Postprint: Structural Characteristics of Chloroplast Genomes and Phylogenetic Analysis of the Genus Aucuba

Authors: Li Juan, Tong Jiayun, Fan Zhichao, Yi Tong

Date: 2023-07-13T00:00:00+00:00

Abstract

To determine the structural composition and sequence variation of chloroplast genomes in Aucuba plants and to reveal interspecific phylogenetic relationships within the genus, this study performed next-generation sequencing on six Aucuba species including Aucuba chinensis and Aucuba japonica var. variegata, as well as the Garrya species Garrya buxifolia. The chloroplast genome sequences were assembled and annotated using bioinformatics software, followed by basic characteristic analysis, sequence comparison, and phylogenetic analysis. The results showed that: (1) The chloroplast genomes of Aucuba plants exhibit a typical circular quadripartite structure, with the six sequences ranging from 157,891 to 158,325 bp in total length, all encoding 114 genes including 80 protein-coding genes, 30 tRNA genes, and 4 rRNA genes; (2) The chloroplast genomes of the six species each contain 29 high-frequency codons, showing a preference for A/U-ending codons, and a total of 100 optimal codons were identified across the six sequences, including 12 shared optimal codons; (3) A total of 270 dispersed repeat sequences, 133 tandem repeat sequences, and 412 SSR loci were detected in the six chloroplast genome sequences; (4) Comparative genomics analysis revealed that the chloroplast genome sequences of this genus are highly conserved; (5) Ten highly variable regions were identified from the chloroplast genomes; (6) Phylogenetic analysis supported Aucuba as a highly supported monophyletic group that is closely related to the genus Garrya. The chloroplast genomes of five Aucuba species and one Garrya species examined in this study were sequenced and assembled for the first time, revealing the phylogenetic relationships within Aucuba and among its species, and providing reference data for the taxonomic identification and phylogenetic studies of Aucuba plants.

Full Text

Preamble

Structural Characteristics and Phylogenetic Analysis of Chloroplast Genomes of *Aucuba* Plants

Juan Li, Jiayun Tong, Zhichao Fan, Yi Tong*

School of Pharmaceutical Sciences, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

Abstract: To determine the structural composition and sequence differences of chloroplast genomes in the genus *Aucuba* and to reveal interspecific relationships within the genus, this study performed next-generation sequencing on six *Aucuba* species including *A. chinensis* and *A. japonica* var. *variegata*, as well as *Garrya buxifolia* from the genus *Garrya*. Bioinformatics software was used to assemble and annotate the chloroplast genome sequences, followed by analysis of basic characteristics, sequence comparison, and phylogenetic analysis. The results showed that: (1) The chloroplast genomes of *Aucuba* plants exhibited a typical circular quadripartite structure, with the six sequences ranging from 157,891 to 158,325 bp in length, all encoding 114 genes including 80 protein-coding genes, 30 tRNA genes, and 4 rRNA genes; (2) All six species had 29 high-frequency codons, showing a preference for A/U endings, and a total of 100 optimal codons were identified, including 12 shared optimal codons; (3) A total of 270 dispersed repeat sequences, 133 tandem repeat sequences, and 412 SSR loci were detected across the six chloroplast genomes; (4) Comparative genomic analysis revealed highly conserved chloroplast genome sequences within the genus; (5) Ten highly variable fragments were screened from the chloroplast genomes; (6) Phylogenetic analysis supported *Aucuba* as a highly supported monophyletic group closely related to *Garrya*. The chloroplast genomes of five *Aucuba* species and one *Garrya* species examined in this study were sequenced and assembled for the first time, revealing the phylogenetic relationships within *Aucuba* and providing reference materials for the taxonomic identification and phylogenetic studies of the genus.

Keywords: *Aucuba*, chloroplast genome, sequence variation, repeat sequences, codon usage bias, phylogenetic analysis

---

Introduction

The genus *Aucuba* Thunb., belonging to the family Garryaceae (Angiosperm Phylogeny Group et al., 2016), comprises evergreen small trees or shrubs. These plants maintain green foliage year-round and produce red fruits in winter, making them excellent ornamental species for garden landscaping (Xiang & Boufford, 2005). Some species in this genus are used in folk medicine, with roots, leaves, and fruits utilized for clearing heat and detoxifying, dispelling wind and

dampness, and promoting blood circulation to remove blood stasis (Editorial Committee of Chinese Materia Medica of the State Administration of Traditional Chinese Medicine, 1997; Jiang, 2005; Ai, 2013; Nanjing University of Chinese Medicine, 2006). Aucubin (AU), first discovered from *Aucuba* plants, exhibits extensive pharmacological activities including antioxidant, anti-aging, anti-inflammatory, and hepatoprotective effects (Zeng et al., 2020).

However, fundamental taxonomic revision and species delimitation of *Aucuba* remain incomplete, which directly impacts future chemical and pharmacological research on the genus and creates difficulties in developing, promoting, and guiding the production of new medicinal *Aucuba* resources. Although *Aucuba* can be easily distinguished from other groups by its flowers, fruits, or overall plant morphology, species identification within the genus is challenging due to complex and variable morphological characteristics (such as leaf shape, margin, and indumentum), lack of effective diagnostic traits for interspecific differentiation, and potential widespread hybridization and polyploidization events. Consequently, species delimitation within the genus is highly problematic, and a comprehensive and systematic interspecific phylogenetic framework urgently needs to be established (Xiang & Boufford, 2005).

During plant evolution, chloroplast genome structure and sequences remain relatively conserved, with consistent numbers, structures, compositions, and arrangements of coding genes, rarely undergoing recombination and other variations. This makes chloroplast genomes widely applicable in plant species identification, phylogenetic studies, and research on species origin (Liu & Huang, 2020). Currently, the National Center for Biotechnology Information (NCBI) contains seven chloroplast genome sequences of *Aucuba* plants. Based on 68 protein-coding genes from these seven chloroplast genomes, Huang et al. (2022) conducted phylogenetic analysis of *Aucuba*, supporting the monophyly of Garryales, Garryaceae, and the genus *Aucuba*.

In this study, using shallow genome sequencing technology and the published chloroplast genome of *A. japonica* (accession NC_{058874}.1) as a reference, we assembled and annotated the chloroplast genomes of *A. japonica* var. *variegata*, *A. omeiensis*, *A. chinensis*, *A. confertiflora*, *A. filicauda*, *A. albopunctifolia* var. *angustula*, and *Garrya buxifolia*. Through comparative sequence analysis and phylogenetic analysis of *Aucuba* chloroplast genomes, we aimed to address: (1) What are the structural characteristics of chloroplast genome sequences in six *Aucuba* species? (2) What are the differences among these sequences? (3) What is the phylogenetic position of *Aucuba*, and what are the relationships among the six *Aucuba* species? This study provides chloroplast genome information as reference material for future research on taxonomic revision, phylogeny, biogeography, molecular identification, and species evolution of *Aucuba* plants.

1. Study Materials

Molecular materials of *Aucuba* plants were primarily collected from fresh leaves in the field across southwestern and southern China distribution areas. For each individual, 2-3 fresh healthy leaves were collected, surface stains and water were removed with wet paper towels and absorbent paper, leaves were cut into pieces and placed in tea bags, then rapidly dried with silica gel. All voucher specimens are preserved in the Herbarium of Guangzhou University of Chinese Medicine (GUCM). The authors identified and confirmed these six *Aucuba* species as *A. omeiensis*, *A. chinensis*, *A. confertiflora*, *A. japonica* var. *variegata*, *A. filicauda*, and *A. albopunctifolia* var. *angustula* (Figure 1 [Figure 1: see original paper]). Dried leaves of *Garrya buxifolia* were obtained from cultivated plants at the UC Botanical Garden (molecular specimen collection number SZ5417). The complete chloroplast genome sequence of *A. japonica* (accession NC_{058874}.1) was downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/) and used as a reference for chloroplast genome assembly and annotation. All newly sequenced chloroplast genome sequences and annotation information have been uploaded to the NCBI database with assigned accession numbers. Collection information and GenBank accession numbers are provided in Table 1 .

Table 1 Collected information and GenBank accession numbers of six *Aucuba* plants

Plant name	Voucher specimen	Collection location	GenBank accession number
Aucuba omeiensis	Li Juan LJ20210329026	Emei Mountain, Emei, Sichuan	OQ348516
A. chinensis	Li Yuling LYL264	Xiangtou Mountain, Huizhou, Guangdong	OQ348513
A. confertiflora	Tong Yi TY20081412	Laohutiao Valley, Napo, Guangxi	OQ348514
A. japonica var. *variegata*	Tong Yi TY20080506	Guilin Botanical Garden, Guilin, Guangxi	OQ348512
A. filicauda	Tong Yi TY20080903	Hongtan Waterfall, Longsheng, Guangxi	OQ348515
A. albopunctifolia var. *angustula*	Li Juan LJ20210329002	Emei Mountain, Emei, Sichuan	OQ362998

Figure 1 Pictures of six *Aucuba* plants. A. *Aucuba omeiensis*; B. *A. chinensis*; C. *A. confertiflora*; D. *A. japonica* var. *variegata*; E. *A. filicauda*; F. *A. albopunctifolia* var. *angustula*.

2.1 Total Genomic DNA Extraction and Sequencing

Total DNA of *Aucuba* plants was extracted from silica-dried leaves using a modified CTAB method (Doyle & Doyle, 1987). After extraction, DNA quality and concentration were assessed using a B-500 ultramicro spectrophotometer (Shanghai Yuanxi Instrument Co., Ltd.) and agarose gel electrophoresis. Qualified DNA samples were sent to Wuhan Huada Gene Technology Co., Ltd. for second-generation sequencing using the DNBSEQ platform, yielding 3 GB of clean data per sample.

2.2 Assembly, Annotation, and Physical Mapping of Complete Chloroplast Genomes

Chloroplast genome assembly was performed using GetOrganelle-1.7.3.5 (Jin et al., 2020) with k-mer values set to 65, 105, and 127, thread count set to 24, and other parameters using default settings. The final assembly command was: `get_{{{organelle}}_{{from}}}_{reads}.py -1 sample_1.fastq.gz -2 sample_2.fastq.gz -F embplant_{pt} output-plastome -R 10 -t 24 -k 65,105,127`. The assembled chloroplast genomes were annotated using PGA-master (Qu et al., 2019) with the *A. japonica* chloroplast genome as reference, and annotation results were manually corrected using Geneious R 9.0.2 (Basic, 2012). Physical maps of chloroplast genomes were drawn using the online tool OGDraw (Lohse et al., 2013).

2.3 Chloroplast Genome Characteristic Analysis

Geneious R 9.0.2 was used to compile information on the length, GC content of each region, and gene annotation results for the complete chloroplast genome sequences, two single-copy regions, and one pair of inverted repeat regions of the six *Aucuba* species.

2.4 Codon Usage Bias and Optimal Codon Analysis

CDS sequences from the chloroplast genomes of six *Aucuba* species were screened, removing duplicate genes and genes shorter than 300 bp. Sequences contained only A, T, C, and G bases, each with a start codon (ATG) and stop

codon (TAG, TGA, or TAA), with no internal stop codons. Each chloroplast genome yielded 52 qualified CDS sequences. CodonW (Peden, 2005) was used to calculate relative synonymous codon usage (RSCU) and effective number of codons (ENC). Data were organized in Excel and visualized as heatmaps using TBtools (Chen et al., 2020). Codons with RSCU > 1 were identified as high-frequency codons (Wang et al., 2018). Codons meeting both high-frequency and high-expression criteria were designated as optimal codons (Sharp & Li, 1987). High-expression genes have lower ENC values, while low-expression genes have higher ENC values. Genes were ranked by ENC value, and the top and bottom 10% were selected as high-expression and low-expression gene sets, respectively. The ΔRSCU (RSCU of high-expression genes minus RSCU of low-expression genes) was calculated for each codon; codons with ΔRSCU > 0.08 were defined as high-expression codons. Optimal codons were then identified as high-frequency codons (RSCU > 1) within the high-expression codon set (Sharp & Li, 1987).

2.5 Repeat Sequence Analysis

Dispersed repeats in chloroplast genomes were identified using the online software REPuter (Stefan et al., 2001), including forward repeats, reverse repeats, complement repeats, and palindromic repeats, with parameters set to repeat unit length ≥ 30 bp and Hamming distance of 3. Tandem repeats were detected using Tandem Repeats Finder (TRF) (Benson, 1999) with default parameters. SSR loci were identified using the Perl script from MISA (Beier et al., 2017) with threshold settings for repeat counts of mononucleotide to hexanucleotide repeats set to 10, 5, 4, 3, 3, and 3, respectively.

2.6 Analysis of IR Boundary Expansion and Contraction

The online software IRscope (Ali et al., 2018) was used to visualize IR boundary expansion and contraction in the six sequences.

2.7 Sequence Comparative Analysis

Using *A. omeiensis* as the reference sequence, the online software mVISTA (Frazer et al., 2004) was employed to perform sequence identity comparisons of the six chloroplast genomes using Shuffle-LAGAN mode to generate visualizations of differences among the six chloroplast genomes.

2.8 Chloroplast Genome Collinearity Analysis

Collinearity refers to the conservation of gene number and order. The Mauve plugin in Geneious R 9.0.2 (Darling et al., 2004) was used to analyze collinearity among the six sequences to visualize consistency in gene arrangement and detect gene rearrangements and inversions within the genus.

2.9 Nucleotide Polymorphism Analysis

DnaSP v6.0 (Rozas et al., 2017) was used for sliding window analysis of the six sequences to detect hypervariable hotspot regions and their sizes, variable site numbers, and to calculate nucleotide diversity (Pi). Aligned sequences were imported into DnaSP v6.0 with parameters set to step size of 200 bp and window length of 600 bp. Fragments with Pi $>$ 0.01 and length $\geq$ 150 bp were selected as hypervariable regions of *Aucuba* chloroplast genomes, and their positions on the chloroplast genome were determined based on gene annotation results.

2.10 Phylogenetic Analysis

To reveal the phylogenetic relationships of *Aucuba*, nine chloroplast genome sequences were used to construct phylogenetic trees, including *Eucommia ulmoides* from Eucommiaceae in Garryales downloaded from NCBI (accession NC_{037948}) and *Garrya buxifolia* (accession OQ348517) assembled in this study as outgroups. Multiple sequence alignment was performed using the MAFFT plugin (Rozewicki et al., 2019) in Geneious R 9.0.2, with uneven ends trimmed before tree construction. Maximum likelihood (ML) and Bayesian inference (BI) methods were used to reconstruct phylogenetic relationships. ML trees were constructed using IQtree (Minh et al., 2020). JModeltest (Posada, 2020) was used to select the best-fit nucleotide substitution model based on the Bayesian Information Criterion (BIC), and MrBayes (Ronquist & Huelsenbeck, 2003) was used to construct BI trees.

3.1 Basic Characteristics of Chloroplast Genomes

The physical maps of the six *Aucuba* chloroplast genomes are shown in Figure 2. All six sequences are typical double-stranded circular DNA molecules with total lengths ranging from 157,891 to 158,325 bp and overall GC contents of 37.7%-37.8%. The LSC region lengths range from 87,210 to 87,563 bp with GC contents of 35.8%-35.9%; SSC region lengths range from 18,531 to 18,580 bp with GC contents of 31.5%-31.6%; IR region lengths range from 26,067 to 26,143 bp with GC contents of 43.0%-43.1% (Table 2).

Table 2 Chloroplast genomic characteristics of six *Aucuba* plants

Characteristic	*A. omeienensis*	*A. chinensis*	*A. confertiflora*	*A. japonica* var. *variegata*	*A. filicauda*	*A. albopunctifolia* var. *angustula*
Genome size (bp)	158,325	158,091	157,891	158,204	158,088	158,050
Length of LSC (bp)	87,563	87,419	87,210	87,494	87,419	87,419
GC content of LSC (%)	35.8	35.9	35.9	35.8	35.8	35.8
Length of SSC (bp)	18,580	18,531	18,531	18,580	18,531	18,531
GC content of SSC (%)	31.5	31.6	31.6	31.5	31.6	31.6
Length of IRs (bp)	26,091	26,071	26,075	26,065	26,069	26,050
GC content of IRs (%)	43.0	43.1	43.1	43.0	43.0	43.1
Total GC content (%)	37.7	37.8	37.8	37.7	37.7	37.7
Total number of genes	114	114	114	114	114	114

Characteristic	*A. omeienensis*	*A. chinensis*	*A. confertiflora*	*A. japonica* var. *variegata*	*A. filicauda*	*A. albopunctifolia* var. *angustula*
Number of protein-coding genes	80	80	80	80	80	80
Number of tRNA genes	30	30	30	30	30	30
Number of rRNA genes	4	4	4	4	4	4

All six chloroplast genomes encoded 114 genes, including 80 protein-coding genes, 30 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes (Table 3).

Table 3 Chloroplast genome composition of six *Aucuba* plants

Gene function	Gene category	Name of gene
Large subunit of ribosome	Ribosomal protein large subunit	*rpl2*a, *$×2*, *rpl14*, *rpl16* *, *rpl20*, *rpl22*, *rpl23* *a×2, *rpl32*, *rpl33*, *rpl36* *||Smallsubunitofribosome|Ribosomalproteinsmallsub *rps2*, *rps3*, *rps4*, *rps7* *a×2, *rps8*, *rps11*, *rps12* *a, b, **×2*, *rps14*, *rps15*, *rps16* *, *rps18*, *rps19 * *a||DNA − dependentRNApolymerase|RNApolymerasesubunit|* *rpoA*, *rpoB*, *rpoC1* *, *rpoC2 * *||RibosomalRNA|RibosomalRNA|* *rrn4.5 * a×2, *rrn5 * *a×2, *rrn16 * *a×2, *rrn23 * *a×2||TransferRNA|TransferRNA|* *trnA − UGC * *a, *×2*, *trnC − GCA*, *trnD − GUC*, *trnE − UUC*, *trnF − GAA*, *trnfM − CAU*, *trnG − UCC * *, *trnH − GUG*, *trnI − CAU * *a×2, *trnI − GUA * *a, *×2*, *trnK − UUU * *, *trnL − CAA * a×2, *trnL − UAA * *, *trnL − UAG*, *trnM − CAU*, *trnN − GUU * a×2, *trnP − GGG*, *trnQ − UUG*, *trnR − ACG * a×2, *trnR − UCU*, *trnS − GCU*, *trnS − GGA*, *trnS − UGA*, *trnT − GGU*, *trnT − UGU*, *trnV − GAC * a×2, *trnV − UAC * *, *trnW − CCA*, *trnY − GUA*

Gene function	Gene category	Name of gene

Note: a indicates genes in inverted repeat regions; b indicates trans-splicing genes; indicates genes containing one intron; ** indicates genes containing two introns; $×$2 indicates duplicated genes.*

3.2 Codon Usage Bias

The relative synonymous codon usage (RSCU) of CDS in the six *Aucuba* chloroplast genomes is shown in Figure 3. RSCU > 1 indicates preferred codons used more frequently; RSCU = 1 indicates no codon usage preference; RSCU < 1 indicates less frequently used codons. Stop codons UAA, UGA, and UAG, as well as the unique codons for tryptophan (UGG) and methionine (AUG), were excluded from analysis as they show no preference. The results revealed that all six sequences contained 59 synonymous codons, with 29 high-frequency codons (RSCU > 1) each. Among these, 28 codons ended with A or U, and one ended with G, indicating that *Aucuba* chloroplast genomes prefer codons ending with A or U.

A total of 100 optimal codons were screened from the six sequences (Table 4), with all but the shared UUG (ending with G) among *A. albopunctifolia* var. *angustula*, *A. japonica* var. *variegata*, *A. filicauda*, and *A. confertiflora* ending with A/U. Twelve optimal codons were shared across all six sequences: AAA, ACU, AGU, CAA, CCU, CGU, GAA, GCU, GGU, GUU, UCU, and UGU. *A. omeiensis*, *A. japonica* var. *variegata*, and *A. confertiflora* each had species-specific optimal codons: CAU, CUU, and GAU, respectively.

3.3 Repeat Sequence and SSR Analysis

REPuter identified a total of 270 dispersed repeat sequences across the six sequences, including 133 forward repeats, 8 reverse repeats, 2 complement repeats, and 127 palindromic repeats (Table 5). Forward repeats were most abundant, followed by palindromic repeats, while reverse and complement repeats were rare, with only one complement repeat detected in *A. chinensis* and *A. albopunctifolia* var. *angustula*, respectively, and none in the other four species.

TRF detected 133 tandem repeats in total across the six sequences, with *A. confertiflora* having the fewest (20) and *A. japonica* var. *variegata* the most (25). Tandem repeats were more abundant in the LSC and IR regions than in the SSC region.

MISA detected 412 SSR loci across the six sequences, including mononucleotide SSRs (367), dinucleotide SSRs (26), trinucleotide SSRs (9), and tetranucleotide

SSRs (10). Mononucleotide A/T repeats predominated, accounting for 86.88%-99.41% of all SSR loci. *A. filicauda* and *A. japonica* var. *variegata* had the most SSRs (73 each), while *A. chinensis* had the fewest (61) (Table 7).

3.4 IR/SC Boundary Expansion and Contraction Analysis

The two IR regions of *Aucuba* chloroplast genomes have four boundaries between LSC and SSC: LSC/IRB, IRB/SSC, SSC/IRA, and IRA/LSC (Figure 4 [Figure 4: see original paper]). These boundaries were relatively conserved but showed minor interspecific differences. The LSC/IRB boundary was located within the *rps19* gene, which extended 33 bp into the IRB region in all six sequences. The SSC/IRB boundary was located within the *ndhF* gene, extending 42 bp into the IRB region in all sequences. The SSC/IRA boundary was located within the *ycf1* gene, extending 1,082 bp into the IRA region in *A. confertiflora* and *A. japonica* var. *variegata*, 1,079 bp in *A. filicauda*, *A. albopunctifolia* var. *angustula*, and *A. omeiensis*, and 1,061 bp in *A. chinensis*. The LSC/IRA boundary was near *trnH*, with a distance of 14 bp in *A. omeiensis* and *A. confertiflora*, and 7 bp in the other four species.

3.5 Sequence Variation Analysis

To compare interspecific differences in chloroplast genome sequences within *Aucuba*, the six sequences were globally aligned using *A. omeiensis* as reference. mVISTA plots revealed high similarity among the six *Aucuba* sequences, with single-copy regions being more conserved than inverted repeat regions and coding regions being more conserved than non-coding sequences (Figure 5 [Figure 5: see original paper]). The *trnC-GCA-petN* region in the LSC region and the *rps7-trnV-GAC* region in the IR region showed lower similarity with varying degrees of variation. rRNA and tRNA gene regions exhibited the highest consistency and were most conserved. The LSC and SSC regions showed more variation than the IR region, indicating that the IR region is more conserved during evolution.

3.6 Chloroplast Genome Collinearity Analysis

Multiple genome alignment detected only one locally collinear block (LCB) among the six sequences (Figure 6 [Figure 6: see original paper]), indicating that gene types, numbers, and arrangement orders are highly consistent within the genus. The chloroplast genomes were completely collinear with no rearrangement or recombination events, further demonstrating the high conservation of chloroplast genomes in this genus.

3.7 Sequence Variation Hotspot Analysis

Sliding window analysis revealed nucleotide diversity (Pi) values ranging from 0 to 0.01889 among the six sequences, with an average Pi of 0.00351 (Figure 7 [Figure 7: see original paper]), indicating high similarity and strong conservation. The IR region showed significantly lower nucleotide diversity (Pi = 0.00085) than the LSC (Pi = 0.00447) and SSC (Pi = 0.00614) regions, with the SSC region showing the highest diversity. Ten hypervariable fragments were identified (Table 8), including seven in the LSC region (*rps16*, *rps16-trnQ-UUG*, *rpoB-trnC-GCA*, *petN-psbM*, *trnC-GCA-petN*, *psbM-trnD-GUC*, *accD-psaI*) and three in the SSC region (*ndhE*, *ndhE-ndhG*, *ycf1*), with *ycf1* showing the highest divergence. Only *rps16*, *ndhE*, and *ycf1* are coding sequences; the other hypervariable fragments are located in intergenic spacers (IGS). These hypervariable regions can serve as candidate DNA barcodes for species identification in *Aucuba*.

The three universal DNA barcode fragments widely used in plant identification (*matK*, *rbcL*, and *trnH-GUG-psbA*) showed low Pi values in *Aucuba* chloroplast genomes and are therefore unsuitable as molecular markers for this genus.

3.8 Phylogenetic Analysis

The best-fit nucleotide substitution model selected by JModeltest using the BIC method was GTR+G for all datasets. BI and ML trees were constructed using MrBayes and IQtree, respectively (Figure 8 [Figure 8: see original paper]). Phylogenetic analysis showed that BI and ML trees based on complete chloroplast genomes exhibited identical topologies. All *Aucuba* species formed a highly supported monophyletic group (BS = 100%, pp = 1) as sister to *Garrya*, together constituting Garryaceae. Within *Aucuba*, two clades were resolved: Clade I (BS = 100%, pp = 1) comprising *A. omeiensis*, *A. confertiflora*, and *A. chinensis*; and Clade II (BS = 100%, pp = 1) comprising *A. japonica* var. *variegata*, *A. japonica*, *A. filicauda*, and *A. albopunctifolia* var. *angustula*.

4.1 Structural Characteristics of Chloroplast Genomes

This study assembled and annotated chloroplast genomes of six *Aucuba* species and conducted comparative sequence analysis. The results showed that the six sequences are highly similar in genome structure and size, gene content, and composition. *Aucuba* chloroplast genomes possess a typical circular quadripartite structure, with total lengths of 157,891-158,325 bp, encoding 114 genes including 80 protein-coding genes, 30 tRNA genes, and 4 rRNA genes, with total GC contents of 37.7%-37.8%. The GC contents of LSC, SSC, and IR regions are 35.8%-35.9%, 31.5%-31.6%, and 43.0%-43.1%, respectively. Like most

angiosperms, the IR region of *Aucuba* chloroplast genomes is more stable than LSC and SSC regions, with the highest GC content.

Codons are crucial for correct genetic information expression. Codon usage bias is specific to different species and even different genes within a species, resulting from combined effects of selection, mutation, and drift during long-term evolution. Closely related species or those in similar environments are more likely to adopt similar codon selection strategies (Romero et al., 2000; Xu et al., 2011). Comparing codon usage bias differences can infer whether genes are under different translational selection pressures, which is important for exploring evolutionary patterns in *Aucuba*. All six sequences contained 59 synonymous codons, with 29 high-frequency codons (RSCU > 1) each. Except for UUG ending with G, the other 28 high-frequency codons ended with A or U, indicating a preference for A/U-ending codons. A total of 100 optimal codons were identified, including 12 shared optimal codons (AAA, ACU, AGU, CAA, CCU, CGU, GAA, GCU, GGU, GUU, UCU, UGU), all ending with A/U, especially U, consistent with findings that dicot plants prefer A/U-ending codons (Kawabe & Miyashita, 2003).

The presence and abundance of chloroplast repeat sequences can increase genetic diversity and are associated with various phylogenetic signals (Adeyemo et al., 2021). SSRs and repeat sequences are widely distributed in chloroplast genomes with rich polymorphic sites, offering advantages such as easy replication and high genetic information content, making them highly reliable for studying genetic diversity, relationships, cultivar identification, and marker-assisted breeding (Adeyemo et al., 2021; Jia et al., 2022). The six chloroplast genomes contained 412 SSR loci, predominantly mononucleotide A/T repeats mainly located in the LSC region, along with 270 dispersed repeats and 133 tandem repeats. The functions of these repeat sequences in *Aucuba* chloroplast genomes require further investigation.

4.2 Comparative Analysis of Chloroplast Genome Structure

Comparative genomic analysis revealed high similarity among *Aucuba* chloroplast genomes, with single-copy regions being more conserved than inverted repeat regions and coding regions more conserved than non-coding regions. The gene types at IR/SC boundaries were identical across species. Gene structure and arrangement were highly similar overall, with complete collinearity and no rearrangement or recombination events. Based on our previous research (unpublished), currently used DNA barcode fragments (*psbA-trnH*, *rbcL*, *matK*, ITS, ITS2) showed insufficient variation for species resolution in *Aucuba*. This study identified ten hypervariable fragments with high variation rates and appropriate lengths from the LSC and SSC regions. These hypervariable regions can serve as potential DNA barcodes for *Aucuba* species identification and, when combined with biparentally inherited nuclear gene fragments, can provide reliable

molecular markers for species identification, hybrid origin, polyploid formation, and phylogenetic analysis.

4.3 Phylogenetic Analysis

Phylogenetic analysis supported *Aucuba* as a highly supported monophyletic group (BS = 100%, pp = 1) closely related to *Garrya*, together forming Garryaceae. Garryaceae is closely related to Eucommiaceae, consistent with APG IV and Huang et al. (2022).

Within *Aucuba*, two highly supported clades were resolved. Clade I comprises *A. omeiensis*, *A. confertiflora*, and *A. chinensis*, characterized as trees with green flowers. Clade II comprises *A. japonica* var. *variegata*, *A. japonica*, *A. filicauda*, and *A. albopunctifolia* var. *angustula*, characterized as shrubs with red flowers. Whether these characteristics represent stable synapomorphies for the two clades requires support from more species and populations within the genus.

Within Clade II, *A. japonica* var. *variegata* and *A. japonica* formed a subclade with close relationship. Comparison of their complete chloroplast genomes revealed identical gene content and positions with only four base variations in intergenic regions, supporting Ranney's (2018) view that spotted foliage in *Aucuba* is likely controlled by nuclear genes with multiple alleles and quantitative inheritance rather than strict maternal inheritance. Therefore, the "variegated leaf" phenotype is not a stable reliable trait, and taxonomic divisions based on this characteristic should be approached with caution.

Xiang & Boufford (2005) synonymized *A. omeiensis* under *A. chinensis* in Flora of China. Our phylogenetic analysis showed *A. omeiensis* is more closely related to *A. confertiflora*, and these two species together form a sister group to *A. chinensis*. Ranney et al. (2018) studied relative genome sizes in *Aucuba* using flow cytometry and found that *A. omeiensis* had a significantly higher 1Cx value (10.6 pg) than other species including *A. chinensis* (7.37 pg) (6.8-7.5 pg), indicating *A. omeiensis* underwent more significant polyploidy-independent genome expansion. Combined with morphological analysis of specimens and field characters, these three species form a species complex with complicated relationships. Therefore, resolution of species delimitation for *A. omeiensis* requires integration of more morphological features, especially cytological evidence, and analysis of nuclear gene data, particularly collinear single-copy nuclear genes.

Aucuba likely experiences widespread hybridization and polyploidization events, while traditional leaf morphological characters used for species identification show extreme variation within the genus, resulting in transitional variation and overlapping types among species with few stable diagnostic features. Many species are difficult to identify accurately, and our understanding of morphological characteristics remains insufficient. Consequently, species with uncertain

identification or unreasonable delimitation were not included in this study. A comprehensive infrageneric classification system and resolution of interspecific relationships in *Aucuba* will require broader sampling, more character evidence (especially chromosome data), and combined analysis of more variable single-copy nuclear genes and chloroplast genomes to clarify taxonomic issues and provide a foundation for future research on medicinal and horticultural uses of *Aucuba* plants.

References

ADEYEMO OA, AYODELE OO, AJISAFE MO, et al., 2021. Evaluation of dark jute SSR markers and morphological traits in genetic diversity assessment of jute mallow (*Corchorus olitorius* L.) cultivars[J]. S Afr J Bot, 137: 290-297.

AI TM, 2013. Medicinal Flora of China[M]. Beijing: Peking University Medicinal Press, 7: 279-287.

ALI AMIRYOUSEFI, JAAKKO HYVÖNEN, et al., 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes[J]. Bioinformatics, 17(34): 3030-3031.

Angiosperm Phylogeny Group, CHASE MW, CHRISTENHUSZ MJM, et al., 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV[J]. Bot J Linn Soc, 181(1): 1-20.

BASIC G, 2012. An integrated and extendable desktop software platform for the organization and analysis of sequence data[J]. Bioinformatics, 28(12): 1647-1649.

BEIER S, THIEL T, MÜNCH T, et al., 2017. MISA-web: a web server for microsatellite prediction[J]. Bioinformatics, 33(16): 2583-2585.

BENSON G, 1999. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucl Acid Res, 27(2): 573-580.

CHEN C, CHEN H, ZHANG Y, et al., 2020. TBtools: an integrative toolkit developed for interactive analyses of big biological data[J]. Mol Plant, 13(8): 1194-1202.

DARLING ACE, MAU B, BLATTNER FR, et al., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements[J]. Genome Res, 14(7): 1394-1403.

DOYLE JJ, DOYLE JL, 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue[J]. Phytochem Bull, 19: 11-15.

Editorial Committee of Chinese Materia Medica of the State Administration of Traditional Chinese Medicine, 1997. Chinese materia medica[M]. Shanghai:

Science and Technology Press, 5: 734-737.

FRAZER KA, PACHTER L, POLIAKOV A, et al., 2004. VISTA: computational tools for comparative genomics[J]. Nucl Acid Res, 32(2): W273-W279.

HUANG Y, FAN L, HUANG J, et al., 2022. Plastome phylogenomics of *Aucuba* (Garryaceae)[J]. Front in Genet, 13: 753719.

JIA SN, ZHANG YM, ZHAO GF, et al., 2022. Comparative analysis of the chloroplast genomes of *Rhododendron przewalskii* and *Rhododendron* plants[J/OL]. Guihaia: 1-17.

JIANG JW, 2005. A Dictionary of Medicinal Plants[M]. Tianjin: Science and Technology: 92.

JIN JJ, YU WB, YANG JB, et al., 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes[J]. Genome Biol, 21(1): 1-31.

KAWABE A, MIYASHITA NT, 2003. Patterns of codon usage bias in three dicot and four monocot plant species[J]. Genes Genet Syst, 78(5): 343-352.

LIU C, HUANG LF, 2020. Application of chloroplast genome in plant systematics and phylogeography[J]. Chin J Plant Ecol, 44(5): 495-509.

LOHSE M, DRECHSEL O, KAHLAU S, et al., 2013. OrganellarGenomeDRAW: a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets[J]. Nucl Acid Res, 41(W1): W575-W581.

MINH BQ, SCHMIDT HA, CHERNOMOR O, et al., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era[J]. Mol Biol Evol, 37(5): 1530-1534.

Nanjing University of traditional Chinese Medicine, 2006. Traditional Chinese medicine dictionary[M]. Shanghai: Science and Technology Press: 334.

PEDEN J, 2005. CodonW version 1.4.2[CP]. Nottingham, UK: University of Nottingham.

POSADA D, 2008. jModelTest: phylogenetic model averaging[J]. Mol Biol Evol, 25(7): 1253-1256.

QU XJ, MOORE MJ, LI DZ, et al., 2019. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes[J]. Plant Methods, 15(1): 1-12.

RANNEY TG, THOMASSON TH, NEILL K, et al., 2018. Ploidy, relative genome size, and inheritance of spotted foliage in *Aucuba* species (Garryaceae)[J]. HortScience, 53(9): 1324-1328.

ROMERO H, ZAVALA A, MUSTO H, 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern

of selective forces[J]. Nucl Acid Res, 28(10): 2084-2090.

RONQUIST F, HUELSENBECK JP, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models[J]. Bioinformatics, 19(12): 1572-1574.

ROZAS J, FERRER-MATA A, SÁNCHEZ-DELBARRIO JC, et al., 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets[J]. Mol Biol Evol, 34(12): 3299-3302.

ROZEWICKI J, LI S, AMADA KM, et al., 2019. MAFFT-DASH: integrated protein sequence and structural alignment[J]. Nucl Acid Res, 47(W1): W5-W10.

SHARP PM, Li WH, 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications[J]. Nucl Acid Res, 15(3): 1281-1295.

STEFAN K, CHOUDHURI JV, ENNO O, et al., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale[J]. Nucl Acid Res, 29(22): 4633-4642.

WANG L, XING H, YUAN Y, et al., 2018. Genome-wide analysis of codon usage bias in four sequenced cotton species[J]. PLoS ONE, 13(3): e0194372.

XIANG QY, BOUFFORD DE, 2005. Flora of China[M]. Beijing: Science Press; St. Louis: Missouri Botanical Garden Press: 222-226.

XU C, CAI X, CHEN Q, et al., 2011. Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium gower ramsey*[J]. Evol Proteins, 7: EBO. S8092.

ZENG X, GUO F, OUYANG D, 2020. A review of the pharmacology and toxicology of aucubin[J]. Fitoterapia, 140: 104443.

Note: Figure translations are in progress. See original paper for figures.