Chloroplast Genome Characteristics and Evolutionary History of Malus sieversii Postprint

Authors: Zhang Jian, Zhang Hongxiang

Date: 2024-12-16T00:00:00+00:00

Abstract

Xinjiang wild apple is an important germplasm resource of the genus Malus, a national second-class protected plant, and also one of the ancestors of cultivated apples. By comparing the structural characteristics of chloroplast genomes among different populations of Xinjiang wild apple, this study aims to elucidate the lineage differentiation pattern and species evolutionary history of Xinjiang wild apple. Whole-genome sequencing was performed on populations from 16 different regions using the Illumina NovaSeq platform, with one representative individual selected from each population. After quality control of the sequencing data, genome assembly and functional annotation were carried out. Subsequently, in-depth structural analysis and lineage differentiation studies were conducted on the assembled genomes. The results showed that the chloroplast genome sequence of Xinjiang wild apple has a total length of 160195~160279 bp, with a typical quadripartite structure. A total of 131 genes were annotated in the chloroplast genome; 48~58 long repeat sequences and 93~101 simple sequence repeats were detected. The chloroplast genomes of Xinjiang wild apple and other Malus species showed low variation in the IR (inverted repeat) region, while the detected variations occurred mainly in non-coding regions. Phylogenetically, Xinjiang wild apple ultimately divided into three lineages, with lineage I mainly distributed in the east, and lineages II and III mainly distributed in the west. The divergence time between lineage I and lineage II was 1.74 Ma, and the divergence time among lineage I, lineage II, and lineage III was 2.28 Ma. The genetic differentiation of Xinjiang wild apple was influenced by Quaternary climate changes. Compared with Xinjiang wild apple distributed abroad, the genetic diversity of Xinjiang wild apple distributed in China is lower. Different protection strategies should be adopted for Xinjiang wild apple distributed in China, with special attention paid to the Tacheng region where genetic diversity is relatively high.

Full Text

Characteristics and Evolutionary History of the Chloroplast Genome in *Malus sieversii*

ZHANG Jian[1,2], **ZHANG Hongxiang**[1,3,4]

[1] State Key Laboratory of Desert and Oasis Ecology, Key Laboratory of Ecological Safety and Sustainable Development in Arid Lands, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, Xinjiang, China

[2] University of Chinese Academy of Sciences, Beijing 100093, China

[3] Xinjiang Key Lab of Conservation and Utilization of Gene Resources, Urumqi 830011, Xinjiang, China

[4] Specimen Museum of Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, Xinjiang, China

Abstract

Malus sieversii, a state-protected species and the progenitor of cultivated apples, is an important germplasm resource within the genus *Malus*. In this study, we aimed to compare the structural characteristics of chloroplast genomes across various populations of *M. sieversii*, clarify the lineage divergence pattern, and trace the evolutionary history of this species. We used the Illumina NovaSeq platform to conduct whole-genome sequencing of individuals from 16 different populations, with one sample representing each population. After conducting quality control on the sequencing data, we performed genome assembly and functional annotation. Subsequently, we conducted a comprehensive structural analysis and lineage differentiation studies on the assembled genomes. The chloroplast genome length in *M. sieversii* ranged from 160,195 to 160,279 base pairs (bp), exhibiting a typical tetrad structure. In total, 131 genes were identified within the chloroplast genome, along with 48–58 long repeats and 93–101 simple sequence repeats. Notably, variations in the IR (inverted repeat) region between *M. sieversii* and other species in the genus were minimal, predominantly occurring in noncoding regions. Phylogenetic analysis revealed that *M. sieversii* clusters into three distinct lineages: lineage I, primarily occupying the eastern part of the distribution range, and lineages II and III, predominantly found in the west. The divergence time between lineages I and II/III was approximately 1.74 million years ago (Ma), while the divergence between lineages II and III was around 2.28 Ma. These findings indicate that the lineage divergences of *M. sieversii* were significantly influenced by climate changes during the Quaternary period. Compared to internationally distributed populations, *M. sieversii* in China shows relatively low genetic diversity. Therefore, tailored conservation strategies should be implemented for *M. sieversii* across different regions, with particular emphasis on protecting genetically diverse populations in the Tacheng area.

Keywords: *Malus sieversii*; chloroplast genome; phylogeny; divergence time;

conservation genetics

1.1 Experimental Materials, Sequencing, and Chloroplast Genome Assembly

We extensively collected *Malus sieversii* samples covering 12 populations in Xinjiang, China, and expanded to neighboring countries including Kyrgyzstan, Kazakhstan, and Tajikistan, totaling 16 populations. From each population, we carefully selected one individual and collected silica-dried leaf samples for subsequent high-throughput sequencing and genomic analysis. The sampling covered the main distribution areas of *M. sieversii* (Table). Collected samples were sent to Shanghai Personal Biotechnology Co., Ltd. for raw data sequencing on the Illumina NovaSeq platform. The generated data were quality-controlled using FastQC v0.11.90. We then used Plast V1.2.9 software for chloroplast genome annotation [13]. After assembly, PGA software [14] was used for manual correction. The assembled chloroplast genomes were uploaded to the NCBI database (accession numbers: PP24956-PP24971).

1.2 Chloroplast Genome Structure and Functional Analysis

After annotating the chloroplast genomes, we used Geneious Prime 2021 to analyze gene composition and function of *M. sieversii* chloroplast genomes. The physical map of the chloroplast genome was drawn using the online OGDRAW software (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

1.3 Repeat Sequences and SSR Loci Analysis

We used the online software REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer) to analyze repeat sequences in the chloroplast genomes, including forward, reverse, complementary, and palindromic repeats. Software parameters were set with a maximum repeat sequence number of 5000, minimum repeat length of 30 bp, Hamming distance of 3, and default edit distance. We used MISA-web [16] to detect simple sequence repeats, with mononucleotide repeats set to 10 times, dinucleotides to 6 times, and trinucleotides to 5 times. Other nucleotide types were set to 3 times, with all other parameters as defaults.

1.4 Comparative Analysis of Chloroplast Genomes

We used CPJSdraw V0.0.1 [17] to compare the boundary information of IR regions among nine *Malus* species. Additionally, we used the online software mVISTA (https://genome.lbl.gov/vista/index.shtml) to compare sequence differences in chloroplast genomes between *M. sieversii* and closely related species. To elucidate the phylogenetic relationship between *M. sieversii* and other *Malus* species, we downloaded 47 chloroplast genomes of *Malus* species

from GenBank (https://www.ncbi.nlm.nih.gov/nuccore) and used *Pyrus ussuriensis* and *Prunus salicina* as outgroups. We first performed alignment using MAFFT(v7.520) [19], then used the resulting fasta file format for analysis. We used IQ-TREE (V2.2.2.6) [24] to select the best model TVM+F+R2 for constructing the phylogenetic tree. Additionally, we used BEAST (v1.10) [20] to estimate divergence times, with fossil calibration points set at 10 Ma and 20 Ma, and chain length set to 100 million.

2.1 Chloroplast Genome Structure, Function, and Characteristics

We analyzed the genome structure, function, and characteristics of 16 samples and found that the *M. sieversii* chloroplast genome had a total length of 160,195–160,279 bp, with an average length of 160,253.56 bp. The chloroplast genome contained 131 genes, including 86 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. *Malus sieversii* chloroplast genomes have a classic circular structure (Fig. [Figure 1: see original paper]). The large single-copy (LSC) region had an average length of 88,352.25 bp, while the small single-copy (SSC) region averaged 26,360.125 bp. Additionally, the IR region length was relatively conserved at 26,353–26,373 bp. The GC content of *M. sieversii* chloroplast genomes averaged 36.7%. We detected 19 genes with introns in the chloroplast genome, with the majority (17 genes) having only one intron and the remaining 2 genes having two introns (Table).

2.2 Repeat Sequences and SSR Loci Analysis

We detected 48–58 repeat sequences in the *M. sieversii* chloroplast genome, among which palindromic repeats appeared most frequently (28–34 times), followed by forward repeats (17–18 times), reverse repeats (30–50 times), and complementary repeats appeared least frequently. These repeats were mainly found in the LSC region and least frequently in the SSC region. Most repeats were detected in non-coding genes, with a small portion detected on the *ycf* gene and primarily located in intron regions. In the SSR loci detection of *M. sieversii*, most samples were found to have 93–101 SSR loci. Among these SSR loci, mononucleotide repeats appeared most frequently, while pentanucleotide repeats appeared least frequently, and no hexanucleotide repeats were detected (Fig. [Figure 2: see original paper]).

2.3 Comparative Analysis of Chloroplast Genomes

Comparative analysis of IR region boundaries among nine *Malus* species showed that the IR region length was relatively conserved, with boundary expansion differences within 12 bp. The IR region boundaries were located between the *rps* gene and the *ycf* gene, with expansion distances of 1,074 bp and 1 bp respectively. As shown in the mVISTA analysis, using cultivated apple as the reference sequence, the chloroplast genomes of *M. sieversii* showed some differences from

other *Malus* species, though fewer differences than other varieties. These differences were mainly concentrated in non-coding regions, with few variations in coding regions. Additionally, the IR region was relatively conserved with minimal differences.

2.4 Phylogenetic Analysis

The phylogenetic tree showed that most nodes had high support rates, indicating credible relationships. The maximum likelihood (ML) tree yielded similar results and is not shown separately. The results revealed that *M. sieversii* had the closest relationship with true apple group species such as cultivated apple (*M. domestica*) and forest apple (*M. sylvestris*), a slightly closer relationship with the *M. baccata* group, a more distant relationship with the *M. floribunda* group, and the farthest relationship with the *M. toringoides* group. The 16 *M. sieversii* populations could be divided into three lineages: lineage I included Kyrgyzstan: YLALT, Xinyuan County, and Emin County; lineage II included Kyrgyzstan: YLALT1, Huocheng County, and Tajikistan; lineage III included Kazakhstan: YLALT2, Tuoli County, and Tajikistan. The Kazakhstan YLALT1 population was relatively unique. Divergence time analysis indicated that the split between *M. sieversii* and related species occurred at approximately 6.57 Ma, while the divergence between lineages I and II/III occurred at 1.74 Ma, and the divergence between lineages II and III occurred at 2.28 Ma (Fig. [Figure 5: see original paper]).

---

3 Discussion and Conclusion

Our annotation and analysis revealed that the *M. sieversii* chloroplast genome contains 131 genes, most of which are protein-coding genes. This gene number differs from previously published chloroplast genomes of other *Malus* species [10], likely due to differences in genome annotation databases used across studies. The chloroplast genome contains 86 protein-coding genes, 37 tRNA genes, and 8 rRNA genes, with the number of rRNA genes being identical to that found in most plant chloroplasts. The GC content of *M. sieversii* chloroplast genomes is 36.7%, indicating relative conservation.

Since chloroplast genomes of *Malus* species have typical tetrad structures, comparison of IR region boundaries can directly reveal structural variations. Our analysis showed minimal variation in IR region length, with no IR region loss observed (Fig. [Figure 3: see original paper]). Comparison between *M. sieversii* and other *Malus* species using mVISTA revealed that sequence variations mainly occurred in non-coding regions, with minimal variation in coding regions. This may be because genes in coding regions are relatively conserved and essential for normal chloroplast function. Repeat sequences play important roles in analyzing chloroplast gene rearrangement, base substitution, phylogeny, and genome evolution [15]. Our study found that the *M. sieversii* chloroplast genome con-

tains four types of repeat sequences, mostly forward repeats with a small portion being complementary sequences, and most repeats appeared in the LSC region. Among SSR loci, mononucleotide repeats were most frequent, while pentanucleotide repeats were least frequent, with mononucleotide repeats occurring far more frequently than dinucleotide repeats, possibly because mononucleotide repeats are more easily unwound, facilitating gene expression [26].

Phylogenetic analysis using chloroplast genomes provides high reliability for plant classification and evolutionary studies. Due to their conserved nature and uniparental inheritance, chloroplast genomes provide clear phylogenetic signals. Additionally, chloroplast genomes contain abundant gene and non-coding region information that helps reveal evolutionary relationships among plants. Our phylogenetic tree analysis clearly showed that 16 populations of *M. sieversii* from different lineages could be divided into three main lineages. The Kazakhstan YLALT1 population was particularly unique, not clustering with other *M. sieversii* lineages but instead grouping separately with *M. baccata*. Similar phenomena were observed in Nikiforova et al.'s study [22], where *M. sieversii* chloroplast genomes clustered with *M. baccata*, and also with *M. halliana*, *M. asiatica*, and *M. adstringens*. These findings may indicate that chloroplast genomes of *Malus* species are not monophyletic, which can occur when using chloroplast genomes for phylogenetic reconstruction. In our study, sample YLALT1 exhibited this characteristic and was therefore not discussed in depth.

Divergence time analysis revealed that the split between *M. sieversii* and related species occurred at approximately 6.57 Ma, while the divergence between lineages I and II/III occurred at 1.74 Ma, and between lineages II and III at 2.28 Ma. These divergence times indicate that the genetic differentiation of *M. sieversii* was strongly influenced by dramatic climate changes during the Quaternary period. However, due to both natural and anthropogenic factors, the already vulnerable *M. sieversii* populations have experienced significant area reduction [29]. To protect the core germplasm resources of *M. sieversii*, we must understand its genetic diversity and lineage divergence patterns to provide theoretical support for future conservation efforts.

Our chloroplast genome research shows that lineage diversity of *M. sieversii* in Xinjiang is significantly lower than in foreign distribution areas. The Ili region contains only one lineage with low genetic diversity, while the Tacheng region contains two lineages with relatively high genetic diversity. Previous studies [7] revealed that Ili and Tacheng belong to different genetic units, indicating significant genetic differences between the two regions, possibly due to their unique natural conditions, geographical environments, or historical evolution. In terms of natural environment, the Ili region receives significantly more precipitation than Tacheng [30], which may contribute to the observed differences. Field investigations confirmed that *M. sieversii* populations in Tacheng are smaller in number than those in Ili, yet Tacheng populations exhibit higher genetic diversity, indicating rich genetic variation and adaptability at the genetic level.

Therefore, to effectively protect and maintain *M. sieversii* populations in these

two regions, we recommend implementing differentiated conservation strategies. For the Ili region, focus may be needed on population size and distribution, as well as environmental factors affecting survival. For the Tacheng region, besides conventional protection measures, special attention should be paid to protecting genetic diversity, as it is crucial for populations to adapt to environmental changes and survive. We recommend conducting detailed genetic resource surveys in Tacheng, collecting seeds and plant samples, and establishing germplasm resource banks for future research and propagation. Appropriate measures such as optimizing in-situ conservation systems should also be implemented to prevent further population decline.

References

[1] Velasco R, Zharkikh A, Affourtit J, et al. The genome of the domesticated apple (*Malus domestica* Borkh.)[J]. *Nature Genetics*, 2010, 42(10): 833-839.

[2] Zhang Xinshi. On the eco-geographical characters and the problems of classification of the wild fruit forest in the Ili valley of Sinkiang[J]. *Acta Botanica Sinica*, 1973, 15(2): 239-253.

[3] Volk G M, Peace C P, Henk A D, et al. DNA profiling with 20K apple SNP array reveals *Malus domestica* hybridization and admixture in *M. sieversii*, *M. orientalis*, and *M. sylvestris* genebank accessions[J]. *Frontiers in Plant Science*, 2022, 13: 1015658.

[4] Zhang Hongxiang, Li Xuesong, Wang Jingcheng, et al. Insights into the aridification history of central Asian mountains and international conservation strategy from the endangered wild apple tree[J]. *Journal of Biogeography*, 2021, 48(2): 332-344.

[5] Zhang Hongxiang, Zhang Menglun, Wang Lina. Genetic structure and historical demography of *Malus sieversii* in the Ili valley and the western mountains of the Junggar Basin, Xinjiang, China[J]. *Journal of Arid Land*, 2015, 7(2): 264-271.

[6] Zhang Hongxiang, Wen Zhibing, Wang Qian. Population genetic structure of *Malus sieversii* and environmental adaptations[J]. *Chinese Journal of Plant Ecology*, 2022, 46(9): 1098-1108.

[7] Zhang Hongxiang, Zheng Tianyong. Effect of habitat fragmentation on the population genetic structure of *Malus sieversii*[J]. *Arid Zone Research*, 2020, 37(3): 715-721.

[8] Maimaiti Mierkamili, Liu Zhongquan, Ma Xiaodong, et al. Survival status, problems and conservation strategies of *Malus sieversii*[J]. *Arid Zone Research*, 2021, 41(12): 2100-2109.

[9] Zhao Yufen. Application progress of chloroplast genome in botany research[J]. *Biology Teaching*, 2022, 47(3): 83-85.

[10] Naizaier R, Qu Z, Wu S, et al. The complete chloroplast genome of *Malus sieversii* (Rosaceae), a wild apple tree in Xinjiang, China[J]. *Mitochondrial DNA Part B*, 2019, 4(1): 983-984.

[11] Chen S F, Zhou Y Q, Chen Y R, et al. Fastp: An ultra-fast all-in-one fastq preprocessor[J]. *Bioinformatics*, 2018, 34(17): 884-890.

[12] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform[J]. *Bioinformatics*, 2009, 25(14): 1754-1760.

[13] Qu X J, Moore M J, Li D Z, et al. PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes[J]. *Plant Methods*, 2019, 15: 1-12.

[14] Geneious Prime[EB/OL]. https://www.geneious.com

[15] Kurtz S, Choudhuri J V, Ohlebusch E, et al. Reputer: The manifold applications of repeat analysis on a genomic scale[J]. *Nucleic Acids Research*, 2001, 29(22): 4633-4642.

[16] Beier S, Thiel T, Münch T, et al. MISA-web: A web server for microsatellite prediction[J]. *Bioinformatics*, 2017, 33(16): 2583-2585.

[17] Li H, Guo Q, Xu L, et al. CPJSdraw: Analysis and visualization of junction sites of chloroplast genomes[J]. *PeerJ*, 2023, 11: e15326.

[18] Frazer K A, Pachter L, Poliakov A, et al. VISTA: Computational tools for comparative genomics[J]. *Nucleic Acids Research*, 2004, 32(Suppl_2): W273-W279.

[19] Katoh K, Standley D M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability[J]. *Molecular Biology and Evolution*, 2013, 30(4): 772-780.

[20] Suchard M A, Lemey P, Baele A G, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10[J]. *Virus Evolution*, 2018, 4(1): vey016.

[21] Ma X, Cai Z, Liu W, et al. Identification, genealogical structure and population genetics of S-alleles in *Malus sieversii*, the wild ancestor of domesticated apple[J]. *Heredity*, 2017, 119(3): 185-196.

[22] Nikiforova S V, Cavalieri D, Velasco R, et al. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line[J]. *Molecular Biology and Evolution*, 2013, 30(8): 1751-1760.

[23] Xiang Y Z, Huang C H, Hu Y, et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication[J]. *Molecular Biology and Evolution*, 2017, 34(2): 262-281.

[24] Minh B Q, Schmidt H A, Chernomor O, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era[J]. *Molecular Biology and Evolution*, 2020, 37(5): 1530-1534.

[25] Ya N L, Yan L L, Chao X, et al. The complete chloroplast genome sequence of *Malus toringo* (Rosaceae)[J]. *Mitochondrial DNA Part B*, 2020, 5(3): 2832-2833.

[26] Galtier N, Piganeau G, Mouchiroud D, et al. GC content evolution in mammalian genomes: The biased gene conversion hypothesis[J]. *Genetics*, 2001, 159(2): 907-911.

[27] Liu B B, Ren C, Kwak M, et al. Phylogenomic conflict analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving diversification, with insights into the complex biogeographic history in the Northern Hemisphere[J]. *Journal of Integrative Plant Biology*, 2022, 64(5): 1020-1043.

[28] Cui Dafang, Liao Wenbo, Yang Haijun, et al. Studies on the floristic composition and genesis of the wild fruit forest in Tian Shan Mountains in China[J]. *Forest Research*, 2006, 19(5): 555-560.

[29] Chu Jiayao, Feng Lingjiao, Hou Yixing, et al. Analysis on population damage of *Malus sieversii*[J]. *Forest Research*, 2022, 40(1): 265-273.

[30] Dong Hanlin, Wang Wenting, Xie Yun, et al. Climate dry-wet conditions, changes, and their driving factors in Xinjiang[J]. *Arid Zone Research*, 2023, 40(12): 1875-1884.

[31] Zhou Xiaodong, Chang Shunli, Wang Guanzheng, et al. Radial growth response of *Picea schrenkiana* to climate change in the middle section of the northern slope of the Tianshan Mountains[J]. *Arid Zone Research*, 2023, 40(8): 1215-1228.

[32] Zhao Zhuoyi, Hao Xingming. Actual evapotranspiration characteristics and attribution in arid Central Asia based on the Priestley-Taylor method[J]. *Arid Zone Research*, 2023, 40(7): 1085-1093.

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