *et al.* Comparative chloroplast genomics of 34 species in subtribe *Swertiinae* (Gentianaceae) with implications for its phylogeny. *BMC Plant Biol.* **23**, 164. <https://doi.org/10.1186/s12870-023-04034-z> (2023).
41. Pan, J. *et al.* Screening and Identification on ITS Sequences of Original Plants from *Ardisia crispa*. *Molecular Plant Breeding*. **18**, 8187-8195. <https://doi.org/10.13271/j.mpb.018.008187> (2020).
42. Wen, Q. Q. *et al.* ITS2 sequence identification of Miao medicine *Ardisia crispa* medicinal materials and their related mixed counterfeits. *Journal of Chinese Medicinal Materials*. **45**, 830-835. <https://doi.org/10.13863/j.issn1001-4454.2022.04.011> (2022).
43. Xie, C. *et al.* Comparative genomic study on the complete plastomes of four officinal *Ardisia* species in China. *Sci Rep.* **11**, 22239. <https://doi.org/10.1038/s41598-021-01561-3> (2021).
44. Doyle, J. DNA protocols for plants: CTAB total DNA isolation. In Hewitt GM, Johnston A, editors. *Molecular techniques in taxonomy*. Berlin: Springer. (1991).
45. Jin, J. J. GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular cp genome using genome skimming data. *BioRxiv*. 256479. <https://doi.org/10.1101/256479> (2018).
46. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* **19**, 455-477. <https://doi.org/10.1089/cmb.2012.0021> (2012).
47. Qu, X. J. *et al.* PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods*. **15**, 50. <https://doi.org/10.1186/s13007-019-0435-7> (2019).
48. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**, 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199> (2012).
49. Lowe, T. M. & Chan P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54-W57. <https://doi.org/10.1093/nar/gkw413> (2016).
50. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**: W59-W64. <https://doi.org/10.1093/nar/gkz238> (2019).
51. Kurtz, S. *et al.* REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids. Res.* **29**, 4633–4642. <https://doi.org/10.1093/nar/29.22.4633> (2001).

52. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **15**, 573-80. <https://doi.org/10.1093/nar/27.2.573> (1999).
53. Beier, S. *et al.* MISA-web: a web server for microsatellite prediction. *Bioinformatics.* **33**, 2583-2585. <https://doi.org/10.1093/bioinformatics/btx198> (2017).
54. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16**, 276–277. [https://doi.org/10.1016/s0168-9525\(00\)02024-2](https://doi.org/10.1016/s0168-9525(00)02024-2) (2000).
55. Amiryousefi, A., Hyvönen, J. & Poczar, P. IRscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics.* **34**, 3030–3031. <https://doi.org/10.1093/bioinformatics/bty220> (2018).
56. Ma, J. Y. *et al.* The complete chloroplast genome characteristics of *Polygala crotalarioides* Buch.-Ham. ex DC. (*Polygalaceae*) from Yunnan, China. *Mitochondrial DNA B Resour.* **6**, 2838-2840. <https://doi.org/10.1080/23802359.2021.1964396> (2021).
57. Katoh, K. *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066. <https://doi.org/10.1093/nar/gkf436> (2002).
58. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* **17**, 754–755. <https://doi.org/10.1093/bioinformatics/17.8.754> (2001).

Figures

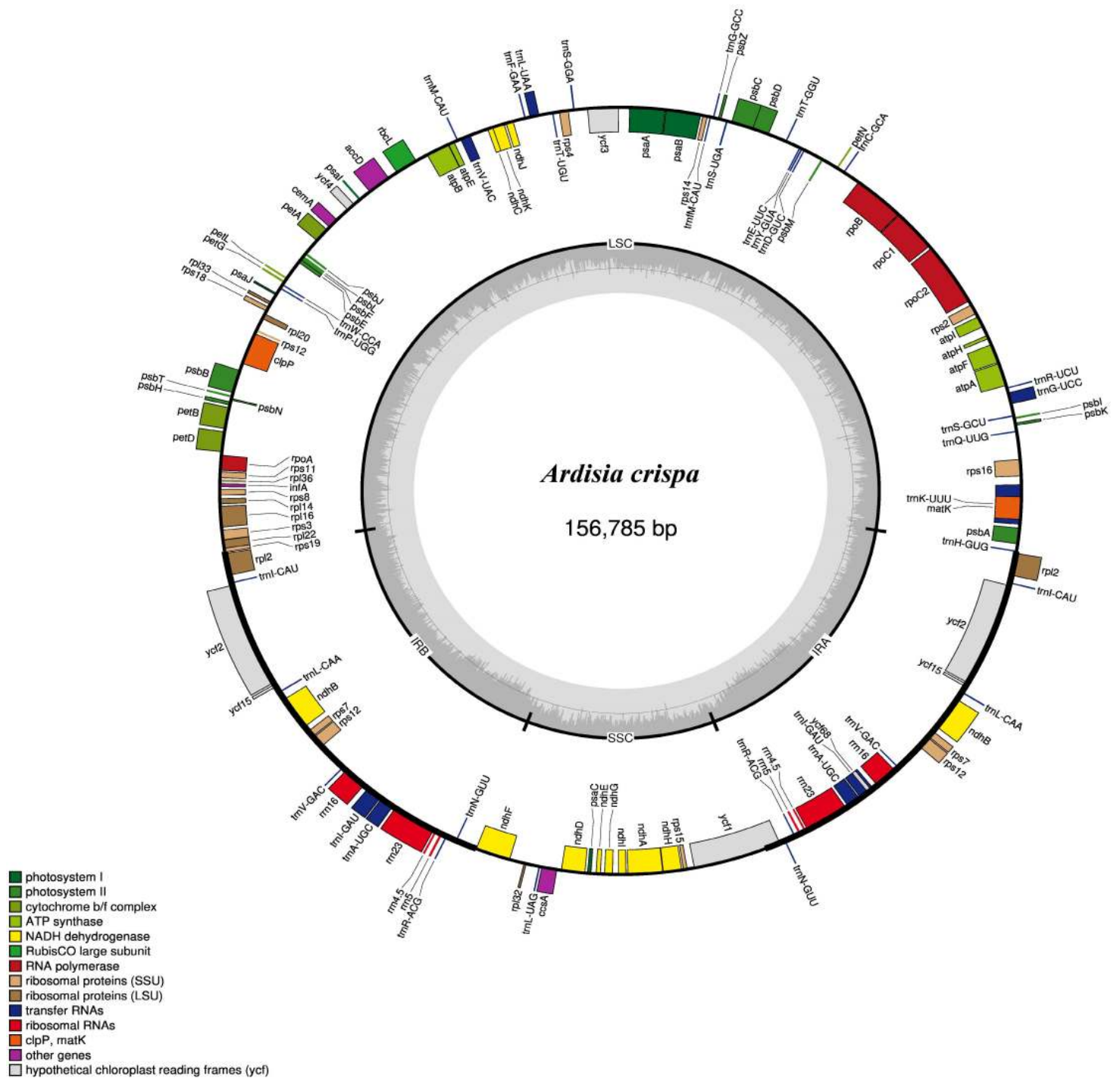


Figure 1

Chloroplast genome map of *A. crispata*. Genes shown inside the circle are transcribed clockwise, whereas genes outside are transcribed counterclockwise. The light gray inner circle shows the AT content, the dark gray corresponds to the GC content.

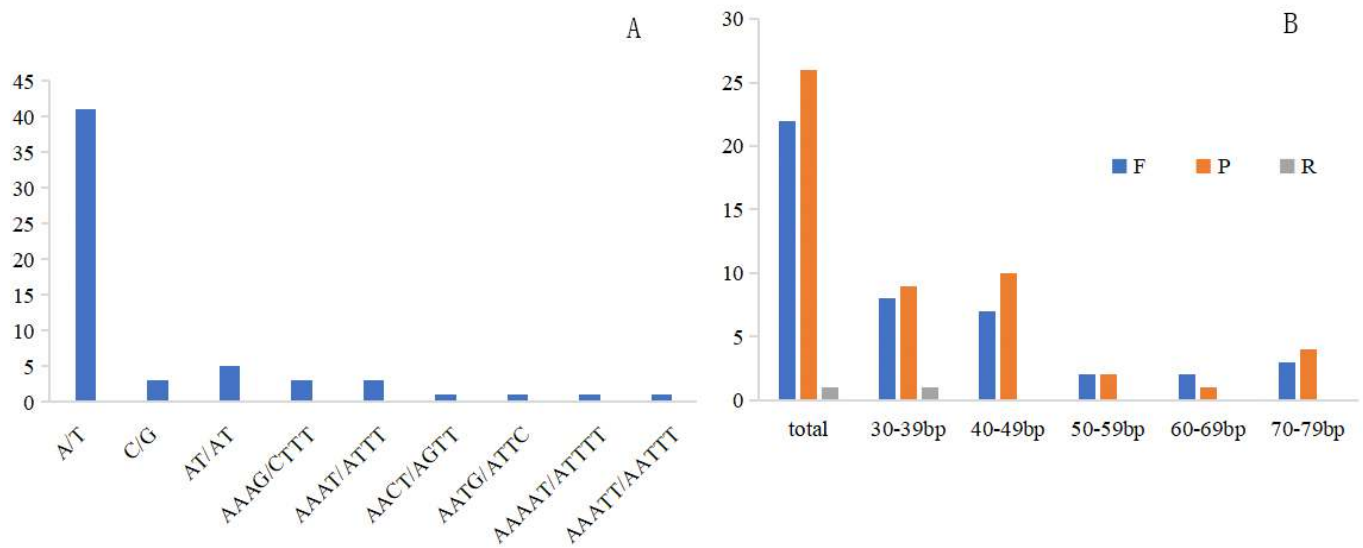


Figure 2

Repeat type and number of *A. crispata* chloroplast genome (A: SSR type B: long repeat type and frequency of use)

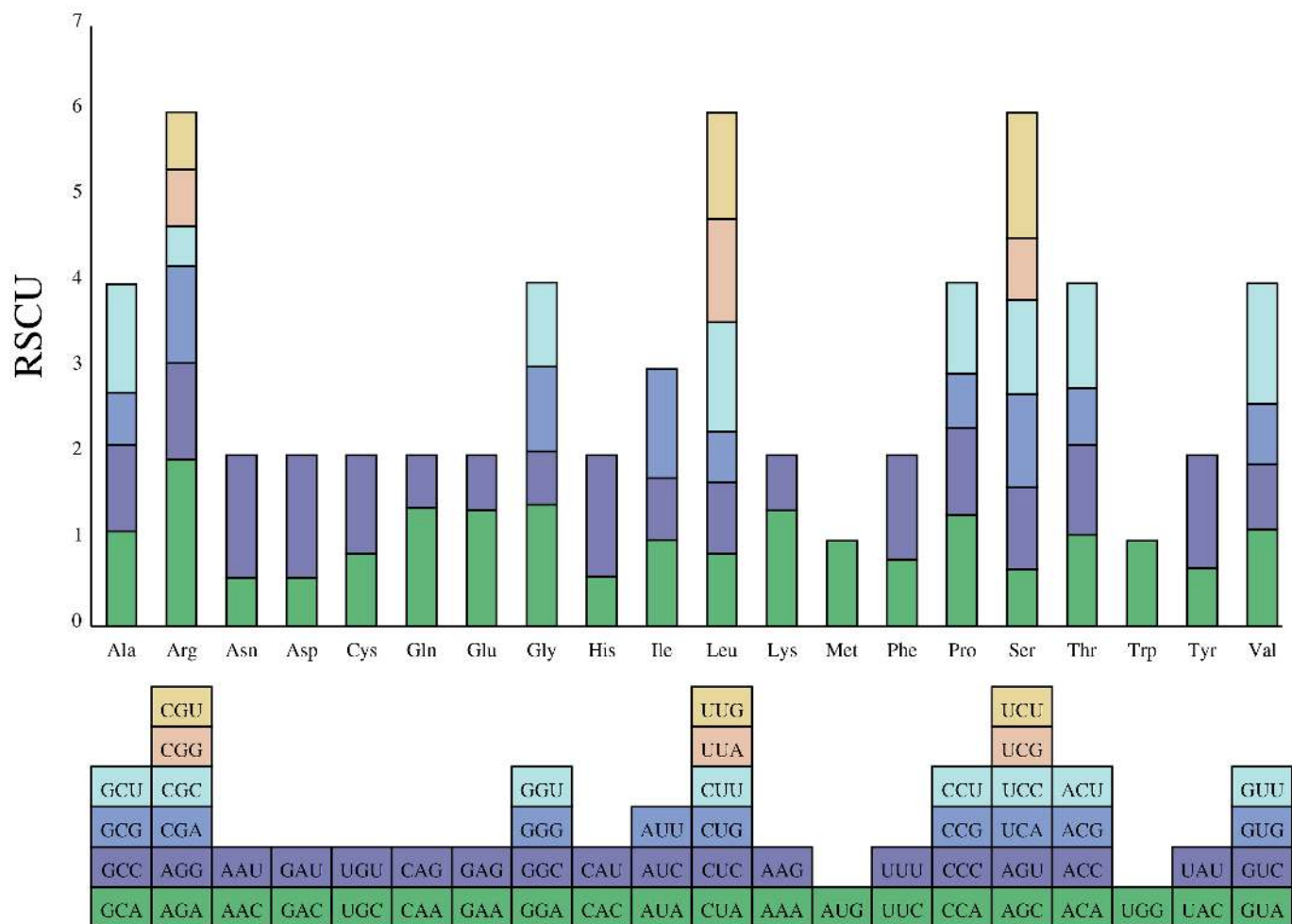


Figure 3

Relative synonymous codon usage (RSCU) for protein-coding genes in *A. crispata*.

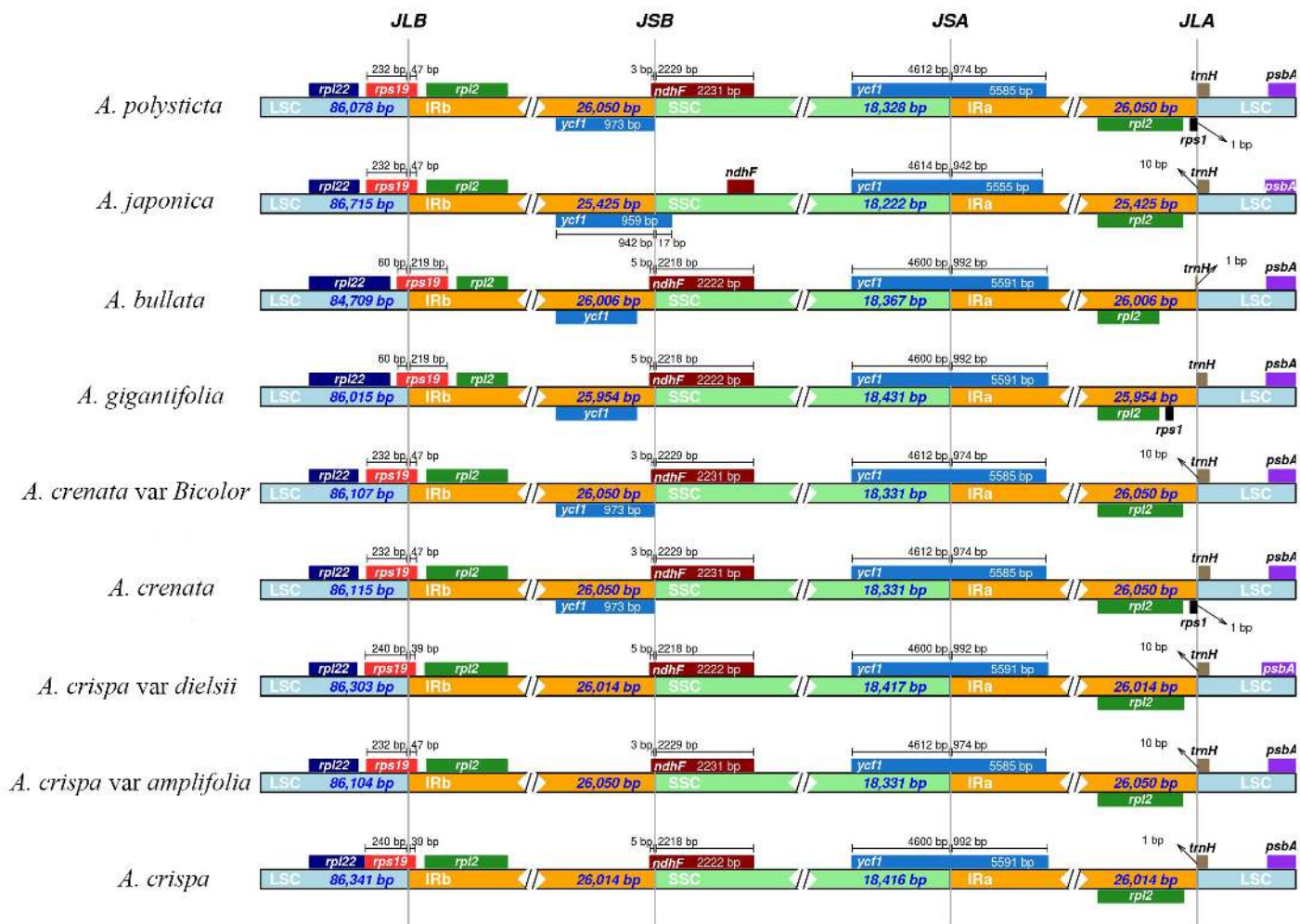


Figure 4

Changes of IR/SC boundary of chloroplast genomes of nine *Ardisia* species

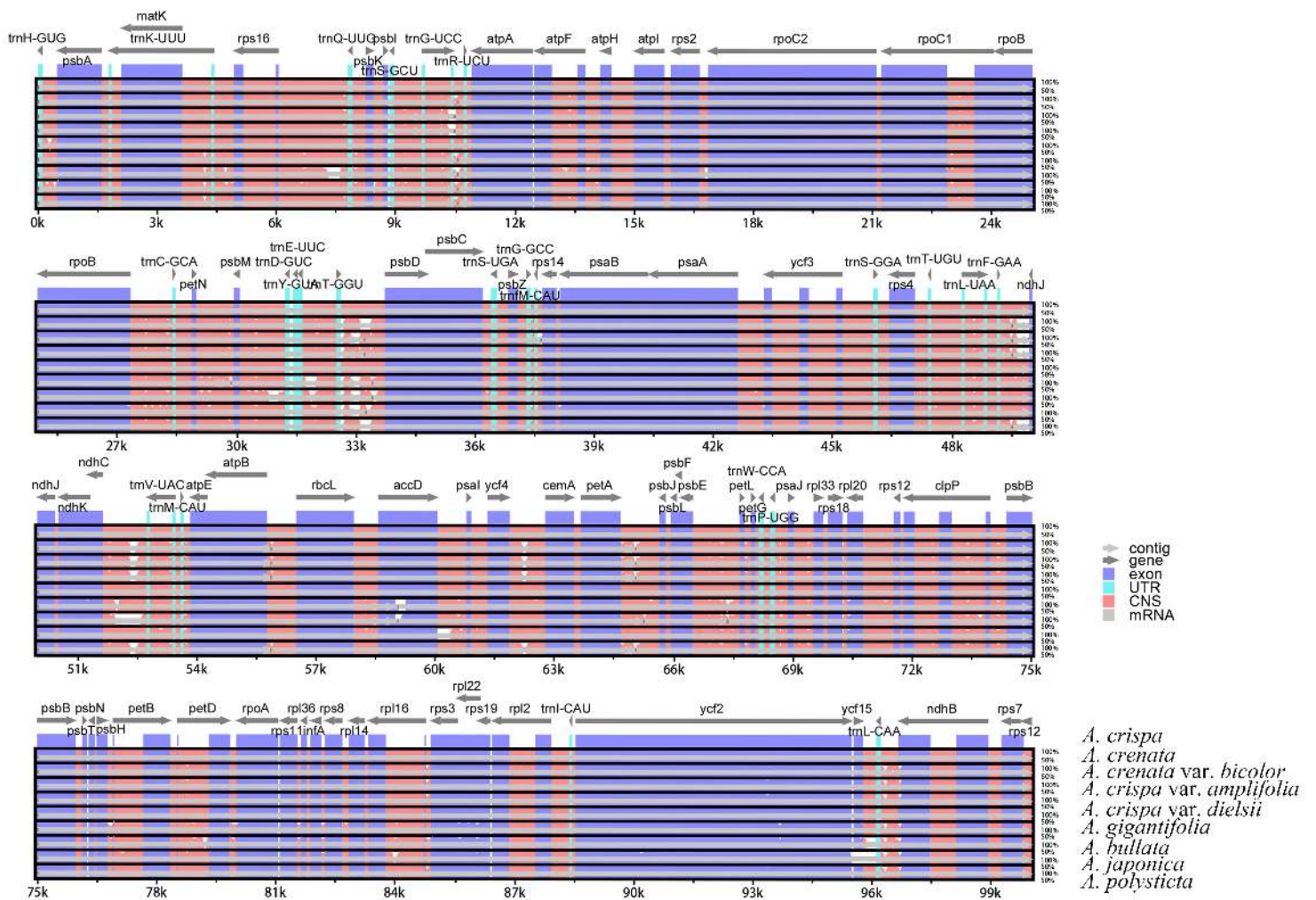


Figure 5

Global alignment analysis of the nine *Ardisia* chloroplast genomes

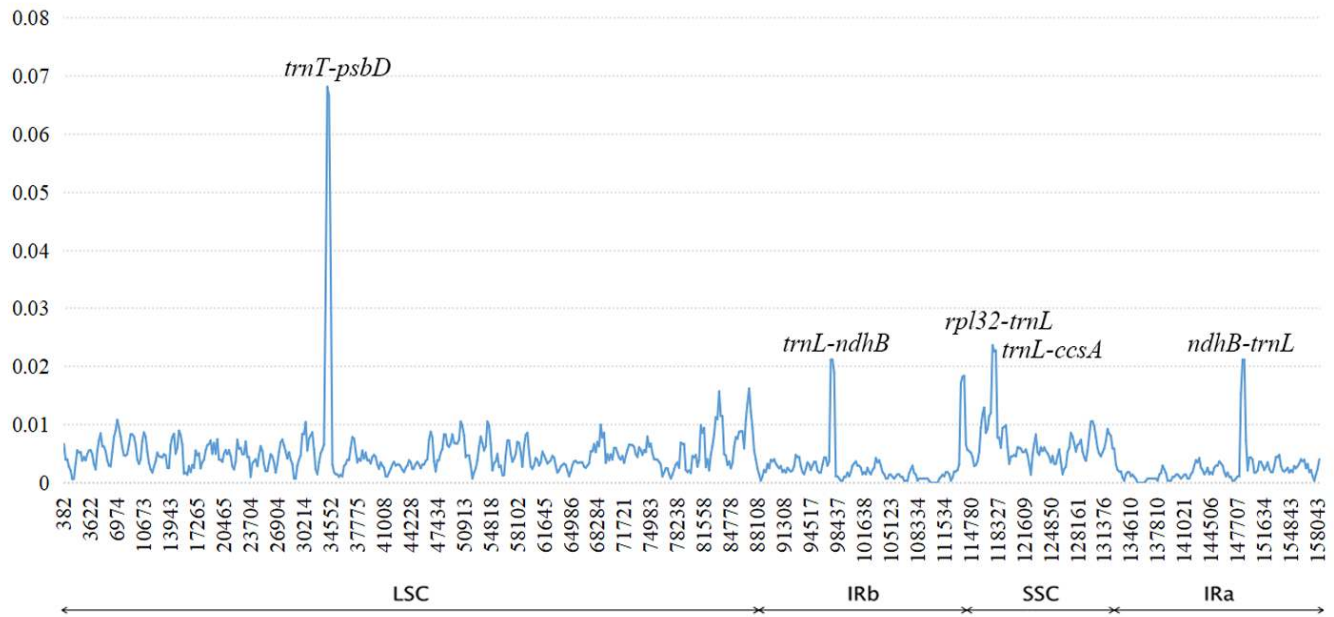


Figure 6

Comparative analysis of nucleotide diversity (Pi) values among the nine *Ardisia* species chloroplast genome sequences.

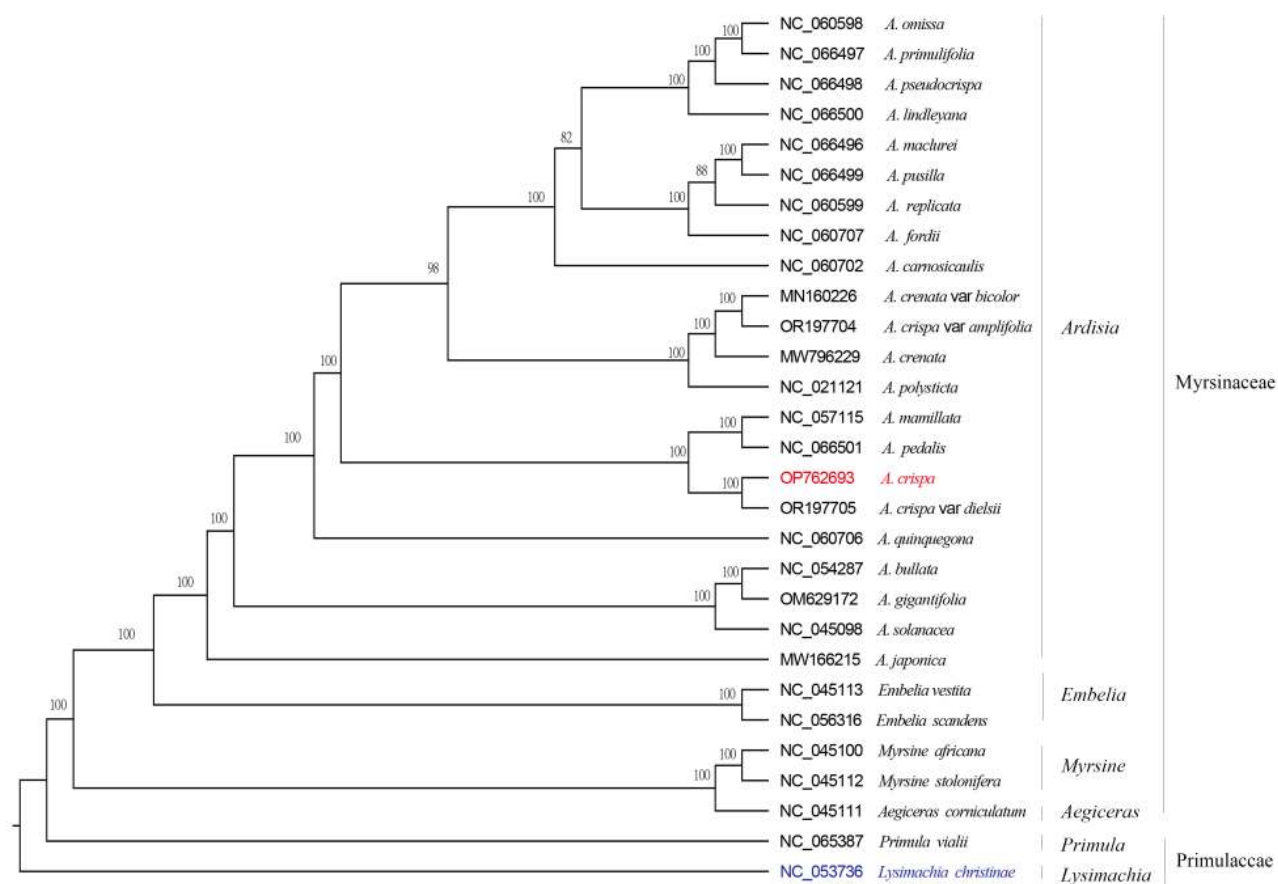


Figure 7

Phylogenetic analysis based on chloroplast genome sequences by BI tree



Figure 8

The plant of *A. crispa* and its growing environment (A: Growing environment B: Living plants C: Flowers)