RESEARCH

Open Access

Deciphering the multi- partite mitochondrial genome of *Crataegus pinnatifida*: insights into the evolution and genetics of cultivated Hawthorn

Lili Zhu[1†], Songfeng Diao[2,3*†], Taishan Li[2] and Juan Guo[3]

Abstract

Flowering plant (angiosperm) mitochondrial genomes are remarkably dynamic in their structures. We present the complete mitochondrial genome of hawthorn (*Crataegus pinnatifida* Bunge), a shrub that bears fruit and is celebrated for its extensive medicinal history. We successfully assembled the hawthorn mitogenome utilizing the PacBio long-read sequencing technique, which yielded 799,862 reads, and the Illumina novaseq6000 sequencing platform, which producing 6.6 million raw paired reads. The *C. pinnatifida* mitochondria sequences encompassed a total length of 440,295 bp with a GC content of 45.42%. The genome annotates 54 genes, including 34 that encode proteins, 17 that encode tRNA, and three genes for rRNA. A fascinating interplay was observed between the chloroplast and mitochondrial genomes, which share 17 homologous sequences sequences that rotal 1,933 bp. A total of 134 SSRs, 22 tandem repeats and 42 dispersed repeats were identified in the mitogenome. Four conformations of *C. pinnatifida* mitochondria sequences recombination were verified through PCR experiments and Sanger sequencing, and *C. pinnatifida* mitogenome is more likely to be assembled into three circular-mapping chromosomes. All the RNA editing sites that were identified C-U edits, which predominantly occurred at the first and second positions of the codons. Phylogenetic and collinearity analyses identified the evolutionary trajectory of *C. pinnatifida*, which reinforced the genetic identity of the hawthorn section. This unveiling of the unique multi-partite structure of the hawthorn mitogenome offers a foundational reference for future study into the evolution and genetics of *C. pinnatifida*.

Keywords Hawthorn, Mitochondrial genome, Homologous RecombinationMulti-partite, Phylogenetic analysis, RNA editing sites

Introduction

Mitochondria are the primary organelles derived from the maternal parent. They are only found in the cytoplasm of most eukaryotic cells, and the mitochondria produce energy to sustain the cells and are known as the "powerhouse" or "energy factory" of the cell [1, 2]. In addition, plant mitochondria are the only organelles that are likely to incorporate foreign DNA [3]. They are also involved in numerous metabolic processes, including programmed cell death [4], proliferation [5], the production of male sterility [6] and the synthesis

[†]Lili Zhu and Songfeng Diao contributed equally to this work.

*Correspondence:
Songfeng Diao
dsf@caf.ac.cn
[1] Academy of Chinese Medical Sciences, Henan University of Chinese Medicine, Zhengzhou 450046, China
[2] Research Institute of Non-Timber Forestry, Chinese Academy of Forestry, Zhengzhou 450003, China
[3] National Resource Center for Chinese Materia Medica, China, Academy of Chinese Medical Sciences, Beijing 100700, China

and degradation of several compounds [7]. Thus, after the first land plant mitogenome of common liverwort (*Marchantia polymorpha*) had been sequenced [8], many plant mitogenomes were sequenced and assembled. This practice has increased dramatically in recent years with the rapid development of high-speed sequencing technologies and improvements in the methods for assembly.

Plant mitogenomes vary dramatically in size, synonymous substitution rates, structural complexity and gene content among the sequenced land plants. The mitogenome sizes range from 66 kb to 11.3 Mb [9, 10], which can be attributed to their repetitive sequences and incorporation of foreign DNA by intracellular and horizontal gene transfer during the process of evolution [11, 12]. Plant mitogenomes are conventionally depicted as circular-mapping molecules, but they are not exclusively limited to a single circular chromosome model. For example, a complex in vivo structure that consisted of both monomeric circles and multimeric forms was found in mitochondrial genomes from the gynodioecious species bladder campion (*Silene vulgaris*) [13]. The highly modified mitogenome of the parasitic plant *Rhopalocnemis phalloides* is uniquely arranged in 21 minicircular chromosomes that vary in size from 4,949 to 7,861 bp, with minicircular and extremely heteroplasmic stuctures [14]. Moreover, the number of genes in the plant mitochondria also varies considerably, and the numbers of protein-coding genes and other genes are generally around 20 – 40 and 20 – 30 [15]. These genes encode relatively conserved proteins, such as ATP synthase, NADH dehydrogenase, ubiquinol cytochrome and cytochrome oxidase, these protein-coding genes (PCGs) have been widely used to resolve evolutionary relationships, whereas others, such as *sdh3*, *sdh4*, and *rps11* are not conserved and have been frequently lost in many angiosperms [16, 17]. Although mitochondrial DNA (mtDNA) in land plants exhibits significant evolutionary plasticity with peculiarities of gene expression, most notably RNA editing and *trans*-splicing, compared with the chloroplast and nuclear genomes, the mtDNA sequence evolved the most slowly with a low rate of mutations [18]. Thus, mitogenomes serve as an attractive reservoir for phylogenetic and evolutionary studies to trace the phylogeny of older land plant clades, which may help to provide new insights into the evolution of species. Although the mtDNA sequences are conserved in multiple species, the open reading frames (ORFs) have been disrupted by frameshift indels, particularly in ATP synthase genes, and are often involved in cytoplasmic male sterility [19], which is useful to improve breeding for heterosis in flowering plants.

Plants in the genus *Crataegus* are known as hawthorns. They are members of an interesting genus that are members of Maloideae subfamily and Rosaceae family, which consists of approximately 280 species and has a very long history of cultivation. There is still much disagreement over the origin and evolution of hawthorns, which were widely believed to have originated in North America or Europe [20, 21]. Currently, a chromosome-level genome of the hawthorn cultivar "Qiu Jinxing" has been assembled, which provided an improved context for understanding the evolution of Maleae [22]. Moreover, the chloroplast genomes of the three Hawthorns from *C. scabrifolia*, *C. chungtienensis* and *C. oresbia* were also reported, and they enriched the chloroplast genome resources of five *Crataegus* species [23]. The mtDNA may have different evolutionary histories compared with the chloroplast cpDNA and nuclear DNA [24]. However, there are no mitogenomes resources of hawthorns to provide novel insights into the evolutionary relationships.

Crataegus pinnatifida is a wild edible shrub that bears fruit and is widely cultivated in North China, it is known as Chinese hawthorn or "Shanzha" (in Chinese) [25, 26]. The desiccative ripe fruits are considered to be a medicinal herb that can improve digestion, invigorate blood circulation, enrich weakness, and provide other positive effects on the heart, lungs and stomach in Traditional Chinese Medicine (TCM) [27]. As an edible shrub that bears fruit, its fruits are also widely used as preserves and teas, and are a popular global dietary supplement [28]. *C. pinnatifida* has been shown to contain a variety of lignans [29, 30], phenylpropanoids [31, 32], flavonoids [33] and triterpenoids [34]. These are all active components that have widespread pharmacological effects, including hepatoprotective [35], cardiovascular protective [36], neuroprotective [37], anti-inflammatory [38] and anticancer effects [39].

This study assembled a complete mitochondrial genome of *C. pinnatifida*, and the organization of mitogenome is characterized by annotation, codon usage, repeat sequence and RNA editing. Gene transfer events between the chloroplast and mitochondrial genomes were also identified. Furthermore, the potential subgenomic structures of the *C. pinnatifida* mitochondrial genome that resulted from homologous recombination mediated by repeats is clearly illustrated. In addition, the mitochondrial genome organization of *C. pinnatifida* was compared to the genomes of previously published species of the Rosaceae to examine patterns of the evolution of organelle genome across this large angiosperm clade. Our data provide basic information on chimeric gene and provides a better understanding of the evolutionary processes of *C. pinnatifida* mitochondrial genomes.

Results

Assembly and structural characters of the *C. pinnatifida* mitogenome

A total of 799,862 reads that represent 13.8 Gb were generated based on the PacBio Sequel II platform, with an average read length of 17.3 Kb, the longest read and N50 length was 58.5 Kb and 18.0 kb (Table S1). To correct the draft mitogenome generated from the PacBio sequencing data, a total of 6.6 million raw paired reads that represented 9.96 Gb were generated using the Illumina novaseq6000 sequencing platform, and 6.2 million clean paired reads that represented 9.35 Gb were obtained for subsequent assembly (Table S2).

The *C. pinnatifida* mitogenome was assembled 440,295 bp long with a GC content of 45.42% (Table 1). The mitogenome is more likely to be assembled into three circular-mapping chromosomes, and their sequences were submitted to the NCBI Genome Database (Accession number, OR448911—OR448913). The largest assembled circular chromosome (Chr1) was 207,313 bp long, the two small assembled circular chromosomes were 119,854 bp (Chr2) and 113,128 bp long (Chr3) (Table 1, Fig. 1b). The obtained PacBio and Illumina reads were mapped to the assembly *C. pinnatifida* mitogenome, the rate of mapped reads is both 100%. The mean depth of coverage of long reads and Illumina reads

Table 1 Basic information of the *Crataegus pinnatifida* mitogenome

NCBI Accession number	Contigs	Type	Length (bp)	GC content (%)
OR448911—OR448913	Mt-Chromosome 1–3	branched	440,295	45.42
OR448911	Mt-Chromosome 1	circular	207,313	45.67
OR448912	Mt-Chromosome 2	circular	119,854	45.75
OR448913	Mt-Chromosome 3	circular	113,128	44.60

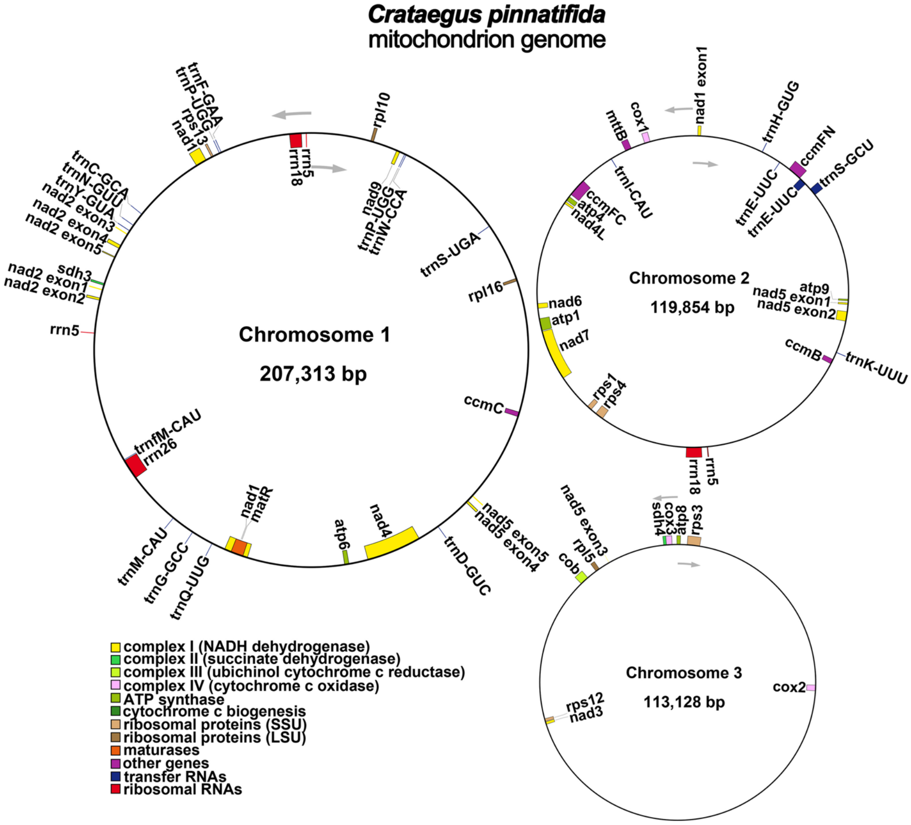

Fig. 1 The assembled mitogenome structure of *Crataegus pinnatifida*. Gene map showing 54 annotated genes of different functional groups. The genes on the outside are transcribed in the clockwise direction, and those inside the circle are transcribed in the counterclockwise direction

mapped to the assembly *C. pinnatifida* mitogenome was 98 X and 51 X, respectively (Table S1, S2, Fig. S1). The multi-partite mitogenome of *C. pinnatifida* was the most similar in size and GC content to those of the Amygdaloideae mitogenomes, such as loquat (*Eriobotrya japonica*), cherry (*Prunus avium*), plum (*Pyrus communis*), and European crab apple (*Malus sylvestris)* among others [2]. These results suggested the quality of assembly *C. pinnatifida* mitogenome was high.

Annotation of the *C. pinnatifida* mitogenome

A total of 39,628 bp were annotated coding sequences, which comprised 9% of the whole mitogenome, while more than 90% were located in intergenic regions. There were 54 genes annotated in the *C. pinnatifida* mitogenome, including 34 PCGs, which were divided into 24 unique core genes and 10 variably present protein genes, as well as 17 tRNA genes and three rRNA genes (Table 2). These unique core genes mostly encode for respiratory proteins. This mitogenome did not have large repeated regions, and most of the annotated genes were single-copy genes. Several of the annotated genes, included *nad4* that encodes NADH dehydrogenase; *trnE-UUC* and *UGG,* that are both tRNA genes; and *rrn18* that is rRNA gene with two copies. Only the *rrn18* rRNA gene has three copies. According to the annotation, those 24 unique core genes could be divided into seven classe, ATP Synthase (five genes), NADH dehydrogenase (nine genes), cytochrome c biogenesis (one gene), ubiquinol cytochromec reductase (four genes), cytochrome c oxidase (three genes), maturases (one gene) and transport membrane protein (one gene). Moreover, 10 non-core genes were identified, including three from the large ribosomal subunit, five from the small ribosomal subunit, and two associated with succinate dehydrogenase.

All three chromosomes of the mitogenome can encode proteins. There are 29 genes on Chr1 (13 complete PCGs, 13 tRNA genes and all three rRNA genes), 23 genes on Chr2 (15 complete PCGs, six tRNA genes and two rRNA genes), and 10 genes on Chr3 (10 complete PCGs). Most of the PCGs were annotated on a single chromosome, while the exons of the gene *nad1* were distributed on Chr1 and Chr2, and the exons of *nad5* were distributed on three different chromosomes (Fig. 1).

Gene transfer between the mitogenome and chloroplast genome

The transfer of DNA sequences from the chloroplast genome to the mitogenome is an important event in the evolution of higher plants. To better understand the intracellular gene or sequence transfer events in *C. pinnatifida*, the individual chloroplast genome of *C. pinnatifida* was assembled into a single, circular mapping molecule that contained a pair of inverted repeat (IR) regions, which were divided by short-single copy (SSC) and long-single copy (LSC) regions (Fig. S2). The mitochondrial genome of *C. pinnatifid*a is approximately 2.7 times larger than those of the chloroplast genome, at 159,657 bp long. The mitochondrial genome was used as a query and compared to the chloroplast genome, and 1,933 bp sequences were transferred from the chloroplast to the mitogenom. These accounted for 0.44% of the mitochondrial genome and 1.21% of the chloroplast. A total of 17 recombinant fragments of mitogenomes (MTPTs) were homologous with specific regions of the chloroplast genome (Fig. 2). The longest homologous recombinant fragment, designated MTPT17, was 646 bp

Table 2 Gene annotation of the *Crateagus pinnatifida* mitogenome

Gene	Group	Gene names
Core genes	ATP synthase	*atp1*, *atp4*, *atp6*, *atp8*, *atp9*
	NADH dehydrogenase	*nad1*, *nad2*, *nad3*, *nad4* (×2), *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*
	Cytochrome c biogenesis	*cob*
	Ubiquinol cytochrome c reductase	*ccmB*, *ccmC*, *ccmFC*, *ccmFN*
	Cytochrome c oxidase	*cox1*, *cox2*, *cox3*
	Maturase	*matR*
	Transport membrane protein	*mttB*
Variable genes	Large subunit of ribosome	*rpl5*, *rpl10*, *rpl16*
	Small subunit of ribosome	*rps1*, *rps3*, *rps4*, *rps12*, *rps13*
	Succinate dehydrogenase	*sdh3*, *sdh4*
rRNA genes	Ribosome RNA	*rrn5* (×3), *rrn18* (×2), *rrn26*
tRNA genes	Transfer RNA	*trnC-GCA*, *trnD-GUC*, *trnE-UUC* (×2), *trnF-GAA*, *trnfM-CAU*, *trnG-GCC*, *trnH-GUG*, *trnI-CAU*, *trnK-UUU*, *trnM-CAU*, *trnN-GUU*, *trnP-UGG* (×2), *trnQ-UUG*, *trnS-GCU*, *trnS-UGA*, *trnW-CCA*, *trnY-GUA*

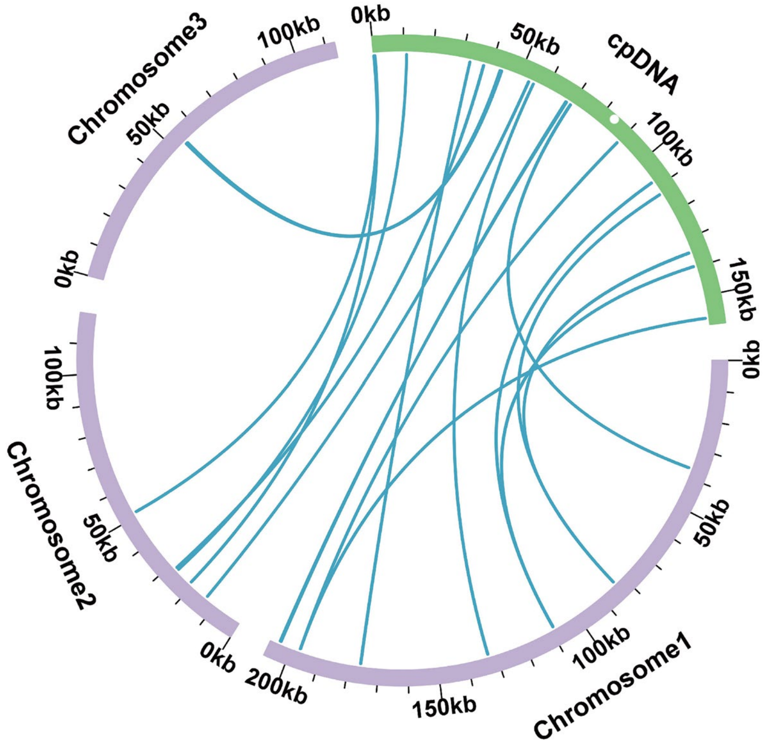

Fig. 2 The distribution of MTPTs of *Crateagus pinnatifida*. The mitochondrial genome and chloroplast genome were marked with purple and green, respectively. The blue ribbons show the MTPTs between the mitochondrial and chloroplast genomes. MTPTs 1–11 were found in chromosome 1, and MTPTs 13–16 were found in chromosome 2. MTPT 17 was found in chromosome 3. MTPTs, recombinant fragments of mitogenomes

long (Table S3). Furthermore, six complete genes were annotated and distributed in the recombinant fragments, which were all tRNA genes, including *trnD-GUC, trnH-GUG, trnI-CAU, trnM-CAU, trnN-GUU* and *trnW-CCA* (Table S4).

Analysis of codon usage in the PCGs in *C. pinnatifida* mitochondrial genome

AUG is a typical start codon in plant mitochondria, but plant mitochondria may also use several other start codons. Most of these PCGs annotated in the *C. pinnatifida* mitogenome have a typical AUG start codon, but the start codon of *cox1*, *nad1* and *rps4* is ACG; that of mttB is AUA and GUG for *rpl16*, which could be the result of RNA editing. In addition, three different typical stop codons were also identified in most of the PCGs, including UAA, UGA and UAG. The codons of mutation sites were also found in the stop codons, the first codon position U was replaced with C. For example, those of *atp6* and *atp9* were CAA and CGA, respectively (Table S4).

The relative synonymous codon usage (RSCU) of the *C. pinnatifida* mitochondrial PCGs was analyzed (Fig. 3, Table S5). RSCU = 1 indicates no preference for codon use, while RSCU > 1 indicates that the codon is a relatively frequently used. As shown in Fig. 3**,** all the codons are present in the PCGs. The most frequently used codons included the stop codon (UAA, 1.55), Tyr (UAU, 1.54), and Ala (GCU, 1.51), whose RSCU values > 1.5. Moreover, the RSCU values of almost all the codons with the third codon position A/T ≥ 1.0. Conversely, the RSCU values of almost all the codons with the third codon position C/G ≤ 1.0 (Table S5). This phenomenon indicates that there is a high content of A/T at the third codon position in the mitogenome of *C. pinnatifida*, which is very similar to what has been reported in the mitogenomes of other land plants. The strong AT bias of the third codon is considered to be a universal phenomenon in higher plants.

Repeat sequence analysis of the *C. pinnatifida* mitochondrial genome

Simple sequence repeats (SSRs) are singular, repeated sequences with one to six bases motifs that are widely used as DNA markers to study genetic diversity and identify species. In this study, 69, 36 and 29 SSRs were identified in Chr1, Chr2 and Chr3 of the *C. pinnatifida* mitogenome, respectively (Table S6). Among the 134 SSRs, more than 50% were monomeric and dimeric repeats. There were slightly fewer tetrameric repeats than monomeric ones in three Chrs. In contrast, there

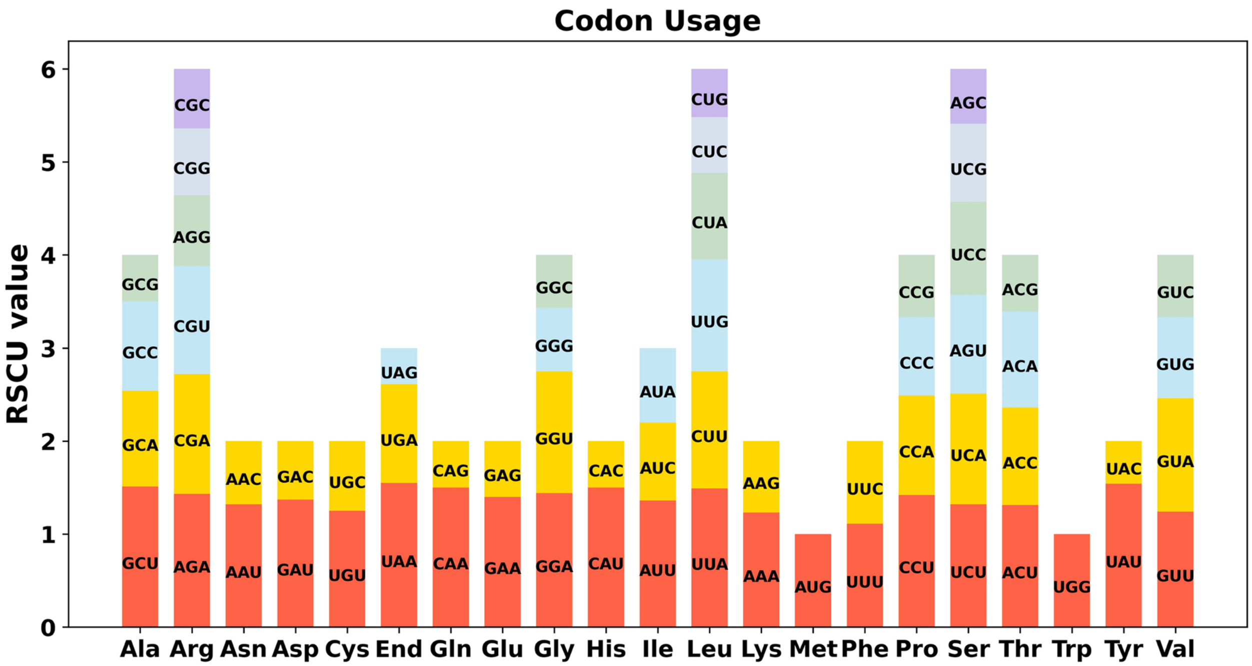

Fig. 3 Codon usage analysis of the PCGs in the *Crataegus pinnatifida* mitochondrial genome. Different codons of the same amino acid were marked with red, yellow, blue, green gray, and purple bars, respectively. The "end" represents stop codons. PCGs, protein-coding genes

were fewere trimeric repeats than the dimeric repeats, and there was only one pentameric; there were no hexameric repeats in the *C. pinnatifida* mitogenome (Fig. 4a). An additional analysis of the SSR repeat units showed that A/T was more prevalent than the other repeat types, and it contributes to the AT richness of the complete *C. pinnatifida* mitogenome.

Tandem repeats, with lengths that range from 7 to 200 base, are widely found in the mitogenomes of plants. A total of 22 tandem repeats were identified in the mitogenome of *C. pinnatifida*, with unit lengths that ranged from 8 to 39 bp. A total of 11 tandem repeats were distributed on Chr1, six on Chr2, and five on Chr3 (Table S7, Fig. 4b). Moreover, many non-tandem long repeat sequences with repeat units ≥ 50 bp, except for the SSRs and tandem repeats, were distributed in plant mitogenome. Those dispersed repeats were classified into forward repeats, palindromic repeats, reverse repeats and complement repeats. ROUSfinder [40] was used to show that 42 dispersed repeats ≥ 50 bp were found in the *C. pinnatifida* mitogenome, including 27 palindromic repeats and 15 forward repeats. Among these repeats, Repeat_1 had the longest sequences of 12,818 bp, followed by Repeat_2 with sequences of 64,971 bp (Table S8). The pairs of the long repeats often cause isomerization of the genome by recombination.

Homologous recombination mediated by repeats

The assembly is composed of six assembled contigs, demonstrating a region of overlap along the linkages (Fig. 5a). It's worth noting that ctg5 (Repeat_1) and ctg6 (Repeat_2) exhibited distinct characteristics suggestive of potential repetitive sequences. To explore the potential subgenomic structures of the *C. pinnatifida* mitochondrial genome, the BLASTN program [41] was used to identify the repeats in three chromosomes. First, a unitig graph that contained two double bifurcating structures (DBS) (Fig. 5a) was archived by extracting the mitochondrial short reads using GetOrganelle [42] combine with performing the de novo assembly of the extracted reads using Unicycler software [43]. To determine whether these two repeats can mediate recombination, we extracted sequences that contained the flanking regions and the repeat sequences. Four alternative conformations were generated by artificially swapping extended 2,000 bp sequences to simulate recombination, and the recombinant sequence represents four potential genomic paths, we then denoted the induced genomic paths as P1-1, P1-2, P2-1 and P2-2 (Fig. 5a). To confirm the presence of these paths within the *C. pinnatifida* mitogenome, we conducted PCR experiments. Eight specific primer pairs based on 1,000 bp sequences contaning repeated and flanking regions sequences were employed to validate P1-1, P1-2, P2-1 and P2-2genomic paths deduced

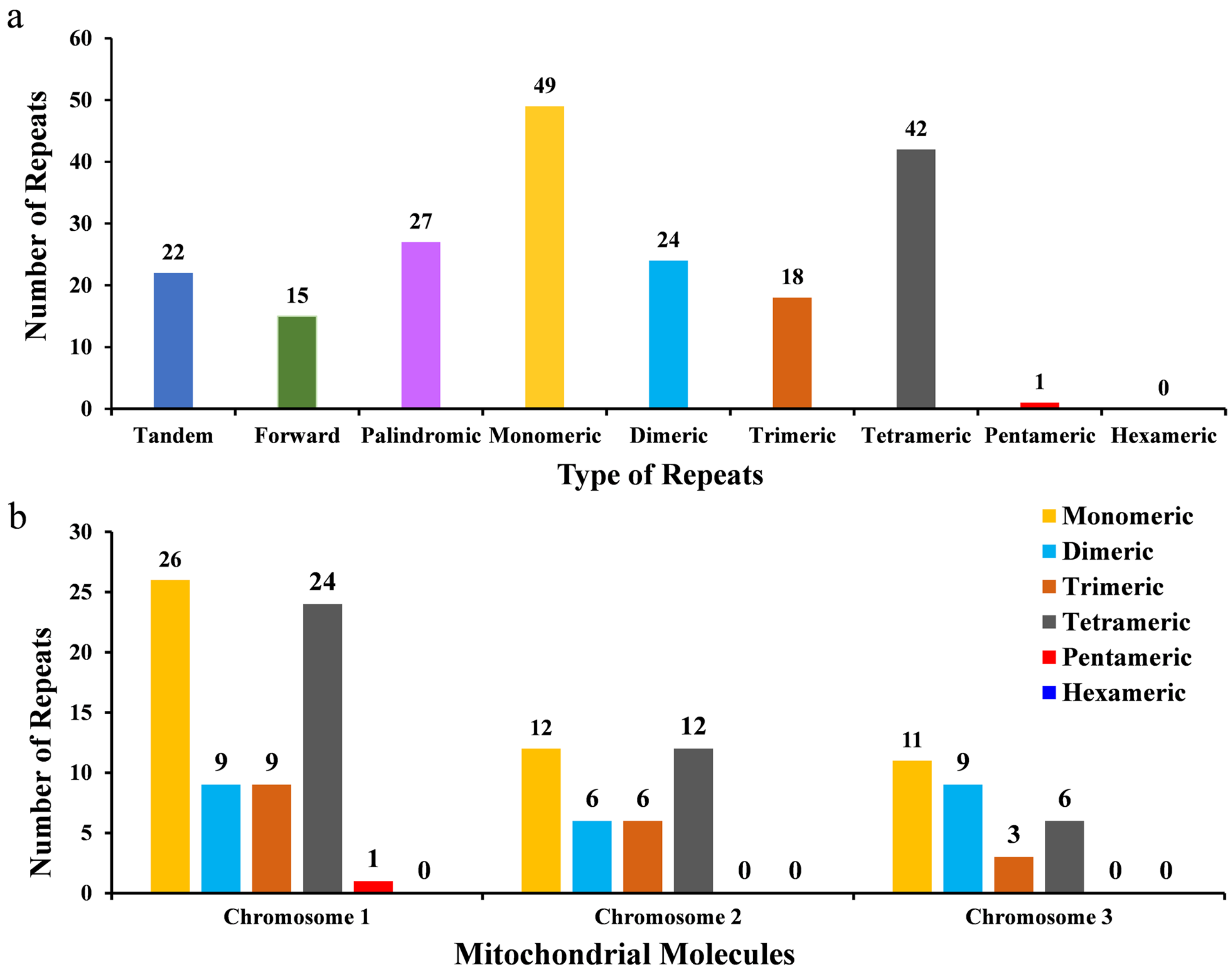

Fig. 4 Repeat analyses in the *Crataegus pinnatifida* mitochondrial genome. **a** Number of different SSR repeat units distributed in three chromosomes. **b** Number of tandem repeats and different long repeat sequences distributed in the three chromosomes. SSR, simple sequence repeats

by switching the flanking regions. The PCR amplifcation results showed eight bands having lengths consistent with those expected (Fig. 5b), and Sanger sequencing results confirmed possible genomic paths (Fig. S3).

There were four possible genomic conformations in *C. pinnatifida* mitochondrial genome. Conformation 1 (C1) presents three independent circular structures, incorporating all six contigs and P1-1 and P1-2 genomic paths both induced by ctg5 and ctg6, respectively (Fig. 5c). Conformation 2 (C2) presents two independent circular structures, ctg2, ctg3, ctg4, P1-2 genomic path deduced by ctg5 and P2-2, P2-1 genomic paths deduced by ctg6 collectively form a bigger circular arrangement, as well as ctg1 and P1-1 genomic path deduced by ctg5 form a smaller circular (Fig. 5d). Similarly, conformation 3 (C3) also showed two circular arrangement, ctg1, ctg3, ctg4, P2-1 and P2-2 genomic paths deduced by ctg5 collectively form a circular arrangement, as well as ctg2 and P1-1 genomic path deduced by ctg6 form a smaller circular (Fig. 5e). Unlike C1, C2 and C3, conformation 4 (C4) presents a master circular structure encompassing all six contigs, P2-1 and P2-2 genomic paths induced by ctg5, and P2-1 and P2-2 genomic paths genomic path induced by ctg6 (Fig. 5f).

Furthermore, the long PacBio reads were mapped to the eight genomic paths to determine the recombination frequency. A total of 34 high scoring pairs (HSPs) were identified that were associated with the four possible conformations (Table S9). Conformations induced by p1-1 and p1-2 genomic paths were denoted as major conformations, on the contrary conformations induced by p2-1 and p2-2 genomic paths were denoted as minor conformations. The percentage of major conformations and minor conformations was 71.43% and 28.57% based on ctg5, respectively (Table 3). In addtion, the percentage of the major and minor conformations was 77.78% and

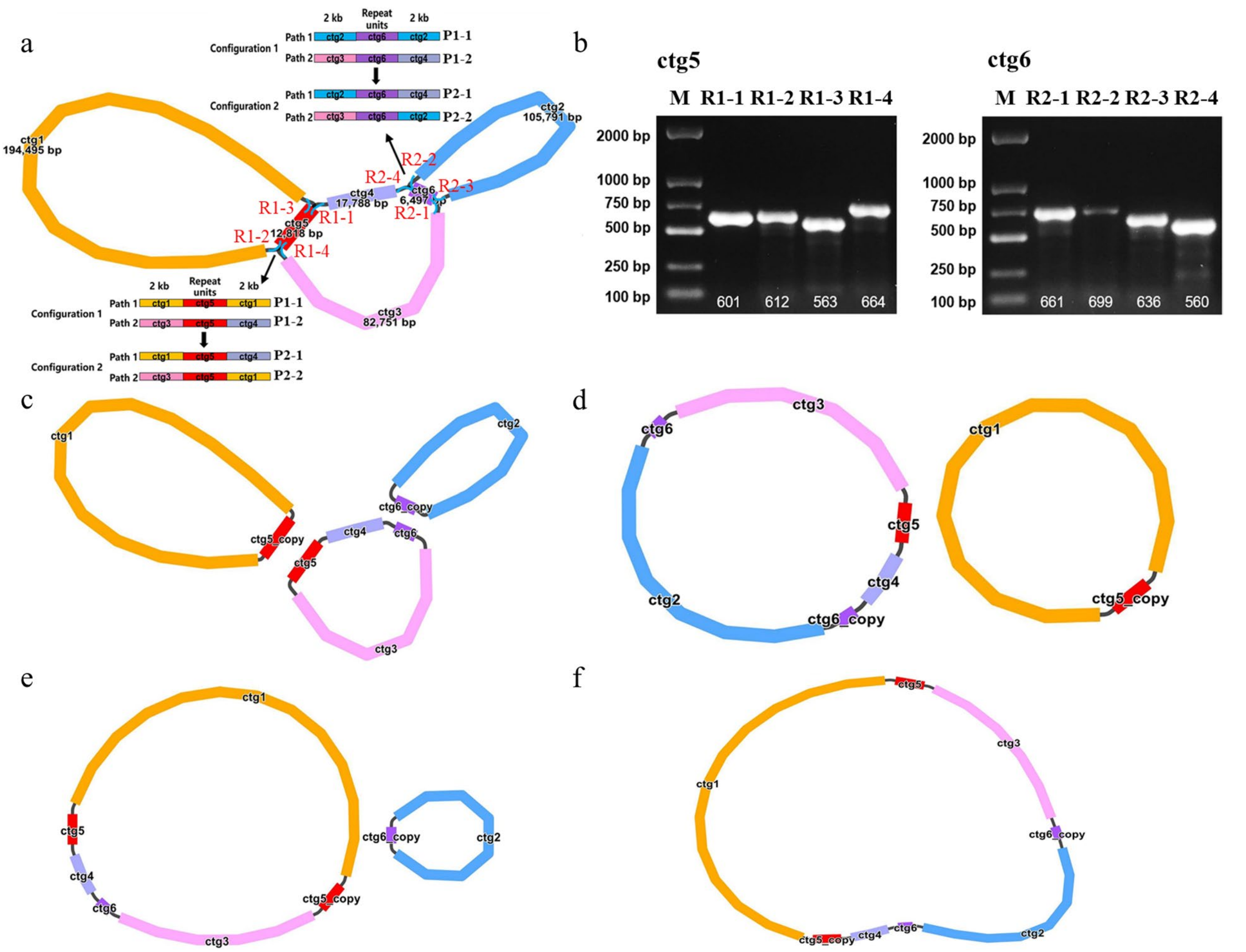

Fig. 5 The graphic assembly and verification of *Crataegus pinnatifida* mitogenome. **a** Unitig graph of the C. pinnatifida mitogenome obtained from the de novo assembly of Illumina reads with Unicycler, ctg5 and ctg6 are the repetitive sequence. **b** Gel electrophoresis image for the PCR products. M. marker; 1–8 represents the electropherogram of eight paths, and the expected lengths of each fragment are shown at the bottom of the gel. **c**-**f** Four potential conformations that result from the rearrangement mediated by two pairs of repeat sequences, respectively

Table 3 The percentage of different genome configurations identified by long-reads in the *Crataegus pinnatifida* mitogenome

ID	Length	Number of Reads Supporting the Paths				Percentage of configuration (%)	
		p1-1	p1-2	p2-1	p2-2	Major	Minor
ctg5	12,818 bp	2	3	2	0	71.43	28.57
ctg6	6,497 bp	10	11	2	4	77.78	22.22

22.22% based on ctg6, respectively (Table 3). Accordingly, C1 has the highest probability of 55.56%, followed by C2 has the probability of 15.87%. Specifically, it seemed that there was no probability of C2 and C4, due to no mapped reads that supported P2-2 genomic path induced by ctg5. Although our results strongly supported the hypothesis that *C. pinnatifida* mitochondrial sequences were separated from each other, and formed three independent circular DNA molecules (Fig. 5b). The three circular-mapping chromosomes of *C. pinnatifida* mitogenome could be rearranged into two circular-mapping chromosomes at a low rate. In addion, three independent circular DNA molecules cloud be rearranged into one bigger circular-mapping chromosome accordding to the PCR amplifcation and Sanger sequencing results. Overall, *C. pinnatifida* is more likely has a multi-partite mitogenome which can be assembled into three circular-mapping chromosomes.

Prediction of the RNA editing sites of pcgs

RNA editing refers to the addition, loss, or conversion of the base in the coding region of the transcribed RNA. This is a common phenomenon in plant mitochondria that regulates the levels of expression of the genes involved in plant growth and development. This study used the predicted PREPACT3 program to predict 456 RNA editing sites within 33 unique PCGs in the mitogenome of *C. pinnatifida*. All the RNA editing sites were C-U edits that occurred in the first (163, 35.75%) and second (293, 64.25%) positions of the codon, whereas no RNA editing was found in the third position, which is similar to the situation in most angiosperms (Table S10). Furthermore, RNA editing caused the substitution of eight amino acids (Aa), and the replacement of serine (S) with leucine (L) occurred most frequently. Among those PCGs, *nad4L* does not have any editing site predicted, while *rpl10* and *ccmFN* have the least (1) and most editing sites (38) predicted, respectively (Fig. 6).

Phylogenetic analysis of *C. pinnatifida*

To understand the evolutionary status of the *C. pinnatifida* mitochondrial genome, a phylogenetic tree of *C. pinnatifida* together with the other 30 species, including 24 species of Rosaceae and two of Rhamnaceae, Cannabaceae and Fabaceae (designated as outgroups), was created. The phylogenetic tree was obtained based on an aligned data matrix of 25 conserved PCGs, including *atp1*, *atp4*, *atp6*, *atp8*, *ccmB*, *ccmC*, *ccmFC*, *ccmFN*, *cob*, *cox1*, *cox2*, *cox3*, *matR*, *nad1*, *nad2*, *nad3*, *nad4*, *nad5*, *nad6*, *nad7*, *nad9*, *rpl16*, *rps3*, *rps4* and *sdh4*, from these species. The taxa from the 24 Rosaceae species were well clustered, and the order of taxa in the phylogenetic tree is strongly consistent with the evolutionary relationships of those species according to the latest Angiosperm Phylogeny Group (APG). As shown in Fig. 7, *C. pinnatifida* was the most closely related to Chinese quince (*Chaenomeles speciosa*), which suggests that the Crataegus and Chaenomeles have a closer evolutionary relationship than the other species of Rosaceae.

Collinearity between the different *Rosaceae* mitogenomes

An analysis of the locally collinear blocks (LCBs, > 0.5 kb) is more suitable to identify large-scale rearrangement, gene gain and gene loss. The mitogenomes of *C. pinnatifida* and the closely related species, including pear (*Pyrus communis*), *Malus sieversii*, Chinese quinc, loquat (*Eriobotrya japonica*), apricot (*Prunus armeniaca*) and the Iturup strawberry (*Fragaria iturupensis*), were compared and subjected to collinear analysis to explore the similarities and differences between the multi-ring and single-ring structures in more detail. A significant number of LCSs was detected among the seven Rosaceae species. However, the multiple synteny plots were not arranged

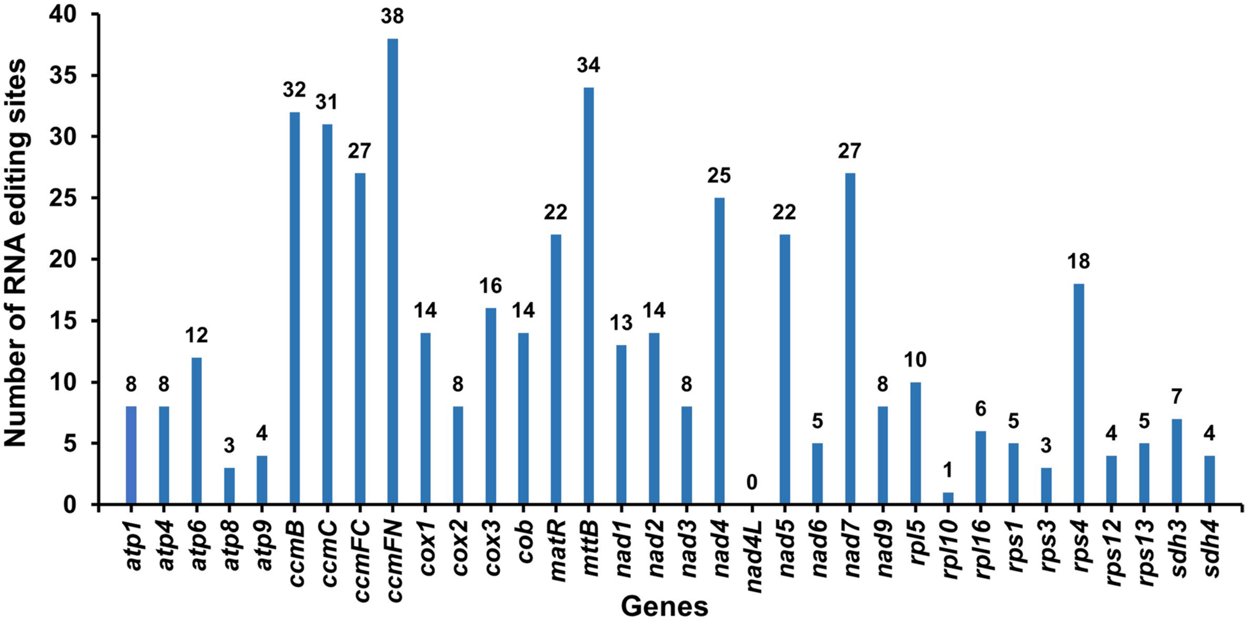

Fig. 6 Number of the RNA editing events in the PCGs of *Crataegus pinnatifida* mitochondrial genome

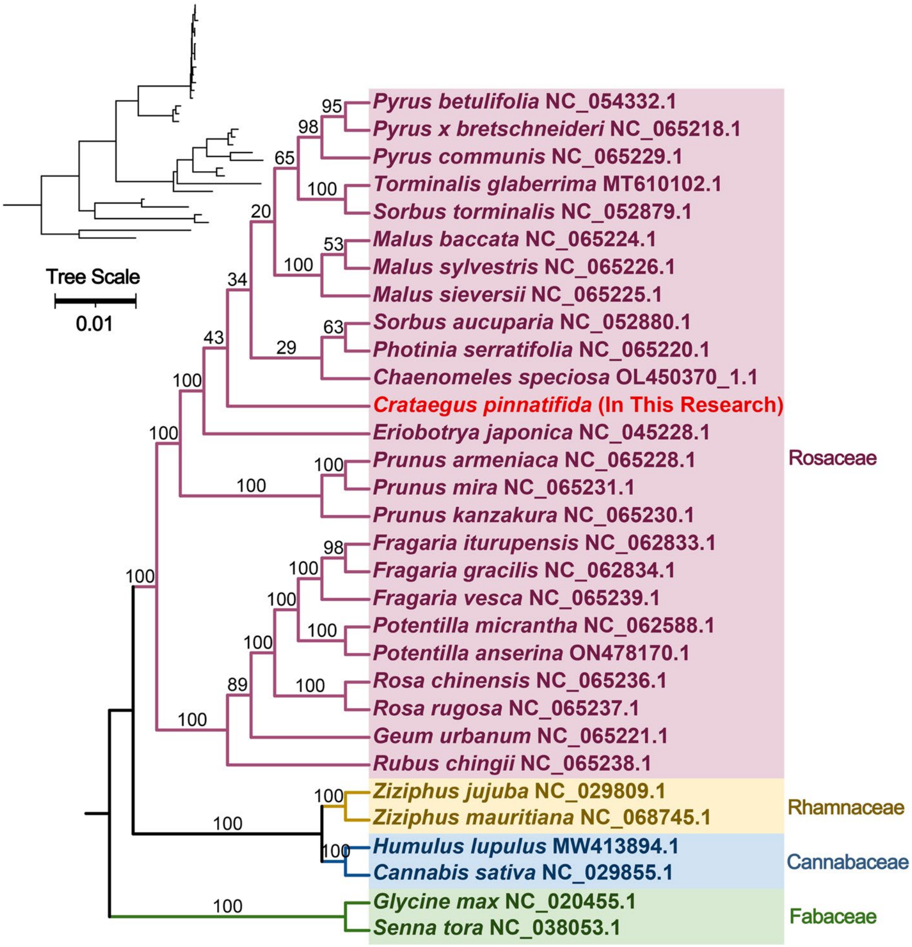
Tree Scale
0.01
95
98
65
100
20
53
100
34
63
29
43
100
100
100
100
100
98
100
100
100
100
89
100
100
100
100
100
100
Pyrus betulifolia NC_054332.1
Pyrus x bretschneideri NC_065218.1
Pyrus communis NC_065229.1
Torminalis glaberrima MT610102.1
Sorbus torminalis NC_052879.1
Malus baccata NC_065224.1
Malus sylvestris NC_065226.1
Malus sieversii NC_065225.1
Sorbus aucuparia NC_052880.1
Photinia serratifolia NC_065220.1
Chaenomeles speciosa OL450370_1.1
Crataegus pinnatifida (In This Research)
Eriobotrya japonica NC_045228.1
Prunus armeniaca NC_065228.1
Prunus mira NC_065231.1
Prunus kanzakura NC_065230.1
Fragaria iturupensis NC_062833.1
Fragaria gracilis NC_062834.1
Fragaria vesca NC_065239.1
Potentilla micrantha NC_062588.1
Potentilla anserina ON478170.1
Rosa chinensis NC_065236.1
Rosa rugosa NC_065237.1
Geum urbanum NC_065221.1
Rubus chingii NC_065238.1
Ziziphus jujuba NC_029809.1
Ziziphus mauritiana NC_068745.1
Humulus lupulus MW413894.1
Cannabis sativa NC_029855.1
Glycine max NC_020455.1
Senna tora NC_038053.1
Rosaceae
Rhamnaceae
Cannabaceae
Fabaceae

Fig. 7 Phylogenetic tree based on 25 conserved PCGs. Two mitochondrial PCGs from sicca senna (*Senna tora*) and soybean, were settled as the outgroups, and the bootstrap support values were recorded at eachnode. PCGs, protein-coding genes

in the same order among the individual mitochondrial genomes (Fig. 8). This suggested that hawthorn and the other six Rosaceae species experienced extremely frequent mitochondrial genome rearrangements to obtain new genomes through evolution. In addition, hawthorn, *M. sieversii*, Chinese quinc, and loquat had a good collinear relationship and shared numerous conserved colinear blocks. Conversely, there were relatively weak locally collinear blocks among hawthorn, pear, apricot and Iturup strawberry.

Discussion

Features of the *C. pinnatifida* mitochondrial genome

Plant mitogenomes are more complex than those of animals and most other eukaryotes because of their variation in size and repeated seqeunces [1]. In this study, we utilized short reads of Illumina novaseq 6000 sequencing in combination with long reads of PacBio RS II sequencing to complete the mitogenome of *C. pinnatifida*. This study produced the first detailed characterization of a complete mitogenome in *Crataegus*. The *C. pinnatifida* mitogenome encompassed a total length of 440,295 bp with a GC content of 45.42%. The *C. pinnatifida* mitogenome had a moderate content of GC and size relative to the mitogenomes of most other Rosaceae species. The GC content and size of *C. pinnatifida* mitogenome are the most similar to the mitogenome of loquat (Size: 434,980 bp, GC content: 45.42%), Cherry ("Glory," Size: 444,576 bp, GC content: 45.62%) and birchleaf pear (*Pyrus betulifolia*) (Size: 432,493 bp, GC content: 45.21%) [2, 44], but they were smaller than those of some gymnosperm, such as a gnetophyte (*Welwitschia mirabilis*) (Size: 978,846 bp, GC content: 53.00%) [45].

Although the genome size plant mitochondrial greatly varied, the number of mitochondrial genes is relatively conserved in the land plant lineage. Most sequences in the *C. pinnatifida* mitogenome are non-coding, while 39,628 bp was annotated PCGs that comprised 9% of the

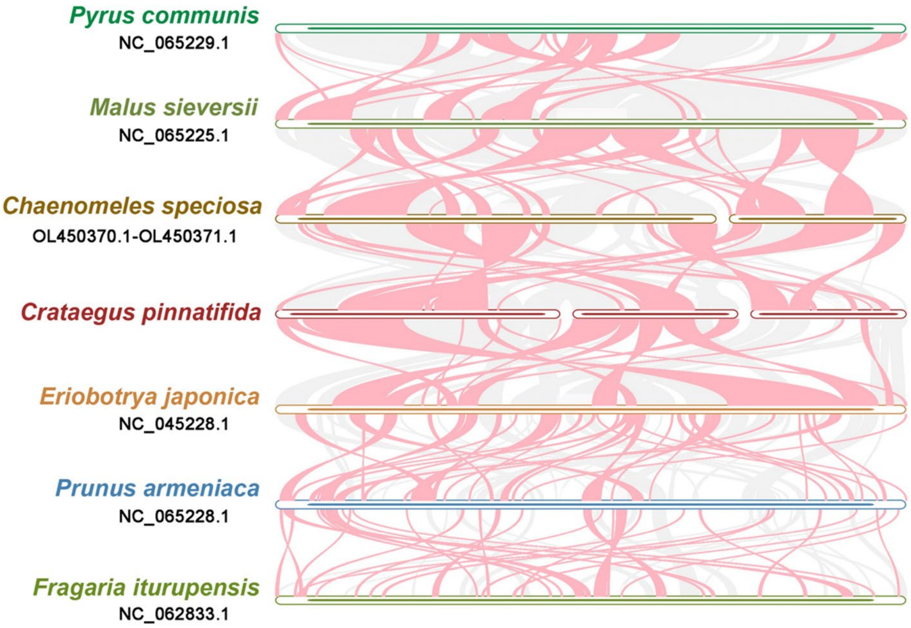

Fig. 8 Genome synteny analysis of seven species of Rosaceae. The ribbons show the LCBs among the seven species. Highly conserved LCBs are shown in gray ribbons, and the pink ribbons indicate where the inversion occurred. The interstices between the LCBs represent the unique fragments in each mitogenome. LCBs, locally colinear blocks

whole mitogenome. This percentage was probably owing to a gradual increase in the duplication of sequences during evolution. Moreover, most PCGs of the *C. pinnatifida* mitogenome had the typical AUG start codon, and the distribution of amino acids was also similar to that of other angiosperms. In contrast, the start codon of *cox1*, *nad1* and *rps4* is ACG, *mttB* is AUA and GUG for r*pl16*. This phenomenon also has been found in other studies. Repeat-induced duplicated genes are widespread in maize (*Zea mays*), sacred lotus (*Nelumbo nucifera*) and some vascular plant mitogenomes, 10 and six duplicated PCGs have been identified in the maize [46] and sacred lotus [47] mitogenomes, respectively. There are varying numbers of copies of *Cox1* in the populations of different species, and it is frequently involved in horizontal gene transfer among angiosperms [48]. However, only *nad4* is found had two copies, and no copies of *cox1* were found in the *C. pinnatifida* mitogenome. This suggests that the recombination of repeat in the homogenization of the gene copies harbored appeared to be less frequent.

The conformations of *C. pinnatifida* mitochondrial genome mediated by repeats and homologous recombination

Plant mitochondrial genomes are conventionally reported as circular structures. Nevertheless, they are far more complex than a simple circular chromosome model and are typically described as multipartite and interconverting even within a single individual [49]. The repeats in mitochondrial DNA provide a solid foundation for intermolecular recombination. They are considered to be one of the most important factors that induces mitochondrial genomes composed of multiple circulars, branched, linear or mixed forms of genomic structure [50, 51]. The repeats are larger and more frequent in the vascular plants. Non-tandem repeats usually include pairs of large repeats (> 1000 bp) that may cause isomerization of the genome by recombination, whereas short- and medium-sized repeated sequences (< 1000 bp) tend to recombine at minor to moderate levels [11, 40, 52].

In *C. pinnatifida* mitogenomes, the number of dispersed repeat sequences, particularly palindromic repeats and forward repeats, was larger than those of the SSRs and tandem repeats (Fig. 4). This indicates that moderately repetitive sequence may be the primary contributor to the variation in observed sizes and structures. This prolific presence of repeats points towards their potential significance not only in size but also in influencing genome reconfiguration. Multiple configurations can be owing to a combination of linear, circular, and branched molecules in the plant mitogenomes. Some studies propose that structural variations within plant mitogenomes may arise from long repeat-mediated recombination. Within the mitogenome of *C. pinnatifida*, our analysis unveiled two pairs of lengthy repetitive sequences, measuring 12,818 bp and 6,497 bp, respectively (Table 3). To determine whether two long repeated sequences can mediate recombination, four alternative conformations were generated by simulating recombinationin *C. pinnatifida* mitogenome (Fig. 5a). Furthermore, four conformations were verified through PCR experiments and Sanger sequencing (Fig. 5b) combined with long PacBio mapped reads (Table S9). Our results suggested that *C. pinnatifida* mitogenome was most likely be formed three independent circular DNA molecules mediated, while three circular-mapping chromosomes could be rearranged into two or one circular-mapping chromosome. This indicates that the mitogenome structure of *C. pinnatifida* is not static but rather dynamically varies among these two conformations. These findings were similar to those found in the mitogenomes of Chinese quince [53], pear [54], Chinese photinia (*Photinia serratifolia*) [55], which suggested that circular arrangements appear to be the common form among the assembled Rosaceae family mitogenome induced by the recombination of repeated sequences[2].

Intergenomic sequence transfers and phylogenetic analysis of *C. pinnatifida* mitochondrial genome

Intergenomic DNA transfer events between different genomes (mitochondrial, nuclear and chloroplast) are a common and foundation phenomenon that occurs during the evolution of plants. Most studies have indicated that the transfer of sequences from the organellar genomes to the nuclear genome and from the nuclear genome and chloroplast genome to the mitogenome were the primary directions [56, 57]. The transfer of sequences was found to be induced by the illegitimate repair of double-stranded breaks or as part of the stress or other responses that induce functional genetic diversity in the receiving genome [58, 59]. In this study, 17 recombinant fragments that were sequences of 1,933 bp long were transferred from the chloroplast to the mitogenome of *C. pinnatifida* and comprised 1.21% of the chloroplast and 0.44% of the mitogenome, respectively (Fig. 2, Table S4). Furthermore, six complete genes that were annotated with transferred sequences that were shared between the nuclear and mitochondrial genomes all encoded tRNA, including *trnD-GUC*, *trnH-GUG*, *trnI-CAU*, *trnM-CAU* , *trnN-GUU* and *trnW-CCA* (Table S4). The transfer of tRNA genes from the chloroplast to mitochondrial DNA is common in angiosperm.

Plant mitochondria genomes are well-known to evolve rapidly compared with other organelles induced by heteroplasmy, genome recombination, and gene chimerism among others [60]. With the development of sequencing technology, the sequences of numerous mitochondria genomes of the Rosaceae family are publicly available. The accession number is listed in Fig. 6. The phylogenetic evolution results derived from mitochondria genomes of the *C. pinnatifida* and other 23 Rosaceae species strongly correspond to the evolutionary relationships of those species according to the latest APG. Hawthorn was placed in the Amygdaloideae clade and is the most closely related to the Chinese quince as a sister lineage to the combined clade of Eriobotrya, Malus, Sorbus and Pyrus, which agreed with the findings of a previous study based on the nuclear genome [22]. Collinearity analyses between C. *pinnatifida* and the mitogenomes of the most closely related species also strongly supported the results of the analysis of phylogenetic evolution. *C. pinnatifida*, *M. sieversii*, Chinese quince, and loquat had a a better collinear relationship and shared more conserved colinear blocks than the other three species of Rosaceae (Fig. 7).

Conclusions

In this study, we assembled and annotated the complete mitochondrial genome of *C. pinnatifida*, a wild shrub that bears edible fruit that are also medicinal, using Illumina short reads and PacBio long reads. The mitochondrial genome of *C. pinnatifida* was the most likely assembled into three circular chromosome structures acrroding to the recombination mediated by repeats. The genome size was 440,295 bp with a GC content of 45.42%, which is similar to those of loquat, cherry, and birchleaf pear. A total of 54 genes were distributed on three chromosomes, and the number of genes positively correlated with the size of each chromosomes. We also identified 17 large fragments that were 1,933 bp long that had been transferred from the chloroplast genome to the mitogenome. In addition, phylogenetic and collinearity analyses based on the mitochondrial genome of *C. pinnatifida* both supported its evolution, and it has a closer genetic evolutionary relationship with Chinese quince, *M.sieversii*, and loquat compared with the other species of Rosaceae. Our findings provide a valuable genetic resource for further studies on the genus *Crataegus*.

Materials and methods

Plant materials, DNA extraction and sequencing

C. pinnatifida plants were collected at the Non-timber Forestry Research and Development Center, the forest planting base in Zhengzhou, Yuanyang County, China (34°55′-34°56′ N, 113°46′-113°47′ E). Fresh leaves were collected, frozen in liquid nitrogen and stored at -80 °C. The genomic DNA was extracted from frozen leaves using the CTAB method [61]. The purity and concentration of DNA were determined using a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and 1% agarose gel electrophoresis, respectively, and stored at -80 °C.

Purified genomic DNA was used to construct a SMRT sequencing library with an insert size of 15 kb. The genome was sequenced using the PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA) at SMART GENOMICS Technology Limited (Tianjin, China). In addition, for leaf samples from the same individual, sequencing libraries of 150 bp paired-end reads with an insert size of 300 bp were constructed using a library construction kit (Illumina, San Diego, CA, USA). The libraries were sequenced using an Illumina Novaseq 6000 sequencing platform by SMART GENOMICS Technology Limited. To ensure the quality of reads, Trimmomatic v 0.39 [62] was used to filter out raw reads that contained adapter contamination and duplication.

Mitogenome assembly and annotation

The mitogenome of *C. pinnatifida* was assembled using the data obtained from the PacBio RS II sequencing in combination with Illumina NovaSeq 6000 sequencing. The contigs were assembled based on Illumina paired-end reads using GetOrganelle v. 1.7.5 [42] with default parameters. Subsequently, Bandage soft visualized these contigs assemblies as an annotated genome displayed as a circular map [63], and the single extended segments from the chloroplast and nuclear genomes were removed. Moreover, the HiFi sequencing data from PacBio were assemble de novo into contigs using SMARTdenovo [64] (https://github.com/ruanjue/smartdenovo) with default settings after a correction process with Canu [65] with the default parameter.

The putative mitochondrial molecule was obtained and assembled by comparing the assembly of short reads and long reads. A major mitogenome of *C. pinnatifida* was assembled by the GetOrganelle v. 1.7.5 pipeline based on the mitochondrial short reads. The GetOrganelle-extended reads were aasembled into a unitig graph using Unicycler (Pacific Biosciences), which was also used to resolve the DBSs in the unitig graph. Finally, the sequences that contained the repeats in DBSs and 2,000 bp sequences upstream and downstream of the repeats were extracted. Subsequently, the flanking regions were switched to form four conformations. To examine the accuracy of the Unicycler, the long-reads were aligned with duplicate fragments in the assembly mitogenome using BWA [66] with default parameters. To examine the mitogenome assembly in more detail, the PacBio long reads and Illumina short reads were mapped to the mitogenome sequences using minimap2 [67] and BWA [66], and samtools [68] was used to calculate the depth of coverage.

The complete mitogenome of *C. pinnatifida* was initially annotated using GESeq [69], and the mitogenomes of *Arabidopsis thaliana* (NC_037304) and tulip tree (*Liriodendron tulipifer*a) (NC_021152.1) were used as the reference genome from GenBank. The PCGs and rRNA genes of the *C. pinnatifida* mitogenome were annotated using BLASTN [41] with parameters. tRNAscan-SE [70] with default parameters was used to predict the tRNA genes. Pseudogenes with clear frame shifts and internal stop codons were annotated manually using Apollo [71].

Analyses of codon usage, sequence repeat and RNA editing site

The RSCU values of *C. pinnatifida* were calculated using MEGA 7.0 [72], and the amino acid compositions of PCGs were analyzed using PhyloSuite [73]. SSRs and tandem repeat sequences were detected using the MISA (https://webblast.ipk-gatersleben.de/misa/) [74], TRE (https://tandem.bu.edu/trf/trf.unix.help.html) [75] and REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer/) [76] programs with the default settings. Non-tandem repeats ≥ 50 bp were identified and curated using a Python script called ROUSFinder [40]. The repeats that contained the overlap regions were manually verified and merged, and visualized using the Circos package [77]. The RNA editing sites in the mitogenome of *C. pinnatifida* were predicted by PREPACT3 online software (http://www.prepact.de/) [78] with a cutoff value = 0.001.

Analyses and validation of the recombination products

The homologous recombination of the non-tandem repeats identified by ROUSfinder was analyzed. BLASTN [41] was used to search for the repeat sequences of *ctg5* and *ctg6*. In addition, the extended sequences on both sides of the repeats that were 2,000 bp long were extracted to predict the presence of possible recombination products around the repeats. Furthermore, the conventional PCR experiments combined with eigut specifc primer pairs were used for amplifying the repeated regions to validate the presence of multiple conformations. The primer sequences for PCR reactions were designed using the Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast) based on the 500 bp upstream and downstream sequences of the repeats combined with 500 bp flanking sequences (Fig. S3), and primer sequences were listed in Table S11. Subsequently, DNA was extracted using using a plant genomic DNA kit (Tiangen, Beijing, China), and the amplifications were performed using Veriti™ Dx 96-Well Fast Thermal Cycler (ThermoFisher). PCR reactions were performed in a 50 μl mixture, comprising 1 μl template DNA, 2 μl each of the forward and reverse primers, 25 μl 2 × Phanta Max Master Mix, and 20 μl ddH2O. After an initial denaturation at 95 °C for 3 min, PCR reactions were conducted for 35 cycles, respectively. Each cycle included denaturation at 95 °C for 15 s, annealing at 55 °C for 15 s, and extension at 72 °C for 15 s, and a final extension step at 72 °C for 5 min. The PCR products were visualized using 1.5% agarose gel electrophoresis (Fig. S4). Subsequently, the single bright bands were excised and sent to Sangon Biotech (Shanghai, China) Co., Ltd. for Sanger sequencing. Sequence alignment was conducted between Sanger and selected sequences using SnapGene. Finally, the PacBio long reads were then mapped to the extracted sequence segments of four conformations, and the repeat-spanning reads were also counted to determine the recombination frequency.

Potential intracellular transfers

To detect the presence of intracellular transfers from the chloroplast genome to the mitogenome, the chloroplast genome of *C. pinnatifida* was also assembled based on the short reads using GetOrganelle v1.7.5 [42]. CPGAVAS2 [79] and CPGView web server [80] were used to annotate the chloroplast genome and check the annotation, respectively. The mitochondrial genome was then mapped to the chloroplast genome using the BLASTN tool [41] with default parameters, and the gene transfer segments of *C. pinnatifida* were visualized by the Circos package [77].

Phylogenetic analysis

The mitochondrial genome data of 25 closely related species from the Rosaceae and six other species from three angiosperm families were downloaded from the GenBank Organelle Genome Resource. The phylogenetic analysis used soybean and sickle senna (*Senna tora*), which are members of the Fabaceae, as an outgroup. The common PCGs were first extracted using PhyloSuite (v1.1.16) [73]. Multiple sequence alignments among these selected PCGs were then performed using MAFFT soft [81]. A maximum likelihood (ML) phylogenetic tree was constructed using IQ-tree [82] with default parameters and 1,000 bootstrap replicates. The best-fit model was chosen as ’GTR + F + R2’ according to the Bayesian Information Criterion (BIC) generated from IQ-tree. The phylogenetic tree was visualized using iTOL software [83].

Collinear analysis with other species in the *Rosaceae*

To explore the mitogenome collinearity relationship between seven Rosaceae species, the nucleotides of the *C. pinnatifida*, pear, *Malus sieversii*, Chinese quince, loquat, apricot and Iturup strawberry were aligned using BLASTN with E-value $\leq 1e^{-10}$, matching rate $\geq$ 80%, and length $\geq$ 500 bp. According to the sequence similarity, the Multiple Synteny Plot of *C. speciosa* with the six Rosaceae species was constructed using MCscanX [84].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05645-w.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Supplementary Material 7.

Supplementary Material 8.

Supplementary Material 9.

Supplementary Material 10.

Supplementary Material 11.

Supplementary Material 12.

Supplementary Material 13.

Supplementary Material 14.

Supplementary Material 15.

Acknowledgements

We’re very grateful to the MitoRun research team for their help with the assembly methods.

Authors’ contributions

LLZ: analyzed the data, prepared figures and tables, wrote the manuscript. SFD:conceived and designed the research, data curation, conceptualization and revised the manuscript. TSL: sampling and carried out the comparative analysis. JG: data curation and revised the manuscript. All authors read and approved the manuscript.

Funding

This research was funded by the Central Finance Forest and Grass Science and Technology Demonstration Project (GTH[2024]2), National Key R&D Program of China (2023YFD1600801) and National Natural Science Foundation of China (32100306).

Availability of data and materials

The assembled mitogenome sequences have been deposited in NCBI GenBank with accession numbers: OR448911 and OR448912, OR448912, and the assembled chloroplast sequences with accession number: OR448914. All data generated by this study are available at the corresponding author upon reasonable request.

Declarations

Ethical approval and consent to participate

We collected fresh leaf materials of *C. pinnatifida* for this study. The study, including plant samples, complies with relevant institutional, national, and international guidelines and legislation. No specifc permits were required for plant collection.

Consent for publication.

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 16 October 2023 Accepted: 27 September 2024
Published online: 07 October 2024

References

1. Moller IM, Rasmusson AG, Van Aken O. Plant mitochondria - past, present and future. Plant J. 2021; 108(4):912–959. https://doi.org/10.1111/tpj.15495.
2. Sun MY, Zhang MY, Chen XN, Liu YY, Liu BB, Li JM, Wang RZ, Zhao KJ, Wu J. Rearrangement and domestication as drivers of Rosaceae mitogenome plasticity. BMC Biol. 2022;20(1):181. https://doi.org/10.1186/s12915-022-01383-3.
3. Wahleithner JA, MacFarlane JL, Wolstenholme DR. A sequence encoding a maturase-related protein in a group II intron of a plant mitochondrial nad1 gene. Proc Natl Acad Sci U S A. 1990;87(2):548–52. https://doi.org/10.1073/pnas.87.2.548.
4. Sychta K, Slomka A, Kuta E. Insights into Plant Programmed Cell Death Induced by Heavy Metals-Discovering a Terra Incognita. Cells-Basel. 2021;10(1):65.
5. Horbay R, Bilyy R. Mitochondrial dynamics during cell cycling. Apoptosis. 2016;21(12):1327–35. https://doi.org/10.1007/s10495-016-1295-5.
6. Xiao SL, Xing JF, Nie TG, Su AG, Zhang RY, Zhao YX, Song W, Zhao JR. Comparative analysis of mitochondrial genomes of maize CMS-S subtypes provides new insights into male sterility stability. BMC Plant Biol. 2022;22(1):469. https://doi.org/10.1186/s12870-022-03849-6.
7. Engqvist MK, Esser C, Maier A, Lercher MJ, Maurino VG. Mitochondrial 2-hydroxyglutarate metabolism. Mitochondrion. 2014; 19 Pt B:275–281. https://doi.org/10.1016/j.mito.2014.02.009.
8. Siqueira SF, Dias SM, Lejeune B, de Souza AP. Marchantia polymorpha mitochondrial orf identifies transcribed sequence in angiosperm mitochondrial genome. Biochim Biophys Acta. 2001;1520(3):203–11.
9. Skippington E, Barkman TJ, Rice DW, Palmer JD. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc Natl Acad Sci U S A. 2015;112(27):E3515-3524. https://doi.org/10.1073/pnas.1504491112.
10. Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 2012; 10(1):e1001241. https://doi.org/10.1371/journal.pbio.1001241.
11. Dong S, Zhao C, Chen F, Liu Y, Zhang S, Wu H, Zhang L, Liu Y. The complete mitochondrial genome of the early flowering plant Nymphaea colorata is highly repetitive with low recombination. BMC Genomics. 2018;19(1):614. https://doi.org/10.1186/s12864-018-4991-4.
12. Wang S, Li D, Yao X, Song Q, Wang Z, Zhang Q, Zhong C, Liu Y, Huang H. Evolution and Diversification of Kiwifruit Mitogenomes through Extensive Whole-Genome Rearrangement and Mosaic Loss of Intergenic Sequences in a Highly Variable Region. Genome Biol Evol. 2019;11(4):1192–206. https://doi.org/10.1093/gbe/evz063.
13. Sloan DB, Müller K, McCauley DE, Taylor DR, Storchová H. Intraspecific variation in mitochondrial genome sequence, structure, and gene content in Silene vulgaris, an angiosperm with pervasive cytoplasmic male sterility. New Phytol. 2012; 196(4):1228–1239. https://doi.org/10.1111/j.1469-8137.2012.04340.x.
14. Yu RX, Sun CY, Zhong Y, Liu Y, Sanchez-Puerta MV, Mower JP, Zhou RC. The minicircular and extremely heteroplasmic mitogenome of the holoparasitic plant. Curr Biol. 2022;32(2):470–9. https://doi.org/10.1016/j.cub.2021.11.053.
15. Møller IM, Rasmusson AG, Van AO. Plant mitochondria - past, present and future. Plant J. 2021;108(4):912–59. https://doi.org/10.1111/tpj.15495.
16. Bi C, Lu N, Xu Y, He C, Lu Z. Characterization and Analysis of the Mitochondrial Genome of Common Bean (Phaseolus vulgaris) by Comparative Genomic Approaches. Int J Mol Sci. 2020;21(11):3778. https://doi.org/10.3390/ijms21113778.
17. Choi IS, Schwarz EN, Ruhlman TA, Khiyami MA, Sabir JSM, Hajarah NH, Sabir MJ, Rabah SO, Jansen RK. Fluctuations in Fabaceae mitochondrial genome size and content are both ancient and recent. BMC Plant Biol. 2019;19(1):448. https://doi.org/10.1186/s12870-019-2064-8.
18. Knoop V. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet. 2004;46(3):123–39. https://doi.org/10.1007/s00294-004-0522-8.
19. Zheng YQ, Liu ZW, Sun YH, Liu GS, Yang AG, Li FX. Characterization of genes specific to sua-CMS in Nicotiana tabacum. Plant Cell Rep. 2018;37(9):1245–55. https://doi.org/10.1007/s00299-018-2309-2.
20. Evans RC, Campbell CS. The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. Am J Bot. 2002;89(9):1478–84. https://doi.org/10.3732/ajb.89.9.1478.
21. Lo EY, Stefanović, S., Christensen, K. I., Dickinson, T. A.. Evidence for genetic association between East Asian and western North American Crataegus L. (Rosaceae) and rapid divergence of the eastern North American lineages based on multiple DNA sequences. Molecular Phylogenetics and Evolution. 2009; 51(2):157–168. https://doi.org/10.1016/j.ympev.2009.01.018.
22. Zhang TC, Qiao Q, Du X, Zhang X, Hou YL, Wei X, Sun C, Zhang RG, Yun QZ, Crabbe MJC et al. Cultivated hawthorn (Crataegus pinnatifida var. major) genome sheds light on the evolution of Maleae (apple tribe). Journal of Integrative Plant Biology. 2022; 64(8):1487–1501. https://doi.org/10.1111/jipb.13318.
23. Wu LW, Cui YX, Wang Q, Xu ZC, Wang Y, Lin YL, Song JY, Yao H. Identification and phylogenetic analysis of five species (Rosaceae) based on complete chloroplast genomes. Planta. 2021;254(1):14. https://doi.org/10.1007/s00425-021-03667-4.
24. Tsutsui K, Suwa A, Sawada K, Kato T, Ohsawa TA, Watano Y. Incongruence among mitochondrial, chloroplast and nuclear gene trees in Pinus subgenus Strobus (Pinaceae). J Plant Res. 2009;122(5):509–21. https://doi.org/10.1007/s10265-009-0246-4.
25. Wu J, Peng W, Qin R, Zhou H. Crataegus pinnatifida: chemical constituents, pharmacology, and potential applications. Molecules. 2014;19(2):1685–712. https://doi.org/10.3390/molecules19021685.
26. Zhang SY, Sun XL, Yang XL, Shi PL, Xu LC, Guo QM. Botany, traditional uses, phytochemistry and pharmacological activity of Crataegus pinnatifida (Chinese hawthorn): a review. J Pharm Pharmacol. 2022;74(11):1507–45. https://doi.org/10.1093/jpp/rgac050.
27. Health NPCoCMo. Pharmacopoeia of the People's Republic of China, 11th ed. Beijing, China: The NMPA Notification; 2020:33.
28. Zhang J, Chai X, Zhao F, Hou G, Meng Q. Food Applications and Potential Health Benefits of Hawthorn. Foods. 2022;11(18):2861. https://doi.org/10.3390/foods11182861.
29. Peng Y, Lou LL, Liu SF, Zhou L, Huang XX, Song SJ. Antioxidant and anti-inflammatory neolignans from the seeds of hawthorn. Bioorg Med Chem Lett. 2016;26(22):5501–6. https://doi.org/10.1016/j.bmcl.2016.10.012.
30. Huang XX, Liu QB, Wu J, Yu LH, Cong Q, Zhang Y, Lou LL, Li LZ, Song SJ. Antioxidant and tyrosinase inhibitory effects of neolignan glycosides from Crataegus pinnatifida seeds. Planta Med. 2014;80(18):1732–8. https://doi.org/10.1055/s-0034-1383253.
31. Zhao P, Guo R, Zhang YY, Zhang H, Yao GD, Lin B, Wang XB, Huang XX, Song SJ. Phenylpropanoid and dibenzofuran derivatives from Crataegus pinnatifida with antiproliferative activities on hepatoma cells. Bioorg Chem. 2019;93: 103354. https://doi.org/10.1016/j.bioorg.2019.103354.
32. Guo R, Lin B, Shang XY, Zhou L, Yao GD, Huang XX, Song SJ. Phenylpropanoids from the fruit of Crataegus pinnatifida exhibit cytotoxicity on hepatic carcinoma cells through apoptosis induction. Fitoterapia. 2018;127:301–7. https://doi.org/10.1016/j.fitote.2018.03.003.
33. Wen L, Lin Y, Lv R, Yan H, Yu J, Zhao H, Wang X, Wang D. An Efficient Method for the Preparative Isolation and Purification of Flavonoids from Leaves of Crataegus pinnatifida by HSCCC and Pre-HPLC. Molecules. 2017;22(5):767. https://doi.org/10.3390/molecules22050767.
34. Wen LR, Guo RX, You LJ, Abbasi AM, Li T, Fu X, Liu RH. Major triterpenoids in Chinese hawthorn "Crataegus pinnatifida" and their effects on cell proliferation and apoptosis induction in MDA-MB-231 cancer cells. Food Chem Toxicol. 2017;100:149–60. https://doi.org/10.1016/j.fct.2016.12.032.
35. Kim E, Jang E, Lee JH. Potential Roles and Key Mechanisms of Hawthorn Extract against Various Liver Diseases. Nutrients. 2022;14(4):867. https://doi.org/10.3390/nu14040867.
36. Guo X, Zhang YT, Lu CH, Qu FX, Jiang XY. Protective effect of hyperoside on heart failure rats via attenuating myocardial apoptosis and inducing

autophagy. Biosci Biotech Bioch. 2020;84(4):714–24. https://doi.org/10.1080/09168451.2019.1685369.

37. Xu Y, Deng T, Xie LJ, Qin T, Sun T. Neuroprotective effects of hawthorn leaf flavonoids in Aβ(25–35)-induced Alzheimer's disease model. Phytother Res. 2023;37(4):1346–65.
38. Tadic VM, Dobric S, Markovic GM, Dordevic SM, Arsic IA, Menkovic NR, Stevic T. Anti-inflammatory, gastroprotective, free-radical-scavenging, and antimicrobial activities of hawthorn berries ethanol extract. J Agric Food Chem. 2008;56(17):7700–9. https://doi.org/10.1021/jf801668c.
39. Shang XY, Guo R, Yu XQ, Lin B, Huang XX, Yao GD, Song SJ. Enantiomeric 8-O-4'-type neolignans from Crataegus pinnatifida exhibit cytotoxic effect via apoptosis and autophagy in Hep3B cells. Bioorg Chem. 2020;104: 104267. https://doi.org/10.1016/j.bioorg.2020.104267.
40. Wynn EL, Christensen AC. Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. G3-Genes Genom Genet. 2019; 9(2):549–559. https://doi.org/10.1534/g3.118.200948.
41. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015;43(16):7762–8. https://doi.org/10.1093/nar/gkv784.
42. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21(1):241. https://doi.org/10.1186/s13059-020-02154-5.
43. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Plos Comput Biol. 2017, 13(6). e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
44. Yang J, Liu N, Zheng XL, Wu JC, Lin XJ, Zheng GH. Phylogenetic relationships and characterization of the complete mitochondrial genome of Eriobotrya japonica in southeast of China. Mitochondrial DNA B Resour. 2019;18(5):269–70. https://doi.org/10.1080/23802359.2019.1692709.
45. Guo WH, Grewe F, Fan WS, Young GJ, Knoop V, Palmer JD, Mower JP. Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol Biol Evol. 2016;33(6):1448–60. https://doi.org/10.1093/molbev/msw024.
46. Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang CY, et al. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics. 2007;177(2):1173–92. https://doi.org/10.1534/genetics.107.073312.
47. Gui ST, Wu ZH, Zhang HY, Zheng YZ, Zhu ZX, Liang DQ, Ding Y. The mitochondrial genome map of Nelumbo nucifera reveals ancient evolutionary features. Sci Rep-Uk. 2016;6:30158. https://doi.org/10.1038/srep30158.
48. Zhang CY, Ma H, Sanchez-Puerta MV, Li L, Xiao JH, Liu ZF, Ci XQ, Li J. Horizontal Gene Transfer has Impacted cox1 Gene Evolution in Cassytha filiformis. J Mol Evol. 2020;88(4):361–71. https://doi.org/10.1007/s00239-020-09937-1.
49. Sloan DB. One ring to rule them all? Genome sequencing provides new insights into the "master circle" model of plant mitochondrial DNA structure. New Phytol. 2013;200(4):978–85. https://doi.org/10.1111/nph.12395.
50. Gualberto JM, Mileshina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A. The plant mitochondrial genome: Dynamics and maintenance. Biochimie. 2014;100:107–20. https://doi.org/10.1016/j.biochi.2013.09.016.
51. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. P Natl Acad Sci USA. 2015;112(33):10177–84. https://doi.org/10.1073/pnas.1422049112.
52. Li JL, Xu YC, Shan YY, Pei XY, Yong SY, Liu C, Yu J. Assembly of the complete mitochondrial genome of an endemic plant, Scutellaria tsinyunensis, revealed the existence of two conformations generated by a repeat-mediated recombination. Planta. 2021;254(2):36. https://doi.org/10.1007/s00425-021-03684-3.
53. Cao P, Huang Y, Zong M, Xu ZL. De Novo Assembly and Comparative Analysis of the Complete Mitochondrial Genome of Chaenomeles speciosa (Sweet) Nakai Revealed the Existence of Two Structural Isomers. Genes-Basel. 2023;14(2):526. https://doi.org/10.1007/s00425-021-03684-3.
54. Fang B, Li JL, Zhao Q, Liang YP, Yu J. Assembly of the Complete Mitochondrial Genome of Chinese Plum (Prunus salicina): Characterization of Genome Recombination and RNA Editing Sites. Genes-Basel. 2021;12(12):1970. https://doi.org/10.3390/genes12121970.
55. Wang Y, Chen SJ, Chen JJ, Chen CJ, Lin XJ, Peng H, Zhao Q, Wang XY. Characterization and phylogenetic analysis of the complete mitochondrial genome sequence of Aythya marila. Sci Rep-Uk. 2023;14(6):1205. https://doi.org/10.3390/genes14061205.
56. Lloyd AH, Timmis JN. The Origin and Characterization of New Nuclear Genes Originating from a Cytoplasmic Organellar Genome. Mol Biol Evol. 2011;28(7):2019–28. https://doi.org/10.1093/molbev/msr021.
57. Qu YS, Zhou PY, Tong CF, Bi CW, Xu LA. Assembly and analysis of the Populus deltoides mitochondrial genome: the first report of a multicircular mitochondrial conformation for the genus Populus. Journal Forestry Res. 2023; 34(3):717–733. https://doi.org/10.1007/s11676-022-01511-3.
58. Leister D. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 2005;21(12):655–63. https://doi.org/10.1016/j.tig.2005.09.004.
59. Hong Z, Liao XZ, Ye YJ, Zhang NN, Yang ZJ, Zhu WD, Gao W, Sharbrough J, Tembrock LR, Xu DP, et al. A complete mitochondrial genome for fragrant Chinese rosewood (Dalbergia odorifera, Fabaceae) with comparative analyses of genome structure and intergenomic sequence transfers. BMC Genomics. 2021;22(1):672. https://doi.org/10.1186/s12864-021-07967-7.
60. Tanaka Y, Tsuda M, Yasumoto K, Terachi T, Yamagishi H. The complete mitochondrial genome sequence of Brassica oleracea and analysis of coexisting mitotypes. Curr Genet. 2014;60(4):277–84. https://doi.org/10.1007/s00294-014-0433-2.
61. Li J, Wang S, Yu J, Wang L, Zhou S. A Modified CTAB Protocol for Plant DNA Extraction. Chinese Bulletin of Botany. 2013;48(1):7.
62. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
63. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015; 31(20):3350–3352. https://doi.org/10.1093/bioinformatics/btv383.
64. Liu H, Wu S, Li A, Ruan J. SMARTdenovo: a de novo assembler using long noisy reads. GigaByte. 2021; 2021:gigabyte15. https://doi.org/10.46471/gigabyte.15.
65. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36. https://doi.org/10.1101/gr.215087.116.
66. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Proc GPD: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
69. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–11. https://doi.org/10.1093/nar/gkx391.
70. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64. https://doi.org/10.1093/nar/25.5.955.
71. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA et al: Apollo: a sequence annotation editor. Genome Biol. 2002; 3(12): research0082.1–0082.14. https://doi.org/10.1186/gb-2002-3-12-research0082.
72. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33(7):1870–1874. https://doi.org/10.1093/molbev/msw054.
73. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX, Wang GT. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20(1):348–55. https://doi.org/10.1111/1755-0998.13096.
74. Beier S, Thiel T, Munch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017; 33(16):2583–2585. https://doi.org/10.1093/bioinformatics/btx198.
75. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.

76. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633–42. https://doi.org/10.1093/nar/29.22.4633.
77. Zhang H, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC Bioinformatics. 2013;14:244. https://doi.org/10.1186/1471-2105-14-244.
78. Lenz H, Hein A, Knoop V. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0. BMC Bioinformatics. 2018; 19(1):255. https://doi.org/10.1186/s12859-018-2244-9.
79. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019; 47(W1):W65-W73. https://doi.org/10.1093/nar/gkz345.
80. Liu S, Ni Y, Li J, Zhang X, Yang H, Chen H, Liu C. CPGView: A package for visualizing detailed chloroplast genome structures. Mol Ecol Resour. 2023; 23(3):694–704. https://doi.org/10.1111/1755-0998.13729.
81. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–780. https://doi.org/10.1093/molbev/mst010.
82. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–274. https://doi.org/10.1093/molbev/msu300.
83. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021; 49(W1):W293-W296. https://doi.org/10.1093/nar/gkab301.
84. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7): e49. https://doi.org/10.1093/nar/gkr1293.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.