Analysis of the chloroplast genome of Malus baccata var. gracilis

Xin Qin
Shandong Agricultural University
Qiang Hao
China National Botanical Garden (North Garden)
Xun Wang
Shandong Agricultural University
Yangbo Liu
Shandong Agricultural University
Chen Yang
Shandong Agricultural University
Mengyi Sui
Shandong Agricultural University
Yawen Zhang
Shandong Agricultural University
Yanli Hu
Shandong Agricultural University
Xuesen Chen
Shandong Agricultural University
Zhiquan Mao
Shandong Agricultural University
Yunfei Mao
Shandong Agricultural University
Xiang Shen (hainandaousyd@163.com)
Shandong Agricultural University

Research Article

Keywords: Malus baccata var. gracilis, chloroplast genome, phylogeny, bioinformatics, comparative genomics analysis

Posted Date: October 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3421026/v1

Additional Declarations: No competing interests reported.

Abstract

Malus baccata var. gracilis has high ornamental value and breeding significance, and comparative chloroplast genome analysis was applied to facilitate genetic breeding for desired traits and resistance and provide insight into the phylogeny of this genus. Using data from whole-genome sequencing, a tetrameric chloroplast genome with a length of 159,992 bp and a total GC content of 36.56% was constructed. The *M. baccata var. gracilis* chloroplast genome consists of a large single-copy area (88,100 bp), a short single-copy region (19,186 bp), and two inverted repeat regions, IRa (26,353 bp) and IRb (26,353 bp). This chloroplast genome contains 112 annotated genes, including 79 protein-coding genes (nine multicopy), 29 tRNA genes (eight multicopy), and four rRNA genes (all multicopy). Calculating the relative synonymous codon usage revealed a total of 32 high-frequency codons, and the codons exhibited a biased usage pattern towards A/U as the ending nucleotide. Interspecific sequence comparison and boundary analysis revealed significant sequence variation in the vast single-copy region as well as generally similar expansion and contraction of the SSC and IR sections for ten analyzed Malus species. *M. baccata var. gracilis* and *Malus hupehensis* were grouped together into one branch based on the phylogenetic analysis of chloroplast genome sequences. The chloroplast genome of *Malus* species provides an important foundation for species identification, genetic diversity analysis, and *Malus* chloroplast genetic engineering. Additionally, the results can facilitate the use of pendant traits to improve apple tree shape.

Introduction

According to Dobránszki et al. (2014), China is the center of origin for the *Malus* genus. *Malus* species are dispersed throughout the north temperate zone in an erratic pattern of Central Asia-Europe and East Asia-North America. *Malus* species can be classified as cultivated or wild (Wang *et al*., 2018; Chen *et al*., 2021). *Malus* species that are grown for human use are more heavily impacted by human activity and are mostly used for raw and processed food, with some utilized as ornamentals (Zhou et al., 2019). Due to their favorable introgression and grafting affinity, wild species are exploited as pollinators, rootstock resources, and breeding material (He et al., 2020; Rong et al., 2022). As an important group of *Malus*, *M. baccata* spp. have high ornamental value with beautiful flowers and fruits, and these plants are also widely used in germplasm evaluation and resistance breeding (Ren et al., 2019). *M. baccata* spp. can be used as a parent to breed cold and black star disease resistant apples (Papp et al., 2020). Pendulousness is an important ornamental trait in woody plants and is used for apple tree shape improvement (Sestras *et al*., 2023). However, limited work has been done to determine the phylogeny and classify the pendulous bush resources, so genetic and genomic studies are needed to effectively conserve, develop, and exploit its germplasm diversity.

The chloroplast genome (cpDNA) contains rich genetic information; is moderate in size, sequence variation rate, and copy number; has a highly conserved genetic structure; is maternally inherited; has good stability in offspring, is not affected by genetic recombination; and evolved independently (Choi et al., 2017), making it suitable for studies of complex evolutionary systems and developmental status

among different species (Liao et al., 2021). Chloroplast genomes are increasingly used in phylogenetic research due to the recent advances of second-generation sequencing technology (Alkan *et al.*, 2011). The few published chloroplast genomes of apple plants include *Malus toringoides*, *Malus hupehensis*, *Malus floribunda*, and *Malus yunnanensis* (Li et al., 2020; Liu et al., 2012; Wick et al., 2015), and these data have greatly facilitated the detection of variation in the genus *Malus* and promote the utilization and conservation of genetic resources.

The goal of this study was to generate the cpDNA sequence of *M. baccata var. gracilis* cpDNA using whole genome sequencing and bioinformatics analysis. The fundamental characteristics, sequence similarity, and evolutionary relationships were further analyzed. The findings of this work can serve as a valuable resource to elucidate the genetic variation characteristics of *M. baccata var. gracilis* cpDNA and *Malus* cpDNA. Additionally, the outcomes of this study will have a positive impact on the preservation and advancement of wild species, such as *M. baccata var. gracilis*, and will facilitate the genetic breeding of drooping traits and the enhancement of apple tree morphology.

Results

Chloroplast genome characterization of *M. baccata var. gracilis*

The chloroplast entire genome of *M. baccata var. gracilis* was successfully acquired by assembly and splicing techniques, resulting in a total sequence length of 159,992 bp (Fig. 1). The genome sequence has a total GC content of 36.56%. The genome is comprised of three distinct regions: a large single-copy region (LSC), a short single-copy region (SSC), and a pair of inverted repeats (IRs, namely IRa and IRb) (Table 1). The lengths of the tetrad sequences were as follows: LSC, 88100 base pairs, spanning from position 1 to 88100; IRb, 26353 bp, spanning from position 88101 to 114453; SSC, 19186 bp, spanning from position 114454 to 133639; and IRA, 26353 bp, spanning from position 133640 to 159992. The GC content of the IRs in *M. baccata var. gracilis* is 42.70%, higher than the overall GC content of the whole chloroplast genome (36.56%).

Table 1
Regions of the *M. baccata var. gracilis* chloroplast genome

Region name	Start	End	Length(bp)	GC(%)
LSC	1	88,100	88,100	34.23
IRa	88,101	114,453	26,353	42.70
SSC	114,454	133,639	19,186	30.40
IRb	133,640	159,992	26,353	42.70

Whole-genome sequence annotation of the chloroplast of *M. baccata var. gracilis* yielded 112 genes (Table 2), including 79 unique protein-coding genes (nine genes were multiple copies), 29 tRNA genes

(eight were multiple copies), and four rRNA genes (four were multiple copies). The set of protein-encoding genes are from 15 distinct gene families. The genome contains 11 genes encoding NADH dehydrogenase subunits, five genes encoding photosystem I subunits, 15 genes encoding photosystem II subunits, six genes encoding cytochrome b/f complex subunits, six genes encoding ATP synthase subunits, one gene encoding the large subunit of ribulose 1,5-diphosphate carboxylase/oxygenase (*rbcL*), four genes encoding DNA-dependent RNA polymerase subunits, nine genes encoding ribosomal large subunit proteins, 12 genes encoding ribosomal small subunit proteins, one gene encoding a maturation enzyme (*matK*), one gene encoding a c-type cytochrome synthase (*ccsA*), one gene encoding a membrane protein (*cemA*), one gene encoding a protease (*clpP*), one gene encoding a subunit of acetyl-CoA-carboxylase (accD), and five genes encoding conserved open reading frames (*ycf1, ycf2, ycf3, ycf4, ycf15*).

Table 2
Coding genes of chloroplast genome in *M. baccata var. gracilis*

Groups of genes	Names of genes
Subunits of NADH-dehydrogenase	*ndhA, ndhB(×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Subunits of photosystem	*psaA, psaB, psaC, psaI, psaJ, psbA, psbB, psbC, psbD, psbE, psbF,*
Subunits of photosystem	*psbH, psbI, psbJ, psbK, psbL, psbM psbN, psbT, psbZ*
Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
Large subunit of rubisco	*rbcL*
Small subunit of ribosome	*rps2, rps3, rps4, rps7(×2), rps8, rps11 rps12(×2), rps14, rps15, rps16, rps18, rps19(×2)*
Large subunit of ribosome	*rpl2(×2), rpl14, rpl16, rpl20, rpl22, rpl23(×2), rpl32, rpl33, rpl36*
DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
rRNA genes	*rrn4.5S(×2), rrn5S(×2), rrn16S(×2) , rrn23S(×2)*
tRNA genes	*rnA-UGC(×2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-GCC, trnG-UCC, trnH-GUG, trnI-CAU(×2), trnI-GAU(×2), trnK-UUU, trnL-CAA(×2), trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU(×2), trnP-UGG, trnQ-UUG, trnR-ACG(×2), trnR-UCU, trnS-GCU(×2), trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC(×2), trnV-UAC, trnW-CCA, trnY-GUA*
Maturase	*matK*

Groups of genes	Names of genes
c-type cytochrome synthesis gene	*ccsA*
Envelope membrane protein	*cemA*
Protease	*clpP*
Subunit of Acetyl-CoA-carboxylase	*accD*
Genes of unknown functions Open Reading	*ycf1(×2), ycf2(×2), ycf3, ycf4, ycf15(×2)*

Simple repeat sequences

Simple sequence repeats (SSRs) or microsatellites are DNA sequences consisting of short, tandemly repeated di-, tri-, tetra-, penta-, or hexa-nucleotide motifs. As shown in Fig. 2, a total of 93 SSR loci were detected in the chloroplast genome of *M. baccata var. gracilis*, of which 68 were single nucleotide repeats, 18 were dinucleotide repeats, and seven were polynucleotide repeats. The monomeric and dimeric forms of SSRs constituted 92.47% of the overall SSRs. Thymidine (T) monomeric repeats constituted 57.35% of the total 68 monomeric SSRs and AT repeats were the predominant kind of dimeric SSRs, accounting for 94.44% of the total. The dispersed repetitive regions within the chloroplast genome were also analyzed. A total of 55 pairs of repeating sequences, each at least 30 in length, were identified. Among these pairings, 22 were found to be palindromic repeats, 27 were identified as forward repeats, one pair exhibited reverse repetitions, and none were complementary repeats. The largest observed palindromic repeat sequence spanned a length of 63 base pairs (bp), the biggest forward repeat sequence was found to be 40 bp in length, and the longest reverse repeat sequence was determined to be 45 bp long.

Codon preference analysis of *M. baccata var. gracilis*

Amino acids preferentially employ codons with a relative synonymous codon usage (RSCU) value greater than 1. According to Fig. 3, the chloroplast coding sequence of *M. baccata var. gracilis* contained 32 codons with RSCU values greater than 1, including UUA (Leu), GCU (Ala), AGA (Arg), ACU (Thr), and UAU (Tyr).

Leucine (Leu) appeared most frequently, followed by arginine (Arg) and serine (Ser). In leucine, the UUA codon was used preferentially, with an RSCU value of 2.04. The preferred codons for arginine and serine were AGA and UCU, and these exhibited a bias towards A/U in the third position. There was a preference for UAA for the termination codon.

Chloroplast genome alignment of *M. baccata var. gracilis* and other *Malus* species

To assess the similarities and differences of the chloroplast genome sequences for different *Malus* species, we performed a global comparison of the newly assembled sequence to the sequences of chloroplast genomes from five *Malus* species (Fig. 4). The chloroplast sequence of *M. baccata var.*

gracilis showed good co-linearity with the chloroplast genome sequences of the five *Malus* species. The chloroplast sequence of *M. baccata var. gracilis* was more similar to those of *M. toringoides*, *M. hupehensis*, and *M. floribunda*, and less similar to those of *M. ioensis* and *M. yunnanensis*.

Analysis of variation hotspots in the cpDNA of *M. baccata var. gracilis* and other *Malus* species

To find the variation hotspot regions of the *M. baccata var.gracilis* chloroplast genome, the cpDNA genome sequences of ten *Malus* species (including *M. baccata var.gracilis*) were compared (Fig. 5). The results showed that the sequences of these species were relatively consistent overall, but there were some regions with large differences. Most regions with large variation were located in the LSC region, including *psbA-trnH(GUG)*, *rps16-trnk(UUU)*, *trnR(UCU)-atpA*, *trnT(GGU)-psbD*, *psbZ-trnG(GCC)*, *trnT(UGU)-rps4*, *rps14-trnM(CAU)*, *trnV(UAC)-ndhC*, *accD-psaI*, *psaJ-rpl33*, *rpl14-rps8*, and *rps3-rpl16*. The identified hotspots are predominantly located in non-coding areas, whereas the coding regions exhibit a higher degree of stability and conservation. Therefore, it is possible that these zones could serve as the foundation for variations among species.

To further assess DNA polymorphism and identify sequence similarities and differences, the chloroplast genome of *M. baccata var. gracilis* were analyzed using DnaSPv6.0 (Fig. 6). Four regions with substantial variability (Pi ≥ 0.008) were identified: *trnR(UCU)-atpA* (Pi = 0.02017), *trnT(GGU)-psbD* (Pi = 0.0105), *trnT(UGU)-trnL(UAA)* (Pi = 0.008), and *ndhF-rpl32* (Pi = 0.00817). The Pi values observed in the LSC and SSC regions were significantly higher than those observed in the IRs. The comparison of these regions, which exhibit substantial variability, can facilitate species identification, molecular marker studies, and investigations into species evolution.

Boundaries of the IR/SC region of the cpDNA of *M. baccata var. gracilis* and other *Malus* species

The structure of the chloroplast genome is characterized by its circular shape, with four regions: LSC-IRb, IRb-SSC, SSC-IRa, and IRa-LSC, also known as JLB, JSB, JSA, and JLA, respectively (Fig. 7). The variability in the length of chloroplast genomes is facilitated by the dynamic processes of expansion and contraction inside the IR regions.

As the IR regions expand and contract, genes close to the boundary are likely to enter the LSC, SSC, or IR regions. The cpDNA boundaries across ten species of the *Malus* genus were analyzed using IRscope. We successfully identified multiple genes that either traversed or were in close proximity to the boundaries separating the IR and SC portions of the cpDNA, including *rpl22*, *rps19*, *rpl2*, *ycf1*, *ndhF*, and *trnH*.

As presented in Fig. 8, *Malus coronaria* exhibited the largest cpDNA genome, 160,295 base pairs (bp), whereas *M. baccata var. gracilis* had a slightly smaller cpDNA size of 159,992 bp. Among the ten cpDNA samples analyzed, the lengths of the LSC region were in the range of 88,100 to 248,274 bp. However, no variation was observed in the lengths of the SSC, IRa, and IRb sections. The distribution pattern of the *rps19* gene at the JLB boundary exhibited consistency across the eight *Malus* species. The length of the *rps19* gene was found to be 159 bp in the LSC region and 120 bp in the IRb region. However, in the *M.*

coronaria sequence, the *rps19* gene is longer on the LSC side (171 bp) and shorter on the IRb side (108 bp). Conversely, in *M. yunnanensis*, the *rps19* gene is longer on the LSC side (210 bp) and shortest on the IRb side (69 bp). The *ycf1* gene in the JSB locus of the *M. ioensis* cpDNA experienced expansion towards the SSC region, leading to an elongation of the *ycf1* gene. The *ycf1* gene located at the IRb/SSC locus did not undergo crossing over at the JSB border in either *M. toringoides* or *M. hupehensis*. All ten selected Malus species exhibited displacement bias in the *trnH* gene at the JLA locus. The aforementioned findings collectively indicate that the expansion and contraction of the border region have a significant impact on the dimensions and structure of cpDNA.

Construction of a phylogenetic tree of the chloroplast genome of *M. baccata var. gracilis*

A phylogenetic tree was created using the chloroplast genomes of 24 plants from the Malus genus and two exo-taxa plants to explore the evolution of *M. baccata var. gracilis* (Fig. 8). As shown in Figs. <link rid="fig8">8</link>-A and 8-B, the 24 *Malus* species can be well classified in the NJ and ML trees, with the outgroup divided into a large branch. *M. baccata var. gracilis* is clustered in a large group with its closest relatives, *M. hupehensis* and *M. sikkimensis*.

The gene *matk*, found in the chloroplast genome, is a solitary gene that undergoes evolution at a modest pace. This gene has been identified as a potential DNA barcode for taxonomic applications (Uchoi et al., 2016). Using *Vitis* as an outgroup, we next constructed a phylogenetic tree based on this gene in *Vitis vinifera* and nine apple species. Phylogenetic trees (ML and NJ trees) were constructed based on the chloroplast *matk* gene for *M. baccata var. gracilis* and nine other apple species, as shown in Figs. <link rid="fig8">8</link>-C and 8-D. The results showed that *M. baccata var. gracilis* clustered with other apple plants on one large branch, and *V. vinifera* occupied a separate branch. *M. baccata var. gracilis* and *M. hupehensis* are closer in evolution, and *M. ioensis* is less similar to other *Malus* plants in the genus.

Discussion

The advancement of high-throughput sequencing technology has led to an increasing number of chloroplast genomes (Green *et al.*, 2011). These chloroplast genomes can be utilized to investigate genetic variation and phylogenetic analysis. (Liu *et al.*, 2011). In this study, we used whole-genome sequencing, gene annotation, sequence alignment and boundary analysis to obtain a 159,992 bp cpDNA with a loop-like structure from a de novo assembly. The total sequence length was comparable to that of the reported chloroplast genomes from *M. toringoides*, *M. hupehensis*, and *M. coronaria* (Li et al., 2020; Zhang et al., 2018; Liu et al., 2018). This suggests that chloroplast sequence lengths are relatively conserved between species within *Malus*, with a low degree of variation.

Analysis of the cpDNA of *M. baccata var. gracilis* yielded four conserved classical regions, consisting of the large LSC (88100 bp), IRB (26353 bp), SSC (19186 bp), and IRA (26353 bp). The whole genome sequence annotation of the chloroplast of *M. baccata var. gracilis* yielded 112 genes, including 79 unique protein-coding genes. Similarly, the cpDNA genome sequences of *M. floribunda* and *M. hupehensis* both have 78 coding genes (Li et al., 2020; Liu L et al., 2018). The variation in gene count among species can

perhaps be attributed to the expansion and contraction of borders within the chloroplast genome. Previous work showed that boundary shifts in the IR region and variation in the gene spacer region mainly affect the length of cpDNA (Maréchal *et al*., 2010) to stabilize the structure of the genome, with differing extents of variation in the location of the IR/SC boundary of cpDNA in different species (Gu et al., 2022). We found no substantial variation in IR length among different apple genera. This suggests that the IR region is actively involved in preserving the structural stability and functional consistency of cpDNA. Four highly variable regions (Pi ≥ 0.008) were detected in the nucleic acid diversity analysis: *trnR(UCU)-atpA* (Pi = 0.02017), *trnT(GGU)-psbD* (Pi = 0.0105), *ndhF-rpl32* (Pi = 0.00817), and *trnT(UGU)-trnL(UAA)* (Pi = 0.008). These are all within the LSC and SSC regions, exhibiting less variability in nucleic acid content within the two IR zones. Boundary analysis revealed that the *ycf1* region is more variable across plants, possibly because the gene spans the junction between the IR and SSC regions, but also possibly because of the presence of a duplicated sequence (Li et al., 2020). This differs from the high nucleic acid variation in the IR region found in analyses of *Larix kaempferi* and *Dracaena draco* plants (Celiński et al., 2020; Chen SF *et al*., 2020). This suggests that chloroplast genomes may evolve at different rates in different regions and among different species, so these regions of high variation can be used as a basis for evolutionary analysis studies of *M. baccata var. gracilis*.

Codon preference is widespread in natural genomes and has evolved over a long period of time, with influencing factors including mutation, selection, gene length, and gene function (Li et al., 2022). Further calculations of the codon preferences of chloroplast-encoding genes in *M. baccata var. gracilis* revealed a bias towards A/U in the terminal position and a high probability of using UAA as a termination codon, with similar results in *Malus zhaojiaoensis* (Wang et al., 2022) and with *M. floribunda*, *Elaeagnus angustifolia*, and *Actinidia latifolia* (Li et al., 2020; Li et al., 2023; Yang et al., 2021). This suggests that codon preference also varies slightly among species, and this variation may be related to variation and natural selection.

This work compared the *M. baccata var. gracilis* chloroplast genome sequence and single-copy gene *matk* sequence with those of other Malus species and outgroups. Previous clustering analysis (Reim *et al*., 2020) show that *M. baccata var. gracilis* is more similar in evolution to *M. hupehensis*, and *M. ioensis* is less similar to the other *Malus* species. The results of this study provide the basis for further determination of the origin and affinities for *Malus* species or other genera of Rosaceae.

Materials and methods

Materials and DNA sequencing

From plants at the National Botanical Garden in Xiangshan, Beijing, China, leaves were collected and a stratified extraction technique was applied. Samples were dried on silica gel and shipped to the State Key Laboratory of Crop Biology at Shandong Agricultural University in Tai'an, Shandong, China, where total DNA was extracted using a modified CTAB technique. (Matthes et al., 2020) DNA quality was assessed

using 1.2% agarose gel electrophoresis (Matthes et al., 2020), and then a 350 bp library was created and double-end sequencing was carried out using the second-generation Illumina Hiseq4000 platform.

Annotation and assembly of the chloroplast genome

To create a full circular chloroplast genome from the second generation DNA sequencing data, the filtered data were assembled using GetOrganelle v1.7.5 software's default settings (Jin et al., 2020). CPGAVAS2 software (Shi et al., 2019) was used to annotate the chloroplast genome, and CPGView (Liu et al., 2023) and OGDRAW (Stephan et al., 2019) were used to visualize the chloroplast genome map. Using the Apollo program, identified annotation mistakes in the chloroplast genome were manually fixed (Lewis et al., 2002). The manually updated chloroplast genome sequences and annotation files were uploaded to the NCBI website (https://www.ncbi.nlm.nih.gov/) using login OQ737737. Using the OGDRAW program (Stephan *et al*., 2019), the chloroplast genomes of Pendula spp. were visualized, with different gene segments delineated by different colors.

Characterization of the chloroplast genome

Following the splicing and annotation of the chloroplast genome, the long repeat sequences were analyzed using the online software REPuter with parameters of a maximum computed repeats value of 1000, a minimal repeat size of 30, and a Hamming distance of 3, as described by Kurtz et al. (2001). Additionally, the identification of simple sequence repeat (SSR) loci was performed using the MISA software, with default parameters as outlined by Beier et al. (2017).

Analysis of chloroplast genome codon preferences

CodonW software was used to analyze the number of codons, frequency of use, and relative use of synonymous codons in the chloroplast genome CDS of *M. baccata var. gracilis*.

Comparative genomic analysis of chloroplast genomes

Comparisons of chloroplast genome similarity were conducted using the Circoletto tool (Darzentas *et al*., 2010). Multiple sequence comparisons were performed using mVISTA (Frazer et al., 2004). Visualization of chloroplast intergenomic linkage sites was accomplished through IRscope comparisons (Amiryousefi et al., 2018). Phylogenetic analysis of various species was conducted using MEGA_11 (Kumar et al., 2018), resulting in the construction of two types of evolutionary trees: a Neighbor-joining tree and a Maximum Likelihood tree. Pi values of homologous sequences were calculated using DnaSP v.6. Highly variable fragments were selected as molecular markers based on the size of Pi values and sequence length (Jung et al., 2021). The chloroplast genomes used for the comparative genomics analysis in this study were acquired from the NCBI database, and the details are listed in Table 3.

Table 3
Chloroplast genomes used for comparative genomics

Sample name	Genus	Accession No.	Size(bp)
Malus coronaria	*Malus*	MN068247.1	160295 bp
Malus floribunda	*Malus*	OM115955.1	160037 bp
Malus hupehensis	*Malus*	MK020147.1	160065 bp
Malus ioensis	*Malus*	MN062004.1	159841 bp
Malus kansuensis	*Malus*	MW018863.1	160133 bp
Malus micromalus	*Malus*	MF062434.1	159834 bp
Malus rockii	*Malus*	MZ984214.1	160117 bp
Malus toringoides	*Malus*	MT483999.1	160093 bp
Malus yunnanensis	*Malus*	MH394387.1	160067 bp
Vitis vinifera	*Vitis*	NC_007957.1	160928 bp

Conclusion

This is the first sequencing and assembly of the chloroplast genome sequence of *M. baccata var. gracilis*. The cpDNA of *M. baccata var. gracilis* is 159,992 bp in length, with a typical tetrameric structure and 112 genes after removal of repeats. The chloroplast genome of *M. baccata var. gracilis* is biased towards the use of A/U containing codons. Assembly annotation, sequence alignment, variation hotspots, boundary features, and evolutionary relationships of cpDNA were performed. The *M. baccata var. gracilis* cpDNA genome can be used to assess genetic diversity and identify different species. Additionally, this analysis can serve as a beneficial reference for the development of cultivars with pendant features in apple breeding. The results obtained in this study can serve as a benchmark for the preservation and advancement of wild apple resources, while also establishing a theoretical framework for future investigations.

Declarations

Acknowledgements

Qin, X., Hao, Q., Mao, Y.F., and Shen, X. planned and designed the research. Wang, X., Mao, Y. F., Liu, Y.B., Yang, C., Sui, M.Y., Zhang, Y.W., performed experiments, conducted fieldwork, analyzed data. Qin, X., Hu, Y.L., Chen, X.S., Mao, Z.Q., Mao, Y.F., and Shen, X. wrote the manuscript. All authors have read and agreed to the published version of the manuscript

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Data will be made available on request. The data supporting the findings of this study are openly available in the GenBank database at https://www.ncbi.nlm.nih.gov/, under accession number OQ737737.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, award number 32072520; Grants: China National Botanical Garden (North Garden) Project: BZ202206; Fruit innovation team project of Shandong Province (CN), award number SDAIT-06-07; Natural Science Foundation of Shandong Province (CN), award number ZR2020MC132.

References

1. Alkan C, Sajjadian S, Eichler E (2021) Limitations of next-generation genome sequence assembly. Nat Methods 8: 61–65. https://doi.org/10.1038/nmeth.1527
2. Amiryousefi A, Hyvönen J, Poczai P et al (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34: 3030–3031. https://doi.org/10.1093/bioinformatics/bty220
3. Beier S, Thiel T, Münch T et al (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16): 2583-2585. https://doi.org/10.1093/bioinformatics/btx198
4. Celiński K, Kijak H, Wiland-Szymańska J (2020) Complete chloroplast genome sequence and phylogenetic inference of the canary Islands dragon tree (*Dracaena draco* L.). Forests 11(3): 309. https://doi.org/10.3390/f11030309
5. Chen S, Ishizuka W, Hara T et al (2020) Complete Chloroplast Genome of Japanese Larch (*Larix kaempferi*): Insights into Intraspecific Variation with an Isolated Northern Limit Population. Forests 11(8): 884. https://doi.org/10.3390/f11080884
6. Choi KS, Park KT, Park S et al (2017) The Chloroplast Genome of *Symplocarpus renifolius*: A Comparison of Chloroplast Genome Structure in Araceae. Genes 8(11): 324. https://doi.org/10.3390/genes8110324

7. Darzentas N (2010) Circoletto: Visualizing sequence similarity with Circos. Bioinformatics 26: 2620–2621. https://doi.org/10.1093/bioinformatics/btq484
8. Dobránszki J, Mendler-Drienyovszki N (2014) Cytokinin-induced changes in the chlorophyll content and fluorescence of in vitro apple leaves. J of Plant Physiol 171: 1472–1478. https://doi.org/10.1016/j.jplph.2014.06.015
9. Frazer KA, Pachter L, Poliakov A et al (2004) VISTA: Computational tools for comparative genomics. Nucleic Acids Res 32 (Suppl. 2): W273–W279. https://doi.org/10.1093/nar/gkh458
10. Geng D (2021) Insights into the effect of human civilization on Malus evolution and domestication. Plant Biotechnology J 19: 2206–2220. https://doi.org/10.1111/pbi.13648
11. Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. Plant J 66: 34–44. https://doi.org/10.1111/j.1365-313X.2011.04541.x
12. Gu X, Zhu M, Su Y et al (2022) A Large Intergenic Spacer Leads to the Increase in Genome Size and Sequential Gene Movement around IR/SC Boundaries in the Chloroplast Genome of Adiantum malesianum (Pteridaceae). Int J Mol Sci 23: 15616. https://doi.org/10.3390/ijms232415616
13. He J, Zhou J, Wan H et al (2020) Rootstock–Scion Interaction Affects Cadmium Accumulation and Tolerance of *Malus*. Frontiers in Plant Science 11: 1664-462X. https://doi.org/10.3389/fpls.2020.01264
14. Jin JJ, Yu WB, Yang JB et al (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21(1): 241. https://doi.org/10.1186/s13059-020-02154-5
15. Jung J, Kim C, Kim JH (2021) Insights into phylogenetic relationships and genome evolution of subfamily *Commelinoideae* (*Commelinaceae* Mirb.) inferred from complete chloroplast genomes. BMC Genomics 22: 231. https://doi.org/10.1186/s12864-021-07541-1
16. Kumar S, Stecher G, Li M et al (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35: 1547–1549. https://doi.org/10.1093/molbev/msy096
17. Kurtz S, Choudhuri JV, Ohlebusch E et al (2001) Reputer: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research. 29(22): 4633 -4642. https://doi.org/10.1093/nar/29.22.4633
18. Lewis SE, Searle S, Harris N et al (2002) Apollo: a sequence annotation editor. Genome Biol 3(12): 1-14. https://doi.org/10.1186/gb-2002-3-12-research0082
19. Li C, Zhou L, Nie J (2023) Codon usage bias and genetic diversity in chloroplast genomes of *Elaeagnus* species (Myrtiflorae: *Elaeagnaceae*). Physiol Mol Biol Plants 29: 239–251. https://doi.org/10.1007/s12298-023-01289-6
20. Li G, Zhang L, Xue P (2022) Comparative Analysis on the Codon Usage Pattern of the Chloroplast Genomes in *Malus* Species. Biochem Genet 61: 1050–1064. https://doi.org/10.1007/s10528-022-10302-8
21. Li YN, Liu YL, Wu P (2020) The complete chloroplast genome sequence of *Malus toringoides* (Rosaceae). Mitochondrial DNA Part B 5(3): 2787-2789.

https://doi.org/10.1080/23802359.2020.1780977

22. Liao L, Zhang W, Zhang B et al (2021) Unraveling a genetic roadmap for improved taste in the domesticated apple. Mol Plant 14: 1454–1471. https://doi.org/10.1016/j.molp.2021.05.018
23. Liu C, Shi LC, Zhu YJ (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13: 715. https://doi.org/10.1186/1471-2164-13-715
24. Liu L, Wang Y, He P et al(2018) Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. BMC Genomics 19(1): 235. https://doi.org/10.1186/s12864-018-4633-x
25. Liu Q, Li X, Li M et al (2020) Comparative chloroplast genome analyses of Avena: Insights into evolutionary dynamics and phylogeny. BMC Plant Biol 20: 406. https://doi.org/10.1186/s12870-020-02621-y
26. Liu S, Ni Y, Li J et al (2023) CPGView: A package for visualizing detailed chloroplast genome structures. Mol Ecol Resour 23: 694-704. https://doi.org/10.1111/1755-0998.13729
27. Matthes N, Westphal K, Haldemann C (2020) Validation of a modified CTAB method for DNA extraction from protein-rich maize feedstuffs. J Consum Prot Food Saf 15: 331–340. https://doi.org/10.1007/s00003-020-01285-y
28. Maréchal A, Brisson N (2010) Recombination and the maintenance of plant organelle genome stability. New Phytol 186(2): 299-317. https://doi.org/10.1111/j.1469-8137.2010.03195.x
29. Moghaddam M, Ohta A, Shimizu M (2022) The complete chloroplast genome of Onobrychis gaubae (Fabaceae-Papilionoideae): comparative analysis with related IR-lacking clade species. BMC Plant Biol 22: 75. https://doi.org/10.1186/s12870-022-03465-4
30. Papp D, Gao L, Thapa R (2020) Field apple scab susceptibility of a diverse Malus germplasm collection identifies potential sources of resistance for apple breeding. CABI Agric Biosci 1: 16. https://doi.org/10.1186/s43170-020-00017-4
31. Ren Y, Yan M, Zhao XQ et al (2019) Characterization of the complete chloroplast genome of *Malus baccata var. xiaojinensis*. Mitochondrial DNA Part B 4(2): 2487-2488. https://doi.org/10.1080/23802359.2019.1638840
32. Rong H, Huang B, Han X (2022) Pedigree reconstruction and genetic analysis of major ornamental characters of ornamental crabapple (*Malus* spp.) based on paternity analysis. Scientific Reports 12: 14093. https://doi.org/10.1038/s41598-022-18352-z
33. Sestras RE, Sestras AF (2023) Quantitative Traits of Interest in Apple Breeding and Their Implications for Selection. Plants 12: 903. https://doi.org/10.3390/plants12040903
34. Shi L, Chen H, Jiang M et al (2019) CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Research 47(W1): W65-W73. https://doi.org/10.1093/nar/gkz345
35. Stephan Greiner (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research 47(W1): W59–W64. https://doi.org/10.1093/nar/gkz238

36. Uchoi A, Malik SK, Choudhary R (2016) RETRACTED ARTICLE: Inferring Phylogenetic Relationships of Indian Citron (Citrus medica L.) based on *rbcL* and *matK* Sequences of Chloroplast DNA. Biochem Genet 54: 249–26. https://doi.org/10.1007/s10528-016-9716-2
37. Wang MR, Chen L (2018) Cryobiotechnology of apple (*Malus* spp.): Development, progress and future prospects. Plant Cell Rep 37: 689–709. https://doi.org/10.1007/s00299-018-2249-x
38. Wang X, Wang D, Gao N et al (2022) Identification of the Complete Chloroplast Genome of *Malus zhaojiaoensis* Jiang and Its Comparison and Evolutionary Analysis with Other *Malus* Species. Genes 13: 560. https://doi.org/10.3390/genes13040560
39. Wick RR, Schultz MB, Zobel J (2015) Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31(20): 3350-3352. https://doi.org/10.1093/bioinformatics/btv383
40. Yang AH, Hu M, Liu TY et al (2021) Characterization and phylogenetic analysis of the complete chloroplast genome of *Actinidia latifolia* (*Actinidiaceae*), Mitochondrial DNA Part B 6(2): 672-673. https://doi.org/10.1080/23802359.2021.1881838
41. Zhang X, Rong CX, Qin L (2018) Complete chloroplast genome sequence of *Malus hupehensis*: genome structure, comparative analysis, and phylogenetic relationships. Molecules 23(11): 2917. https://doi.org/10.3390/molecules23112917
42. Zhou T, Fan J, Zhao M et al (2019) Phenotypic variation of floral organs in *Malus* using frequency distribution functions. BMC Plant Biol 19: 574. https://doi.org/10.1186/s12870-019-2155-6

Figures

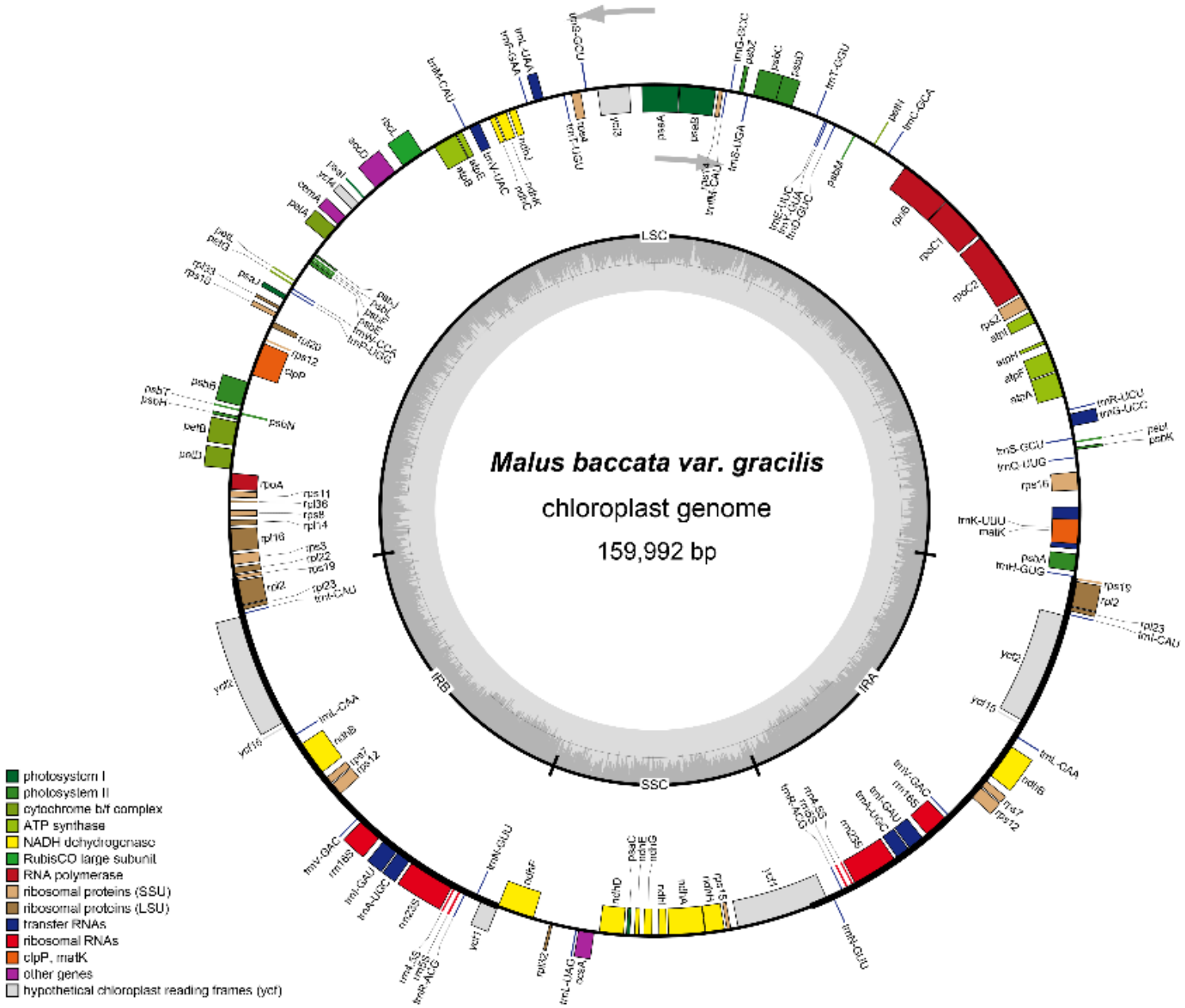

Figure 1

Chloroplast genome mapping in *M. baccata var. gracilis*

Arrows indicate the direction of transcription, and the shaded areas show GC/AT content

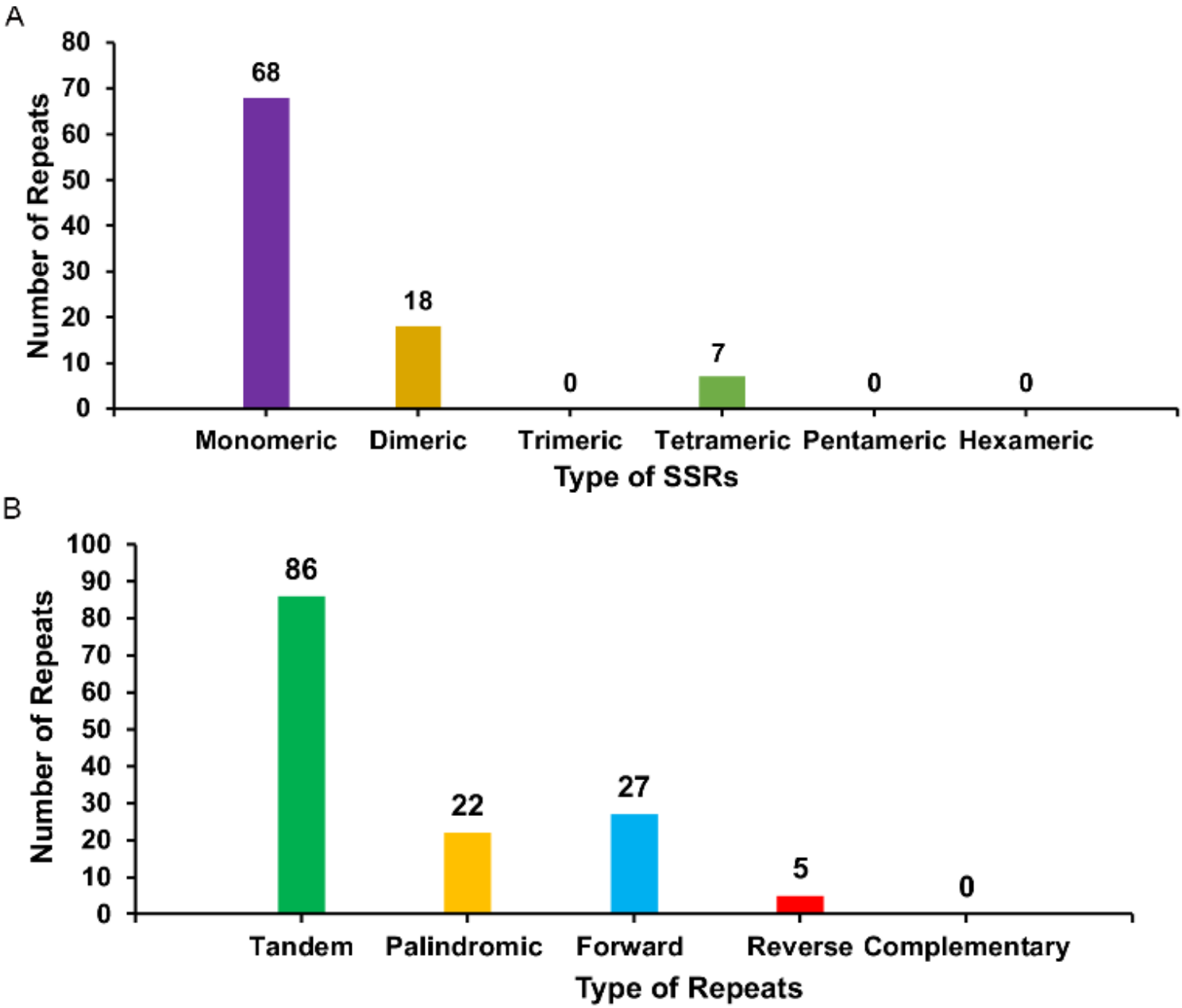

Figure 2

Repeat sequence analysis of the chloroplast genome of *M. baccata var. gracilis*

(A) The horizontal coordinate represents the type of SSRs, the vertical coordinate represents the number of repeats, and the purple bar represents single SSRs. The orange bar represents dimer SSRs and the green bar represents tetramer SSRs. (B) The abscissa represents the type of the repeating sequence, and the ordinate represents the number of repeats. The green bar represents tandem repeats, and the yellow bar represents palindromic repeats, the blue bar indicates forward repeats, and the red bar indicates reverse repeats

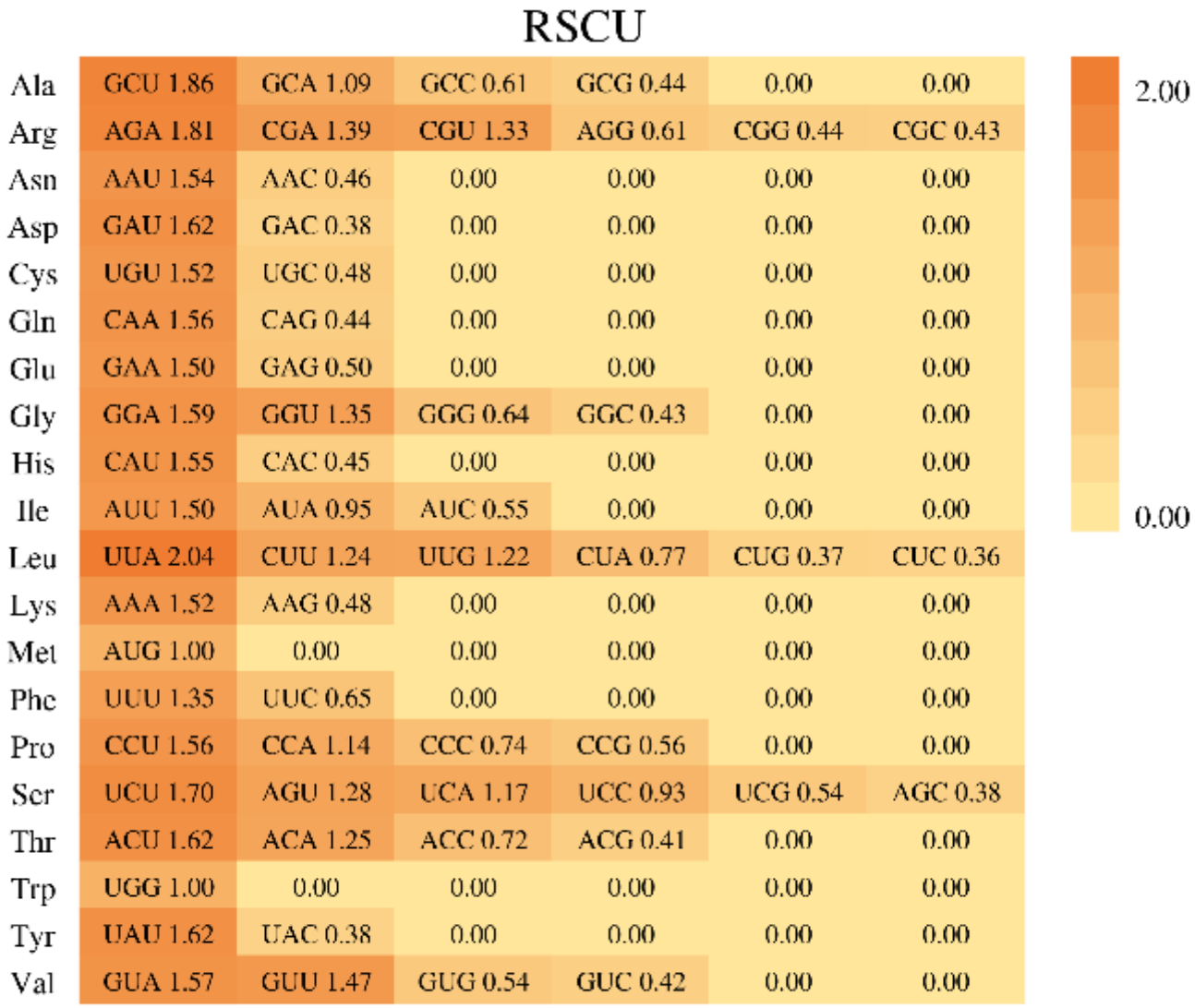

Figure 3

Codon preference heat map of the chloroplast genome of *M. baccata var. gracilis*

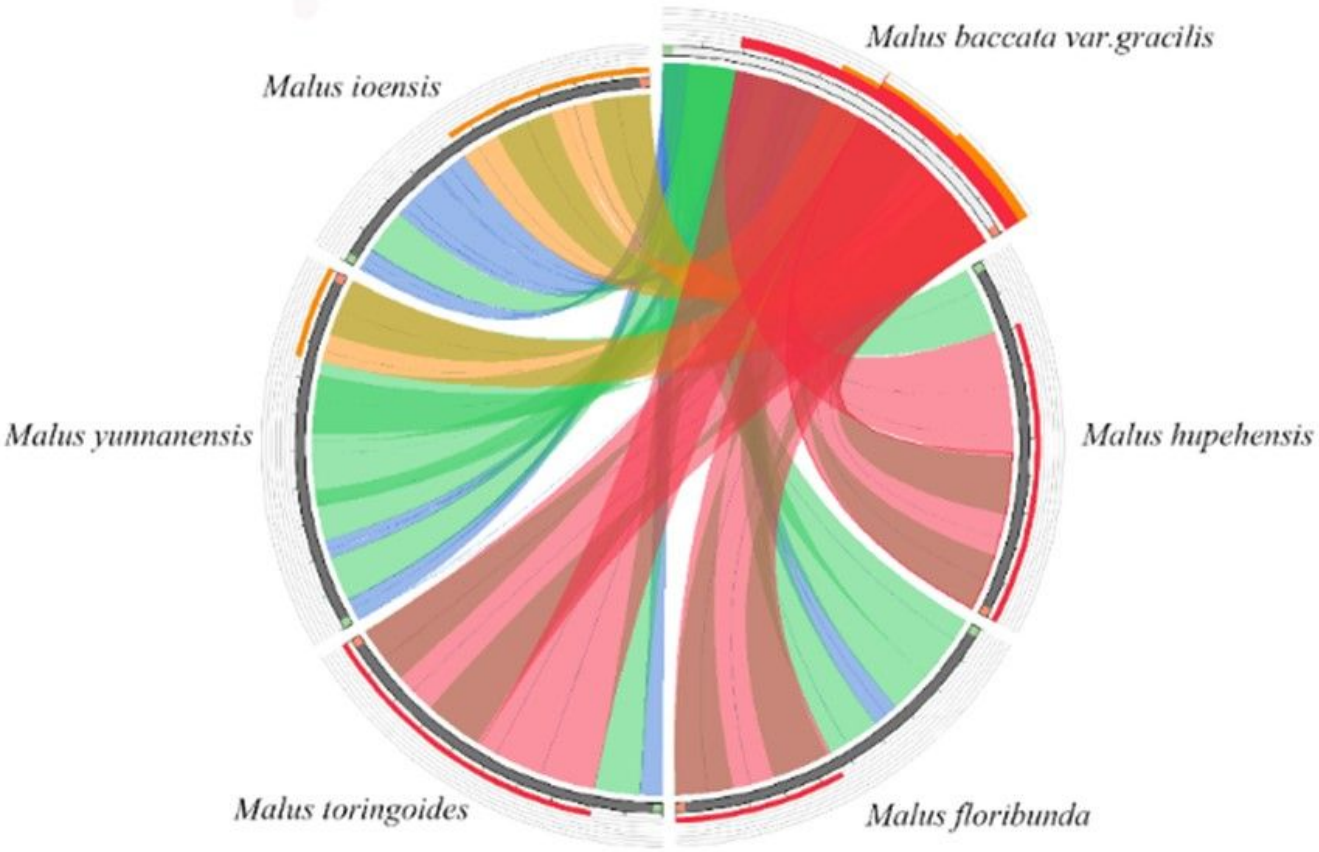

Figure 4

Similarity of the *M. baccata var. gracilis* chloroplast genome to other chloroplast genome sequences.

Various colors are used to signify distinct areas of comparison, each with corresponding scoring rates. The score-to-maximum ratios were assigned color codes, with blue representing ratios less than or equal to 0.25, green representing ratios less than or equal to 0.50, orange representing ratios less than or equal to 0.75, and red representing ratios greater than 0.75.

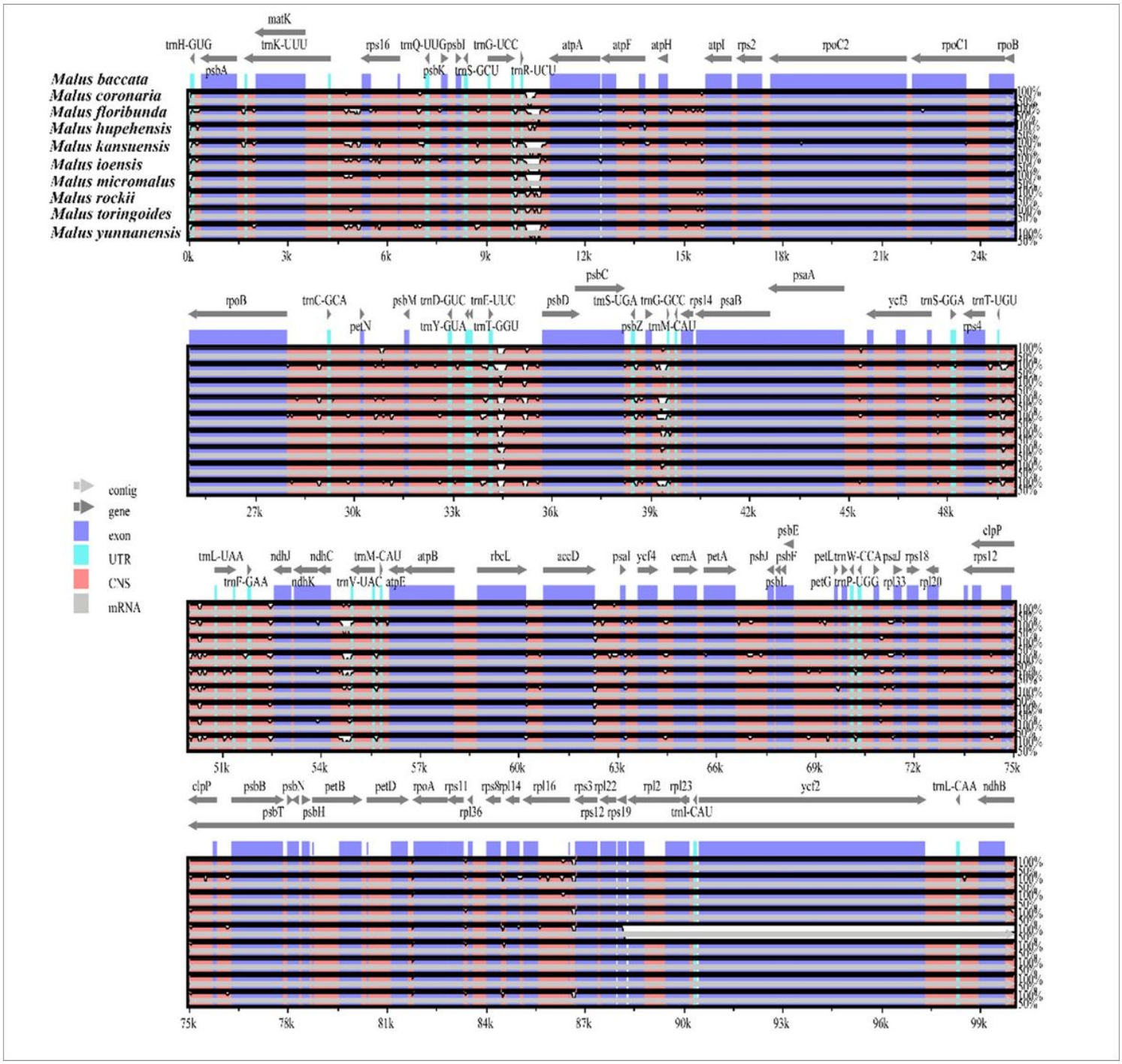

Figure 5

Sequence alignment of chloroplast genomes

The horizontal axis represents the sequence length of the chloroplast genome, and the vertical axis shows the percentage of similarity

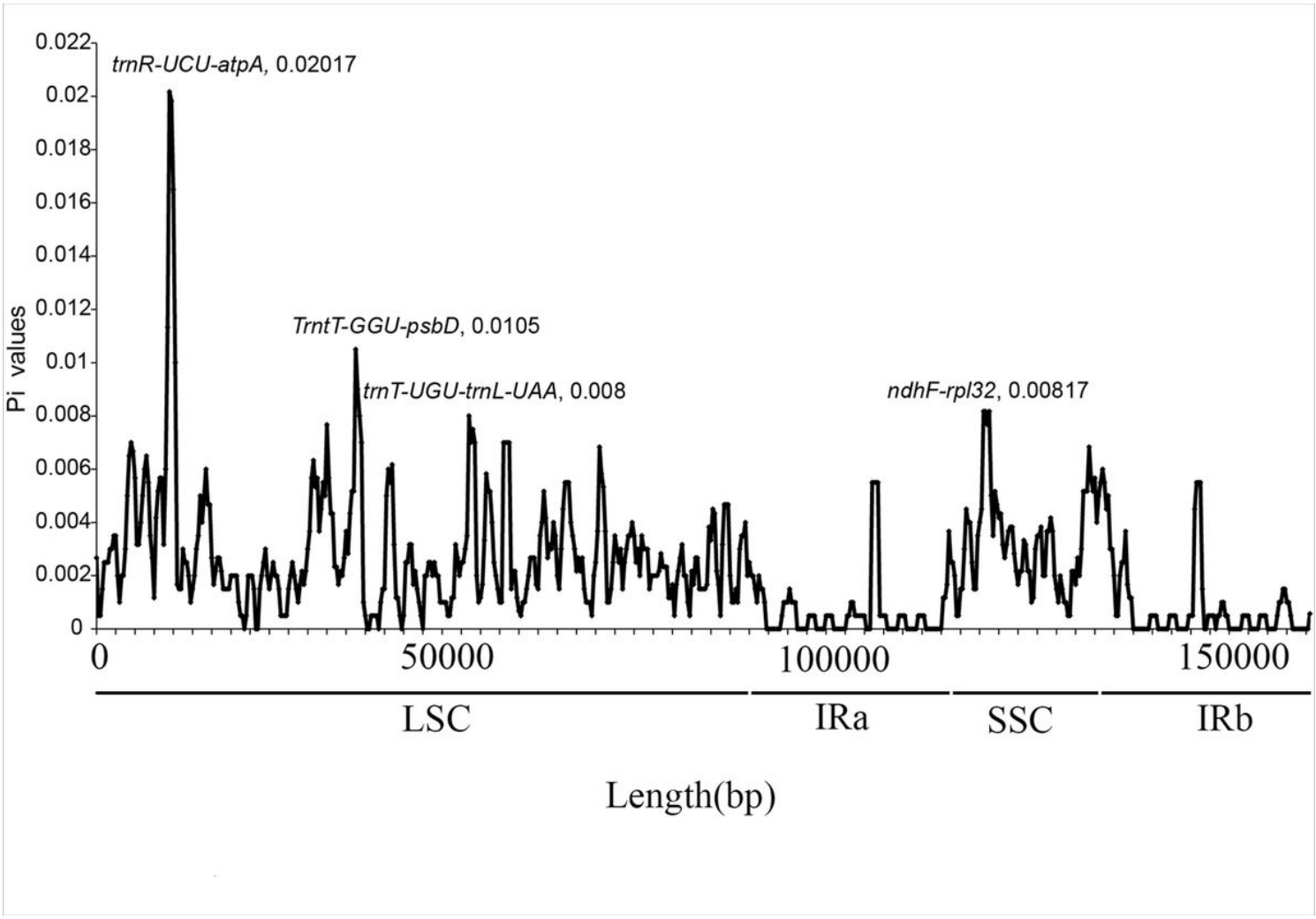

Figure 6

Nucleic acid diversity in the chloroplast genome of *M. baccata var. gracilis*

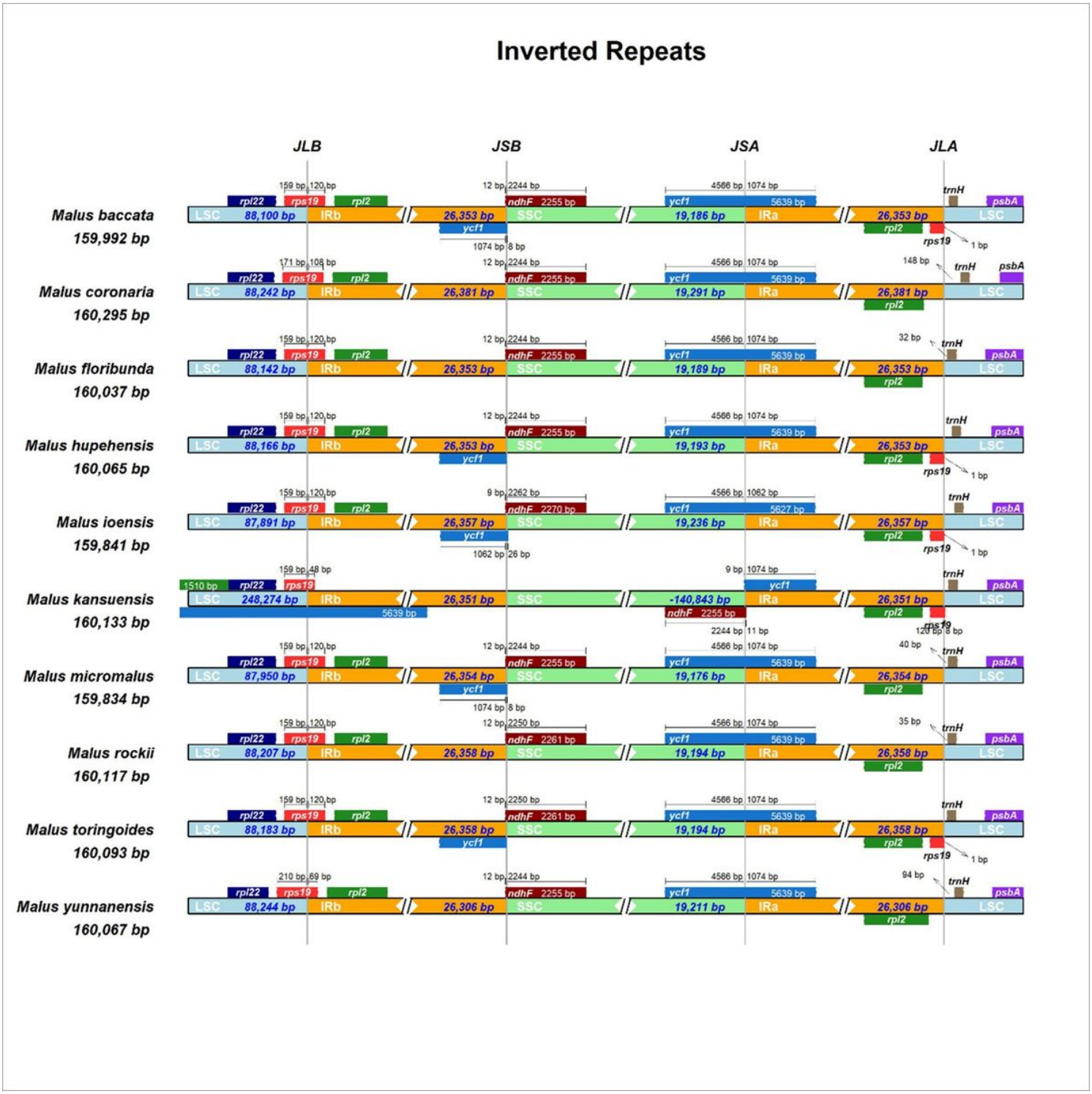

Figure 7

Boundary analysis of chloroplast genome in *M. baccata var. gracilis*

JLB, JSB, JSA, and JLA are the four junction sites in the boundary region of chloroplast genome

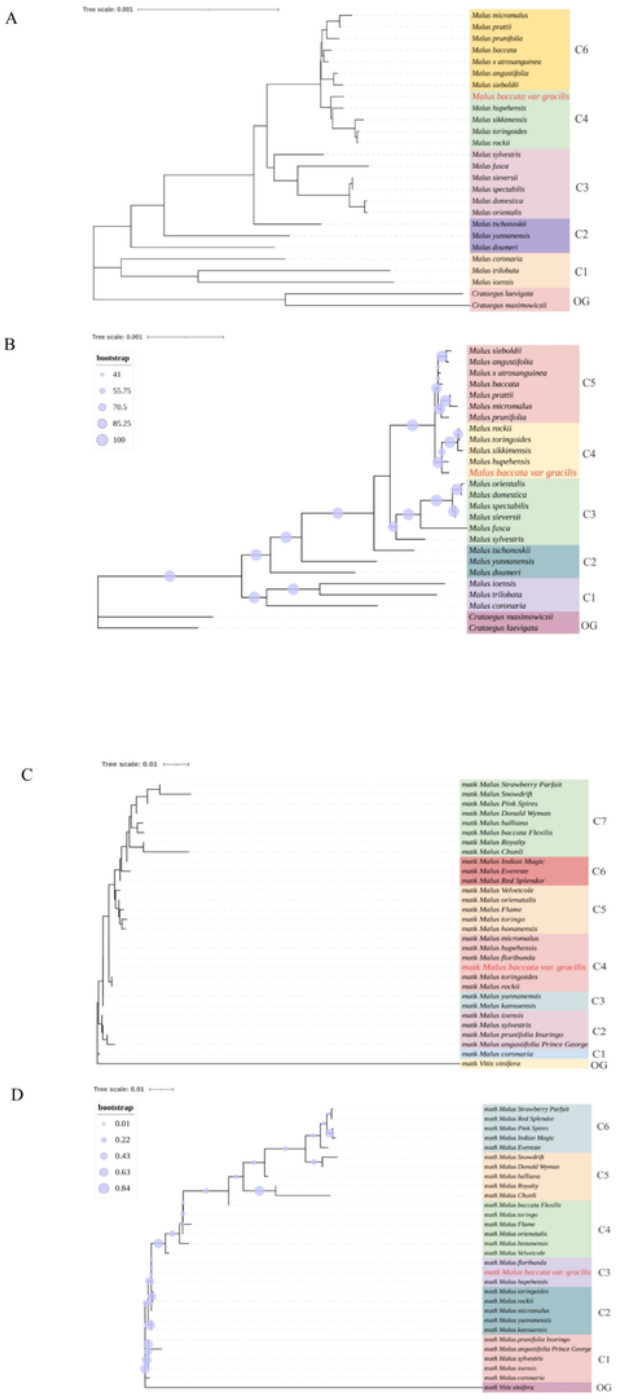

Figure 8

A: NJ tree based on chloroplast genome construction; B: ML tree based on chloroplast genome construction; C: NJ tree based on single copy chloroplast gene *matK*; D: ML tree based on single copy chloroplast gene *matK*

Note: OG is an outgroup, C is a subgroup of the genus *Malus*. *Malus baccata var.gracilis* is highlighted in red