Cytoplasmic genomes of *Jasminum sambac* reveal divergent sub-mitogenomic conformations and a large nuclear chloroplast-derived insertion

Jingping Fang[1,2*†], Aiting Lin[4†], Hansong Yan[4†], Liqing Feng[1,3], Shaoqing Lin[1], Patrick Mason[2], Linwei Zhou[1], Xiuming Xu[4], Kai Zhao[1], Yongji Huang[3*] and Robert J. Henry[2*]

Abstract

Background *Jasminum sambac*, a widely recognized ornamental plant prized for its aromatic blossoms, exhibits three flora phenotypes: single-petal ("SP"), double-petal ("DP"), and multi-petal ("MP"). The lack of detailed characterization and comparison of *J. sambac* mitochondrial genomes (mitogenomes) hinders the exploration of the genetic and structural diversity underlying the varying floral phenotypes in jasmine accessions.

Results Here, we *de novo* assembled three mitogenomes of typical phenotypes of *J. sambac*, "SP", "DP", and "MP-hutou" ("HT"), with PacBio reads and the "HT" chloroplast (cp) genome with Illumina reads, and verified them with read mapping and fluorescence in situ hybridization (FISH). The three mitogenomes present divergent sub-genomic conformations, with two, two, and four autonomous circular chromosomes ranging in size from 35.7 kb to 405.3 kb. Each mitogenome contained 58 unique genes. Ribosome binding sites with conserved AAGAAx/AxAAAG motifs were detected upstream of uncanonical start codons TTG, CTG and GTG. The three mitogenomes were similar in genomic content but divergent in structure. The structural variations were mainly attributed to recombination mediated by a large (~5 kb) forward repeat pair and several short repeats. The three jasmine cp. genomes showed a well-conserved structure, apart from a 19.9 kb inversion in "HT". We identified a 14.3 kb "HT"-specific insertion on Chr7 of the "HT" nuclear genome, consisting of two 7 kb chloroplast-derived fragments with two intact *ndhH* and *rps15* genes, further validated by polymerase chain reaction (PCR). The well-resolved phylogeny suggests faster mitogenome evolution in *J. sambac* compared to other Oleaceae species and outlines the mitogenome evolutionary trajectories within Lamiales. All evidence supports that "DP" and "HT" evolved from "SP", with "HT" being the most recent derivative of "DP".

---

†Jingping Fang, Aiting Lin and Hansong Yan contributed equally to this work.

*Correspondence:
Jingping Fang
jinphia@fjnu.edu.cn
Yongji Huang
scolandry@foxmail.com
Robert J. Henry
robert.henry@uq.edu.au

Full list of author information is available at the end of the article

Conclusion The comprehensive characterization of jasmine organelle genomes has added to our knowledge of the structural diversity and evolutionary trajectories behind varying jasmine traits, paving the way for in-depth exploration of mechanisms and targeted genetic research.

Keywords *Jasminum sambac*, Mitochondrial genome, Repeat-mediated recombination, Subgenomic configurations, Chloroplast-derived fragment, Phylogenetic tree

Background

Jasmine (*Jasminum sambac* (L.) Aiton) is a perennial evergreen erect-shrub with the white sweet-scented flowers used for multiple purposes such as an ornamental, in scented tea, as an essential oil and food flavor [1]. Jasmine plants, known under their common name as "Molihua", have long been cultivated in China (over 2000 years) due to their ornamental and economic value. *J. sambac* is a dicotyledonous species belonging to the Oleaceae family under the order Laminales. Jasmine varieties present varying morphological features that can be roughly classified into three main types based on the petal phenotype: single-petal ("SP", cv. *Unifoliatum*), double-petal ("DP", cv. *Bifoliatum*), and multi-petal ("MP", cv. *Trifoliatum*), which can be differentiated by the number of flowers, floral fragrance and stress-resistance. In commercial production systems, Jasmine is propagated asexually leading the accumulation of deleterious mutations in jasmine genomes, rendering them vulnerable to both biotic and abiotic stress. The "DP" or "MP" jasmine typically has a diploid genome (2n=2x=26) whereas the "SP" has a chimeric composition of diploid and triploid cells in various tissues, which probably leads to the poorer resistance and fertility in "SP" varieties [2]. Specifically, "SP" jasmine flowers tend to be highly aromatic, whilst the "DP" jasmines have more flowers and stronger resistance to biotic and abiotic stressors, making them the leading commercially cultivated type. Unlike "SP" and "DP", "MP" jasmine is not suitable for commercial tea production due to the longer duration of flowering, reduced flower number, increased incidence of flower deformity, weaker floral aroma and disease resistance. Their larger staggered multi-petal phenotype, however, is considered one of the most highly valuable commercial floral attributes in floriculture. The "MP" jasmine cultivar "Hutoumoli" (hereinafter referred to as "HT") is one of the most popular ornamental flowers in China for its myriad of aesthetic uses. The flowers of "HT" harbor large white staggered petals with mellow elegant aroma, and are widely used in bonsai gardening. The "HT" genotype has long been recognized as unstable or mutable, yet a clear genetic basis for this has not been demonstrated. The evolution of "HT" remains enigmatic, with uncertainty surrounding whether it originated from "DP" or "SP."

Mitochondria are known as the "powerhouse" of eukaryotic cells where aerobic respiration takes place. Around 1.4 billion years ago, mitochondria came into existence through a process called endosymbiosis. During this process, primitive cells engulfed single-celled α-proteobacteria ancestors and assimilated them into the host cell [3]. The endosymbiont "domestication" within the host cell experienced a drastic elimination of genome size and coding capacity as a consequence of gene loss or extensive migration of genes from the endosymbiont to the nucleus [4]. Only a small fraction (0.5–1.2%) of the original gene content have been preserved within modern mitochondrial genomes (hereafter, mitogenomes or mtDNA) [5]. It is well known that intracellular and horizontal gene transfers (IGTs and HGTs) have occurred during evolution and are still ongoing today [6, 7]. Such exogenous sequences help shape the plant mitogenomes that we have seen today. In addition, multiple pieces of evidence suggest that plant mitogenomes are evolving rapidly in structure but slowly in sequence, especially at a slower tempo in gene sequences compared to their chloroplast or nuclear counterparts [8, 9]. The mitogenome is typically inherited maternally [10]. The slow-paced evolution and uniparental inheritance of plant mtDNA provide an attractive reservoir of phylogenetic information to trace evolutionary events. Many factors, such as widespread incomplete lineage sorting, molecular convergence, heterogeneity of evolutionary rates, and reticulate evolution (hybridization and HGT) can lead to multiple origins of mitochondria and incongruent phylogenetic signals from cytoplasmic and nuclear genomic compartments [11, 12].

Plant mitogenomes come in every shape and size, exhibiting high variability in terms of size (ranging from ~66 kb in *Viscum scurruloideum* to ~12 Mb in *Larix sibirica*), structural organization, gene order and content, as well as repeat structure [13]. Unlike compact single-circular mitogenomes in animals, land plants possess a versatile range of genomic configurations: they are elastic in length and evolve rapidly in structure. Many lines of evidence from sequencing data and electron microscopy show that plant mtDNAs display striking structural diversity with architectures shifting among linear, circular, branched, multi-chromosomal and complex molecules of different sizes [13–16]. The massive proliferation of noncoding content, the richness in repetitive sequences, together with the active integration of IGTs and HGTs [7] appear to contribute considerably to mitochondrial genome diversity [15, 17, 18]. The various-sized repeats can participate in extensive homologous recombination

(HR), which is thought to be largely responsible for structural diversification and multiple configurations of the plant mitogenomes [14]. Repeat-mediated HR could actively transform one single master ring into an equimolar collection of interchangeable subgenomic rings, and maintain the mitogenome in a dynamic equilibrium [14]. The presence of multiple-ring mitogenomes induced by repeats have been reported in many plant species, such as *Silene*, cottonwood [19], tea tree [20], soybean [21], and cucumber [22]. Recombination could be driven via large and small repeated sequences. Large repeats (>500 bp) typically promote frequent and reversible reciprocal HR events that do not interfere with gene structures, preserving the mitogenome in a highly dynamic entity [18, 23]. But in certain cases, the reshuffling of mitogenomes by repeat-mediated recombination through HR can create expressed chimeric cytoplasmic male sterility (CMS)-relevant genes [18]. Small repeats (<500 bp) could also mediate infrequent, irreversible and asymmetric recombination events, resulting in a low number of recombinant molecules, the so-called 'sublimons' [24].

In spite of the economic and ornamental importance of jasmine flowers, advances in genetics and genomics have lagged behind other horticultural flowers. Not until 2021 was the first *J. sambac* "MP" genome published using a combination of Illumina, Nanopore and HiC sequencing data [25]. Thereafter, additional *J. sambac* genomes have been assembled for the varieties "SP" [2, 26] and "DP" [2, 27–29] using PacBio long-reads. These jasmine genome resources provide valuable molecular data for functional genomics and genetic breeding research. The complexity of repeated sequences, nuclear-mitochondrial (NUMT) sequences, mitochondrial plastid transferred fragments (MTPTs), and the complex mixture of physical forms including linear, branched linear and multiple-ring conformations [7, 14] pose a challenge to the complete assembly of plant mitogenomes. The Oleaceae family compasses ca. 900 species in 28 genera. However, only 14 complete mitogenomes of 8 species compared to 344 chloroplast (cp.) genomes of 147 species in Oleaceae were available on GenBank as of August, 2023. Mitochondrial-plastid phylogenomic incongruence often arises from their distinctly different evolutionary tempos and patterns, despite their coexistence in the same cell [30]. Unravelling more mitogenomes in the Oleaceae family will facilitate better understanding of evolutionary history within Oleaceae and across Lamiales lineages from a mitochondrial perspective, particularly where some nodes are as yet not fully resolved.

Here, we *de novo* assembled the mitogenomes of three *J. sambac* accessions "DP", "SP", and "HT" into one of the physical forms — multi-circular conformations by mining mtDNA reads from PacBio libraries generated as part of the effort to sequence entire jasmine genomes. The

assembly of a single-circular cp. genome of "HT" was also completed. The accurate assembly of three multi-circular jasmine mitogenomes were confirmed by bioinformatics analysis and fluorescence in situ hybridization (FISH) experiments. Recombination behavior among three jasmine mitogenomes was assessed based on the repeated sequence analysis to illustrate the structural complexity and dynamic conformation. The availability of complete jasmine mitogenomes enables us to explore the molecular mechanisms underlying structural mutability in the jasmine mitogenome and gain a more holistic view of gene content, genome architecture and arrangement, and selection pressure among Oleaceae species. We also resolved a high-resolution of phylogenetic relationship among Lamiales species based on single-copy genes of mtDNA to expand our understanding of the evolutionary history within Lamiales. This study will contribute to our understanding of the genetic diversity of organelle genomes behind varying phenotypes in jasmine accessions.

Results and discussion

The atypical multi-circular structure of jasmine mitogenomes

Approximately 21.67 Gb of PacBio HiFi data composed of 1,273,938 reads were obtained with an estimated coverage of 44x for the "HT" genome. A total of 51.80 Gb clean Illumina PE reads representing 106x of "HT" genome equivalents were generated after filtering. Homologous sequence search extracted 110,503 (1.27 Gb) "SP", 65,066 (889.06 Mb) "DP" and 126,207 (1.27 Gb) "HT" potential mitochondrial homo-reads from the PacBio read pool for subsequent *de novo* assembly of mitogenomes. The search for candidate mitogenome contigs resulted in 7 Flye contigs for "HT", 2 Canu contigs for "DP", and 5 Canu contigs for "SP" final assemblies. The backbone contigs used for the three mitogenome assemblies and the genome coverage of each mitocontig are listed in Table S1.

The "HT" mitogenome was finally assembled into four circular chromosomes with no gaps with sizes of 280,079 bp, 103,659 bp, 89,506 bp, and 35,701 bp, totaling 508,945 bp and with a GC content of 44.98% (Fig. 1; Table 1, and Table S2). In contrast, the "DP" mitogenome was assembled into two independent, gap-free circular mtDNA molecules (405,270 bp and 103,659 bp, totaling 508,929 bp), and the "SP" mitogenome into two circular molecules (404,522 bp and 130,744 bp, totaling 535,266 bp), with overall GC contents of 44.98% and 44.50%, respectively. Alignment of PacBio long reads revealed no breakpoint across the mitogenomes (Fig. 2a-c). The average depth was in the range of 1291~1587x for "HT" mtDNAs molecules, 495~611x for "DP", and 1001-1009x for "SP", respectively (Fig. 2c and S1).
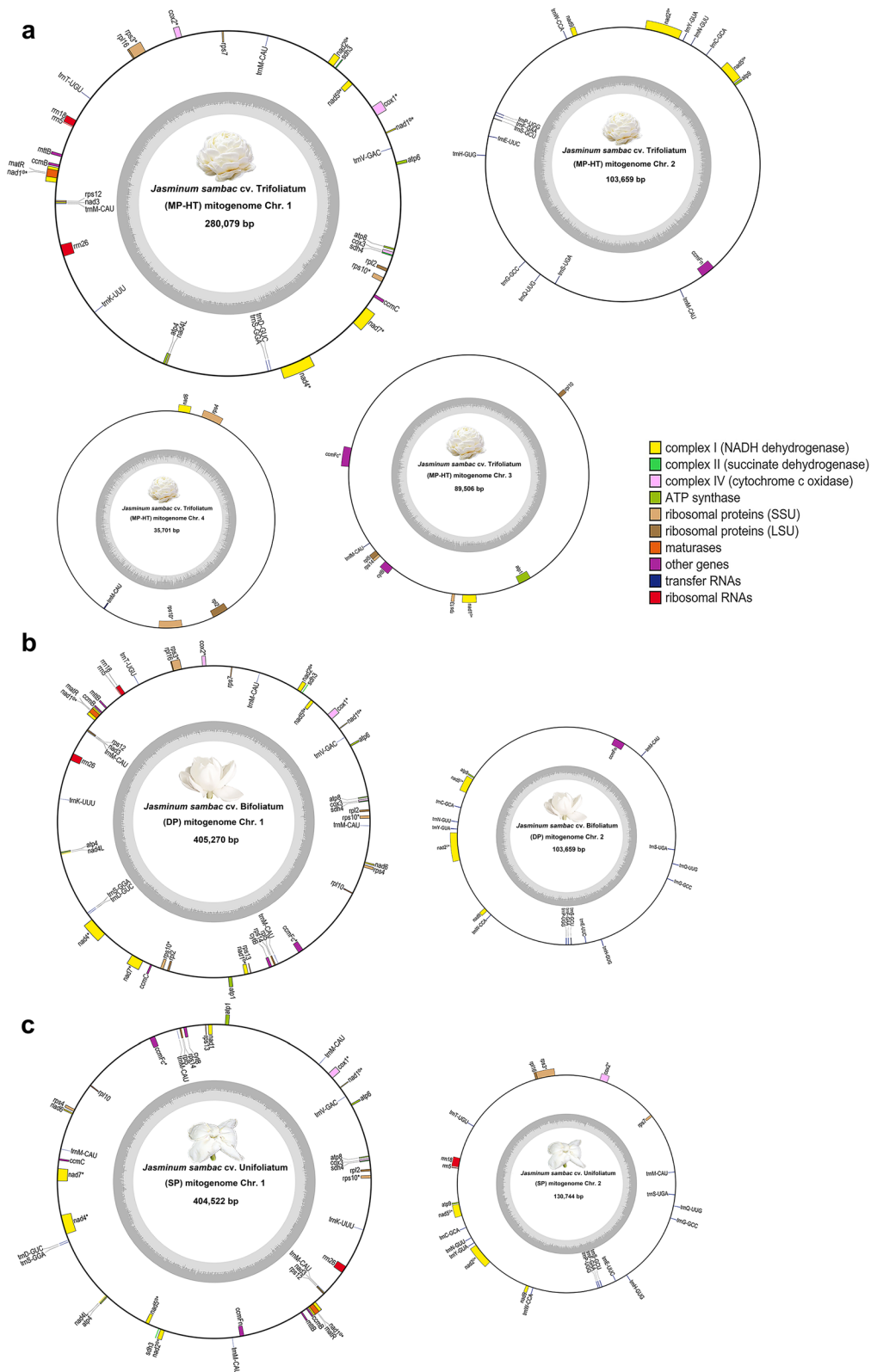
Fig. 1 Circular maps of the multi-circular mitogenomes of **a.** *Jasminum sambac* cv. Trifoliatum; **b.***J. sambac* cv. Bifoliatum; **c.***J. sambac* cv. Unifoliatum. Genes are colored according to different functional groups. Genes facing outside and inside of the circle are transcribed in the clockwise and counter-clockwise directions, respectively. The dark-grey inner circle represents the GC content of each chromosome. Intron-containing genes are indicated with asterisks (*). "ѳ" indicates trans-spliced introns

Table 1 The statistics of general mitogenomic features of three representative jasmine plants

Feature	MP (Multi-petal)	DP (Double-petal)	SP (Single-petal)
Genome size (bp)	508,945	508,929	535,266
GC content (%)	44.98	44.98	45.00
No. of PCGs	39 (37 unique)	39 (37 unique)	37 (37 unique)
No. of exons	62	62	59
Combined length of PCGs (bp)	32,946 (6.47%)	32,946 (6.47%)	31,674 (5.92%)
Average length of PCGs (bp)	845	845	856
GC content of PCGs (%)	42.78	42.78	42.64
Intergenic region (bp)	475,999 (93.53%)	475,983 (93.53%)	503,592 (94.08%)
GC content of intergenic region (%)	45.13	45.13	45.15
No. of *Cis*-spliced introns	12	12	11
Cis-spliced genes	ccmFc, cox1, cox2, nad4, nad7, rps10, rps3		
Cis-spliced introns (bp)	16,601	16,601	15,793
No. of *Trans*-spliced *introns*	7	7	7
Trans-spliced genes (No. of introns)	nad1 (2), nad2 (3), nad5 (2)		
Trans-spliced introns (bp)	11,151	11,151	11,151
No. of tRNA genes	22 (18 unique)	22 (18 unique)	23 (18 unique)
Combined length of tRNAs (bp)	1,674 (0.33%)	1,674 (0.33%)	1,747 (0.33%)
Average length of tRNAs (bp)	76.09	76.09	75.96
No. of rRNA genes	3 (3 unique)	3 (3 unique)	3 (3 unique)
Combined length of rRNAs (bp)	4,963 (0.98%)	4,963 (0.98%)	4,963 (0.93%)
Average length of rRNAs (bp)	1654.33	1654.33	1654.33

FISH-mapping revealed that all four mitochondria-derived probes from each "HT" mitochondrial chromosome mainly produced signals in the cytoplasm of "HT" cells, while 5 S rDNA probes as a control produced two clear signals on interphase nucleus chromatids (Fig. 2d). The cytoplasmic distribution of the "HT" mitogenome confirms the accuracy of this assembly. In "DP" and "SP" cells, similar distribution patterns of mtDNAs were observed, with 5 S rDNA probes generating two and three clear signals on the chromatids, respectively (Fig. S2-S3). These findings ascertain the cytoplasmic distribution of three newly-assembled mitogenomes, and also confirm the triploidy in "SP" cells. These together validate the accurate assembly of three gap-free multi-circular jasmine mitogenomes.

Mitogenomic features and annotation

The three jasmine mitogenomes shared most gene features, each containing 58 unique genes that were scattered singly or in nests, including 37 protein-coding genes (PCGs), 18 tRNA genes and 3 rRNA genes (*rrn5*, *rrn18* and *rrn26*) (Fig. 1; Table 1, and Table S3). All PCGs were single-copy genes in "SP", while two ribosomal protein genes (*rpl2* and *rps10*) were found to be doubled in "HT" and "DP". The tRNA gene *trn(f)M-CAU* was quintupled in "HT" and "DP", and sextupled in "SP", and they were randomly distributed across the mitogenomes. The *cox3* overlapped with the upstream *sdh4* gene by 72 bp, and the *rpl16* showed an overlap of 22 bp with the upstream *rps3* gene. In all three mitogenomes, each circular chromosome had protein-coding capability, but the majority of genes were located on chromosome 1 (mtChr1). Three genes (*nad1*, *nad2* and *nad5*) containing *trans*-spliced introns were found. *Nad2* and *nad5* were split across two distinct chromosomes, while "HT" had an additional *nad1* separated by mtChr1 and mtChr3. In total, the three mitogenomes contained 7 *trans*-spliced introns with combined length of 11,151 bp. In "HT" and "DP", 8 PCGs harbored 12 *cis*-spliced introns totaling 16,601 bp, while in "SP", 7 PCGs (except one *rps10* copy) comprised 11 *cis*-spliced introns with 15,793 bp in length. Three *cis*-spliced introns were found in both *nad4* and *nad7*, while the remaining genes each contained one.

The combined length of coding regions was identical in "HT" and "DP" mitogenomes (39,538 bp), comprising 32,946 bp (6.47%) PCGs, 1,674 bp (0.33%) tRNA genes and 4,963 bp (0.98%) rRNA genes (Table 1), implying the closer evolutionary relationship between "HT" and "DP". In contrast, non-coding regions represented 93.53% of the mitogenomes. The total length of coding regions shrunk slightly to 38,384 bp in "SP", containing 31,674 (5.92%) PCGs, 1,747 (0.33%) tRNAs and 4,963 (0.93%) rRNAs, whereas intergenic spaces constituted the remaining 94.08% of this mitogenome. More specifically, PCGs had an average length of 845 bp for "HT" and "DP", and 856 bp for "SP", with an overall GC content of 42.64–42.78%. The tRNAs and rRNAs had an average length of 76 bp and 1,654 bp, respectively, in the three mitogenomes.
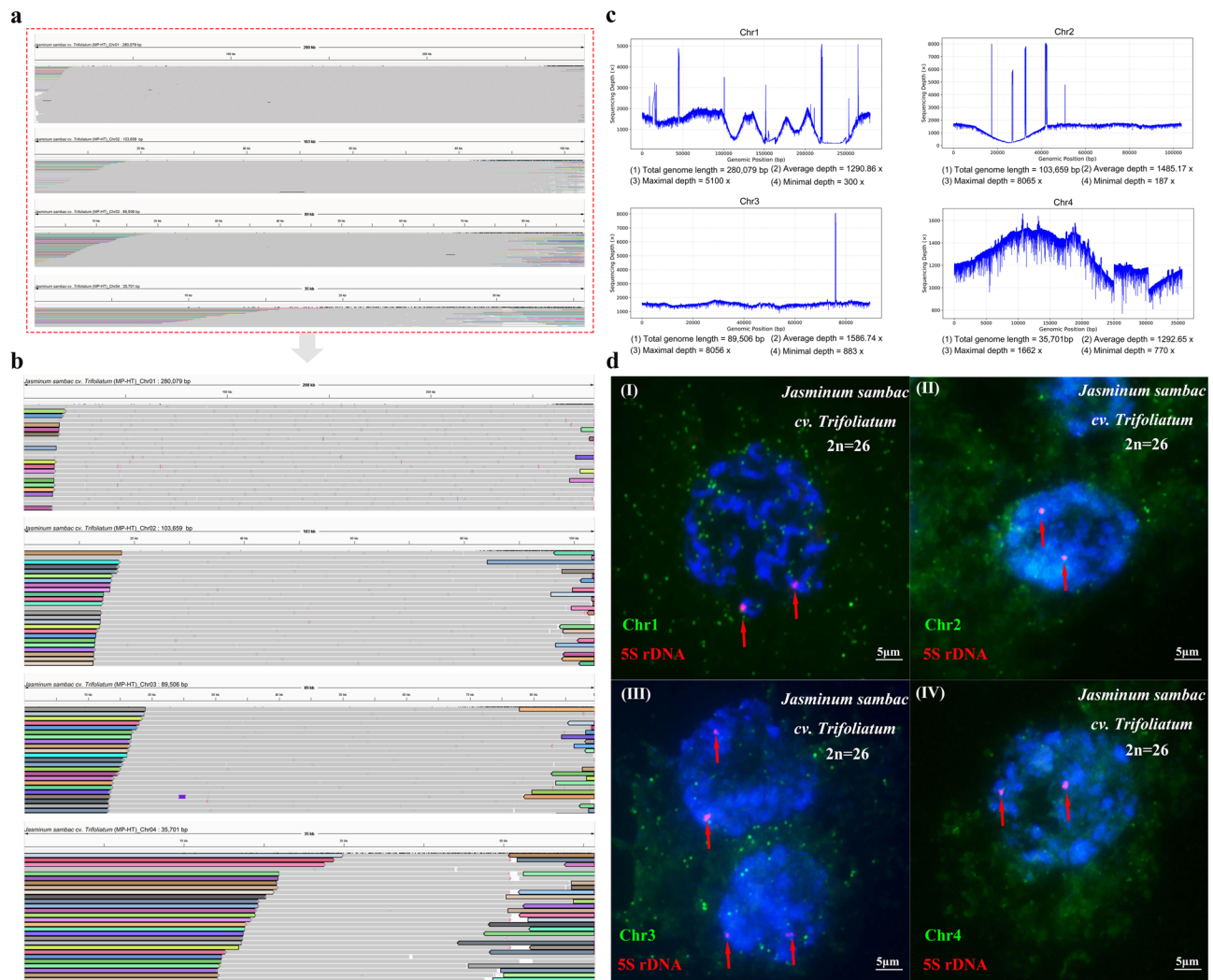
Fig. 2 Verification of the mitogenome assembly and the distribution of mitochondrial DNAs (mtDNAs) in *Jaminum sambac* cv. *Trifoliatum* "HT". (**a-b**) The accuracy and circularity of the assembly was manually checked in IGV. The alignment of PacBio reads revealed no breakpoints across the chromosomes. Single reads aligning to both ends simultaneously were color-coded identically, confirming the circularity of the four chromosomes. (**c**) The mapping depth of PacBio reads along the "HT" mitogenome chromosomes where every position was covered at least 187x. The average depth was in the range of 1290.86-1586.74x. (**d**) FISH mapping of mtDNAs in *J. sambac* "HT" cells. FISH mapping of probes from mtChr1 (**I**), mtChr2 (**II**), mtChr3 (**III**) and mtChr4 (**IV**) in the cells of *J. sambac* cv. *Trifoliatum*. Signals of the mtDNAs (in green color) were mainly detected in cytoplasm based on all of the four probes, while 5 S rDNA (in red) probes as a control were detected on chromatids of interphase nucleus. The red arrows indicate the signals of 5 S rDNA

A total of 37 unique PCGs included nine, two, one, three, five and four genes (24 genes) encoding mitochondrial respiratory chain complexes I, II, III, IV, V, and cytochrome c biogenesis, respectively; Four and seven genes encoding large and small subunit ribosomal proteins, respectively (Table S3). All 37 unique PCGs were functionally annotated to the Nr (37), GO (22), COG (27), KEGG (24) and Swiss-Prot (37) databases, respectively (Table S4-S7).

Codon usage of PCGs and uncanonical initiator codons

The potential codon usage and codon-anticodon recognition pattern on 37 unique PCGs of *J. sambac* mitogenomes were estimated (Table S8, Fig. S4). The jasmine mitogenome contained 61 nucleotide triplets (codons) for 20 different amino acids. Met and Trp were encoded by single codon CAU and CCA, respectively. Other amino acids were encoded by 2 to 6 codons. In total, 10,927 (for "HT" and "DP") and 10,521 (for "SP") codons encoded by 37 unique PCGs were detected in the three mitogenomes. AT-rich codons dominated (60.17%), with A/T most frequently occurring at the third position (62.23%), followed by the second (57.14%) and first positions (52.45%). The relative synonymous codon usage (RSCU) analysis showed that Leu, Ser and Arg were the most frequently used amino acids, whereas Trp and Met were the least common (Fig. S4). The GCT-**AGC** codon (Ala) with the highest RSCU value was the highly
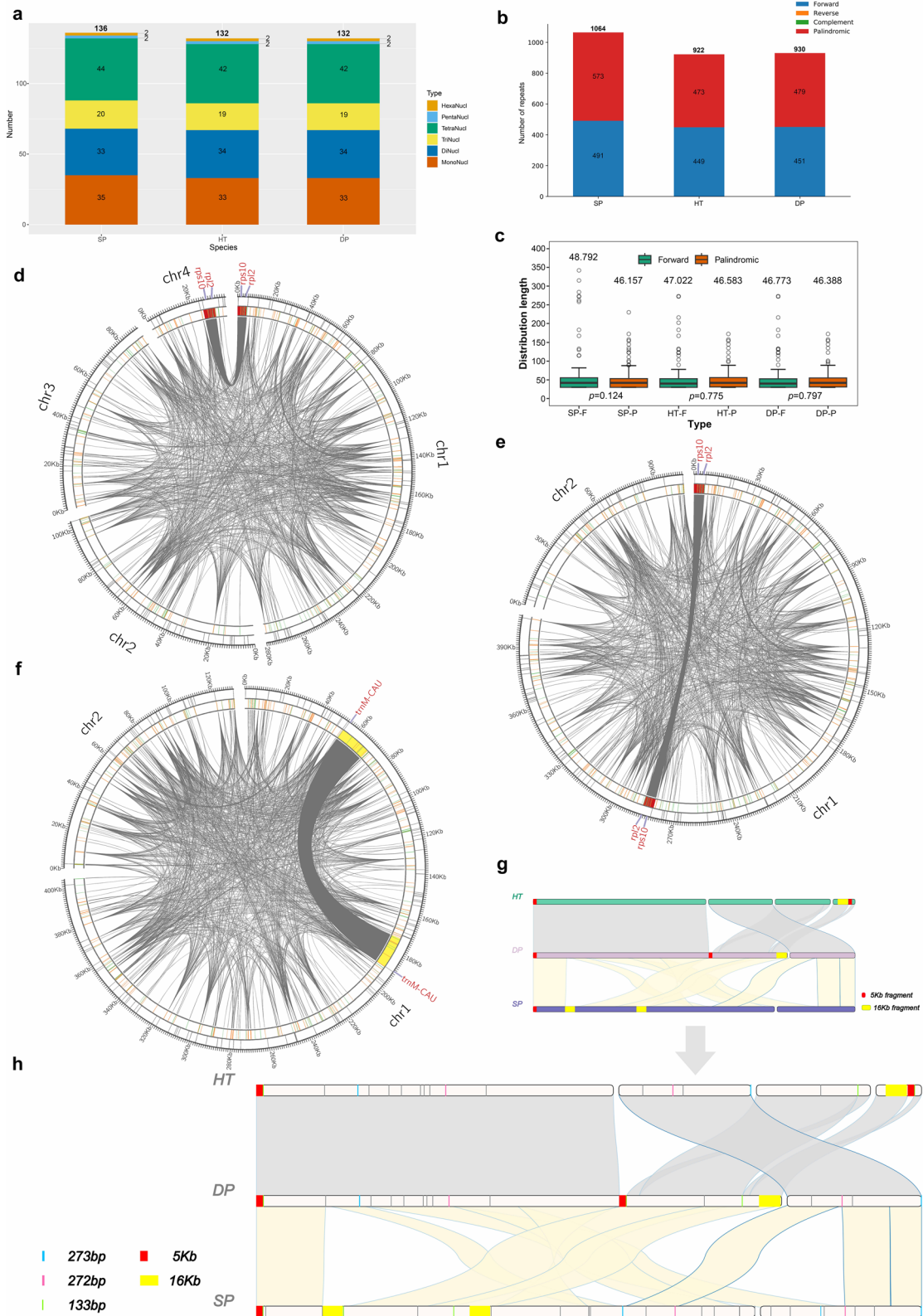
Fang *et al. BMC Plant Biology*   (2024) 24:861

Page 7 of 23

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Analysis of microsatellites and long repeated sequences in jasmine mitogenomes. **(a)** Distribution of microsatellites by the motif length, and **(b)** Number and type of dispersed repeat-pairs in three different mitogenomes "HT", "DP" and "SP". No reverse and complement repeats were found. **(c)** Box plot illustrating the length distribution of forward and palindromic repeats in "SP", "HT", "DP". The circles represent potential outliers, where the longest values, including 5,352 bp, 5,421 bp, and 16,267 bp, were removed. The mean values and statistical significance are shown. No significant differences in the length distributions among "SP", "HT", and "DP". **(d-f)** Schematic representation of microsatellites and dispersed repeats in the mitogenomes of "HT" (d), "DP" (e), and "SP" (f). The outer circle shows the distribution of microsatellites with grey short bars. The inner circle shows the distribution of forward and palindromic long repeats with green and orange short bars, and the syntenic repeat pairs were connected with grey lines. The long ~ 5 kb repeat pair and ~ 16 kb repeat pair were highlighted in red and yellow respectively. The interval scale was 1 kb. **(g)** Syntenic comparisons showing the highly variable mitogenome structure among "HT", "DP" and "SP". **(h)** The collinear repeated pairs present in the rearrangement breakpoints among three mitogenomes are displayed. The long (> 1 kb) and short (> 30 bp) repeat-pairs might be involved in the formation of subgenomic conformations and structural variations through homologous recombination (HR). Three pairs with intermediate lengths (133 bp, 272 bp, and 273 bp) are shown in green, pink and blue color, respectively, while the rest pairs (< 100 bp) are plotted in grey

preferred codon, followed by TAT-**AUA** (Tyr) and CAT-**AUG** (His). The RSCU results also indicated that codons with A/T at the third position were more frequently used than those with G/C at the third position (Tables S8).

Nearly all PCGs in jasmine mitogenomes used the canonical ATG start codon, except for four genes, *rps4*, *mttB*, *nad4L* and *rpl16*, which started with three atypical initiator codons: *rps4* and *mttB* with TTG, *nad4L* with CTG, and *rpl16* with GTG. These exceptions (TTG, CTG and GTG) have long been known as translational initiation codons in other species [31–33]. The mitochondrial *rpl16* gene starting with GTG were consistently observed among all Oleaceae mitogenomes, as shown in *Arabidopsis rpl16*, even if ATG was located upstream (Fig. S5), further confirming the previous finding in Oleaceae [34]. This change was likely induced by premature truncation resulting from RNA editing as explained in *Arabidopsis* and *Petunia rpl16* [35]. Two internal stop codons were detected in the upstream region of the start codon GTG of *rpl16* in all Oleaceae species (Fig. S5). Both internal stop codons cannot be corrected at the RNA level as indicated by RNA-editing sites [34]. An A-rich region with the AAGAAx/AxAAAG motifs, followed by a Shine-Dalgarno (SD) motif-like sequence (AGG), was detected at 6-bp upstream of GTG (Fig. S5). Intriguingly, we also detected the AAGAAx/AxAAAG motifs situated upstream between ATG and the other alternative initiator codons (TTG and CTG) in *rps4*, *mttB*, and *nad4L* (Fig. S6-S8). Two to eight internal stop codons were detected between the ATG and these start codons. These findings are concordant to a recent study, which identified an mTRAN-mRNA interaction governing mitochondrial translation initiation in land plants [36], highlighting conserved A/U-rich motifs, such as AAGAAx/AxAAAG, in the 5' regions of mitochondrial mRNAs as ribosome binding sites directly targeted by mTRAN proteins. All PCGs terminated in the common stop codons TAA, TAG, or TGA.

Repeat-mediated recombination in jasmine mitogenomes

A total of 132, 132 and 136 simple sequence repeats (SSRs) were non-randomly distributed with densities of 259.36, 258.37 and 254.08 SSRs/Mb for "HT", "DP" and "SP", respectively (Table S9, Fig. 3a). Only four SSRs were located in a CDS regions [*nad1*-(T)10, *matR*-(TCTAG)3, *rrn18*-(GAAA)3 and *rrn26* -(CT)5] that exhibited a strong bias towards AT-rich, while the remainder were in intergenic spacers or introns. Overall, the nucleotide composition of SSR motifs were strongly biased toward AT-rich (69.70-70.59%) in mitogenomes (Table S10). All SSRs, ranging from 12 to 18 bp in each mitogenome, fell within the class II group (< 20 bp). The most predominant repeat length was 12 bp (50.76–52.21%), followed by 10 bp (34.56–34.85%). Tetranucleotides were the most frequent SSR motifs (31.82–32.35%), while penta- and hexanucleotides were the least common (1.47–1.52%) (Fig. 3a).

A total of 922, 930 and 1,064 pairs of dispersed repeats were identified with a minimum size of 30 bp in "HT", "DP" and "SP" mitogenomes, respectively, consisting of forward repeats and palindromic repeats (Fig. 3b and c; Table S11-S13). No reverse or complement repeats were found. The most common repeat length was 30 bp. The longest dispersed repeat pair in "HT" and "DP" mitogenomes were forward repeats measuring 5,352 bp and 5,421 bp, respectively. In "HT", this largest forward repeat pair (5,352 bp) was separately situated in Chr1: 1–5,352 bp and Chr4: 25,018–30,369 bp (Fig. 3d, Table S11 and S14). In "DP", this largest pair (5,421 bp) was detected bp in Chr1 (Fig. 3e, Table S12 and S14). The longest repeat pairs included duplicated ribosomal protein genes (rpl2 and rps10) in "HT" and "DP" mitogenomes. In contrast, the longest repeat pair in "SP" was a palindromic repeat pair of nearly 16,267 bp, consisting of two identical sequences with 99.994% identity in Chr1, containing no PCGs except for *trnM-CAU* in each repeat (Fig. 3f, Table S13 and S14). Our PacBio-based assemblies reflected that subgenomic conformations in the jasmine mitogenomes were abundant and extensively divergent from the mono-chromosomal master-circle conformation (Fig. 3g). It is worth noting that we could not find any clue to the existence of a single master ring in our three mitogenomes. The subgenomic conformations in vivo may exhibit significantly greater stability than the

single master ring in multipartite mitogenomes [37]. Remarkably, as mentioned above, a ~ 5 kb syntenic repeat pair was found in "HT" and "DP" (Fig. 3d, e, g), indicating that the presence of active recombinable large repeats was the major factor contributing to the highly variable structural rearrangement and organization between the two mitogenomes, e.g. from di-ring to tetra-ring. There is an exception as found in the mitogenome of *Nymphaea colorata* that exhibited an extremely low level of recombination frequency, despite possessing two large repeats [38]. Similarly, in our study, a ~ 16 kb repeat pair was identified in Chr1 of "SP", which was singly present in the other two mitogenomes (Fig. 3f, g). However, we did not observe any occurrence of homologous recombination mediated by this large repeat pair. The activity of homologous recombination involving large repeats may exhibit polymorphism across three *J. sambac* varieties. This finding also suggests that the "HT" mitogenome is structurally and genetically closer to "DP" than to "SP", and likely evolved from "DP". More cases in the "SP" cultivar should be examined to see if there are other interchangeable isomeric and subgenomic rings in "SP" mitogenomes.

Intriguingly, we found that almost all the rearrangement breakpoints were closely connected with repeats by conducting collinearity analysis. All collinear repeat pairs linked to breakpoints across three mitogenomes were displayed (Fig. 3h). In contrast to large repeats, short repeats (< 500 bp) are thought to be subject to not-infrequent recombination [14, 39, 40]. However, in our case, we noticed that some short-repeat pairs ranging from 36 bp to 273 bp in three jasmine mitogenomes, especially three pairs with intermediate lengths: 133 bp, 272 bp, and 273 bp, were likely involved in mediating homologous recombination (Fig. 3h; Table S11-S13), similar to the phenomenon in *Glycine max* [21] and *Silene vulgaris* [23]. These short repeats acted just like large repeats, promoting homologous recombination events albeit with lower efficiency. In the near future, we aim to reconstruct mitogenome architectures in additional jasmine varieties to examine patterns of structural configurations, repeat-mediated rearrangements, and their potential regulation of CMS occurrence.

Syntenic analysis and selection pressure in oleaceae

We compared all seventeen complete mitogenomes of Oleaceae available in GenBank. The mitogenomes ranged in size from 508,929 bp in *J. sambac* to 848,451 bp in *Ligustrum quihou* (Table S15). Co-linearity in genome structure and gene placement among these Oleaceae mitogenomes was assessed (Figs. 3g and 4, and S9, and Table S15). Mitogenome organization and gene order were highly variable among all Oleaceae species, while the PCG content was broadly conserved. All species in Oleaceae shared an identical complement of 37 distinct

PCGs, although the copy number of some genes varied from species to species (Table S15). As reported in other plant mitogenomes [41], the major variations among Oleaceae species are in the gene content and copy number of ribosomal proteins. We detected that *rpl23*, which codes for ribosomal protein L23, was present in three genera of Oleaceae (*Hesperelaea*, *Osmanthus*, *and Olea*) but was lost in other genera. This is supported by the phylogenetic tree, where these three genera appear closely clustered together, diverging from the other genera (Fig. 4). *Osmanthus* and *Hesperelaea* also had five and one additional genes derived from chloroplast respectively (Table S15), consistent with their grouping in a single cluster, while Olea formed a separate adjacent cluster (Fig. 4). We inferred that the shared losses/gains could be traced back to an early common ancestor. Three common rRNA genes (*rrn5*, *rrn18* and *rrn26*) were found in all Oleaceae species but the number and type of unique tRNAs varied from 15 to 20 (19 to 27 including copies) (Table S15). All thirty-seven common PCGs in Oleaceae were further selected for the positive selection analysis (Table S16). Four positively selected PCGs (*nad5*, *rpl10*, *rps7*, and *sdh3*) were found with significant posterior probabilities for codon in the BEB test, although *p*-values of positive selection were not significant in all gene groups (p-value > 0.05).

The phylogenetic relationships among Oleaceae species were resolved based on the entire mitogenome sequences (Figs. 4 and 5a). This tree was largely in accordance with traditional taxonomy that divides these species into three tribes: Jasmineae, Forsythieae and Oleeae. With respect to mitogenome organization, many structural rearrangements have taken place among lineages, even within a given species (Fig. 4). Barely no syntenic gene blocks were arranged together among different tribes, illustrating the genomic diversity within Oleaceae. Within *J. sambac* and *Olea europaea* species, the highly variable gene order was observed. A Mauve alignment among four jasmine species showed that they shared eleven locally collinear blocks (LCBs) (Fig. S9). Despite the conserved gene sequences and content, the four *J. sambac* mitogenomes exhibited multiple genome rearrangements (e.g., single-ring vs. multi-ring) and shuffled gene orders, reflecting the fast-evolving nature of the jasmine mitogenome, likely driven by homologous recombination of repeat elements.

Phylogenetic inference in lamiales

To explore the phylogenetic position of *J. sambac* within Oleaceae, the phylogenetic analysis was conducted based on codon sequences of 37 common single-copy PCGs from 17 species (Fig. 5a). The concatenated DNA codon sequence length was 31,407 bp for each mitogenome. The ML tree (scored under GTR+I+G model)
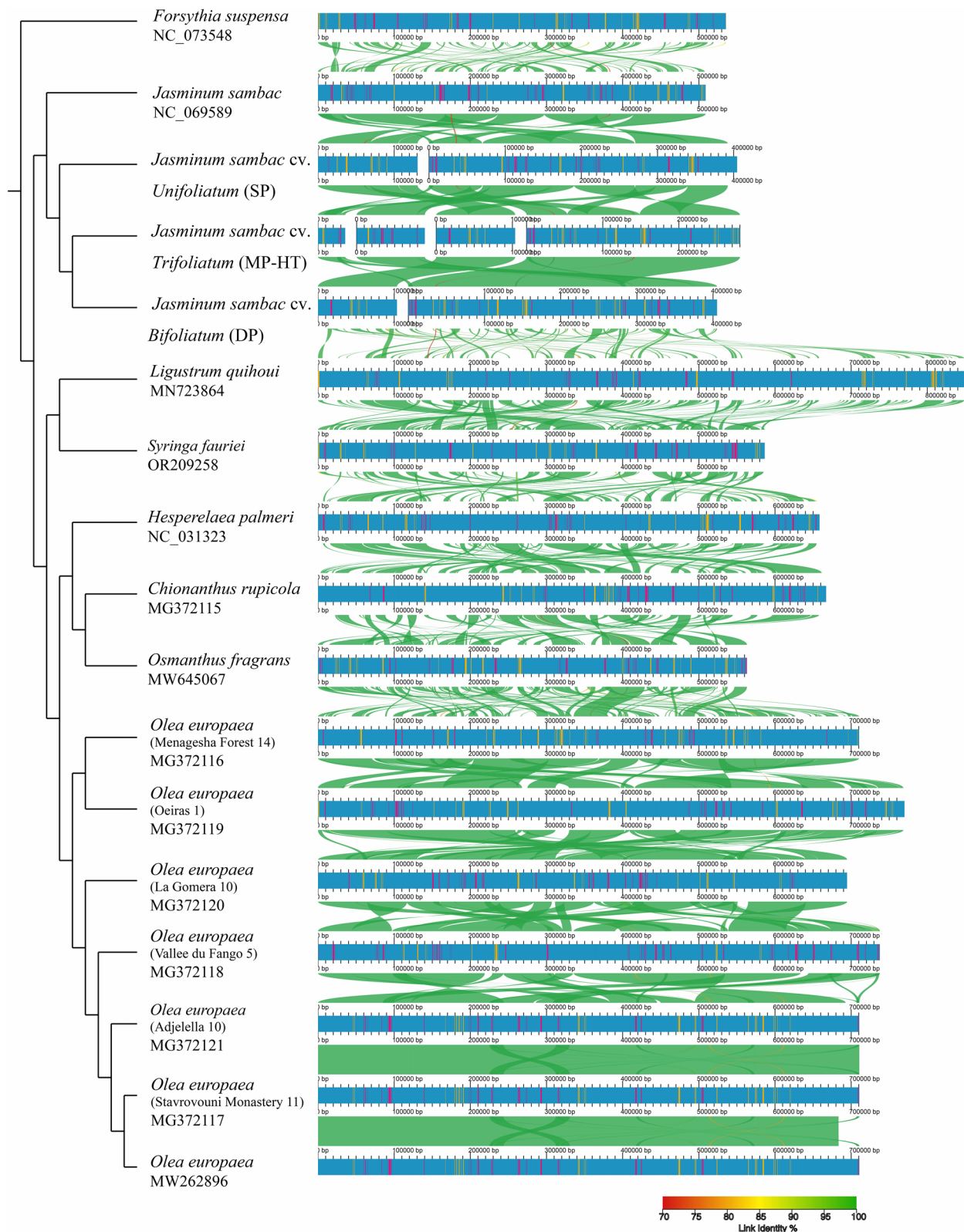
Fig. 4 Syntenic comparisons of seventeen mitochondrial genomes in Oleaceae. Based on AliTV software, syntenic comparisons of linear mitochondrial maps relative to a maximum-likelihood (ML) phylogenetic tree of 17 Oleaceae species were shown. Pairwise comparisons expressed as percentage of nucleotide similarity are depicted between panels, which connect different homologous genomic regions. The x-axis denotes the site of the feature on the mitogenome. Genes are highlighted in orange and red colors, representing they are transcribed in the clockwise and counterclockwise directions, respectively. The ML phylogenomic tree constructed with 37 common single-copy protein-coding genes from 17 mitogenomes is shown on the left
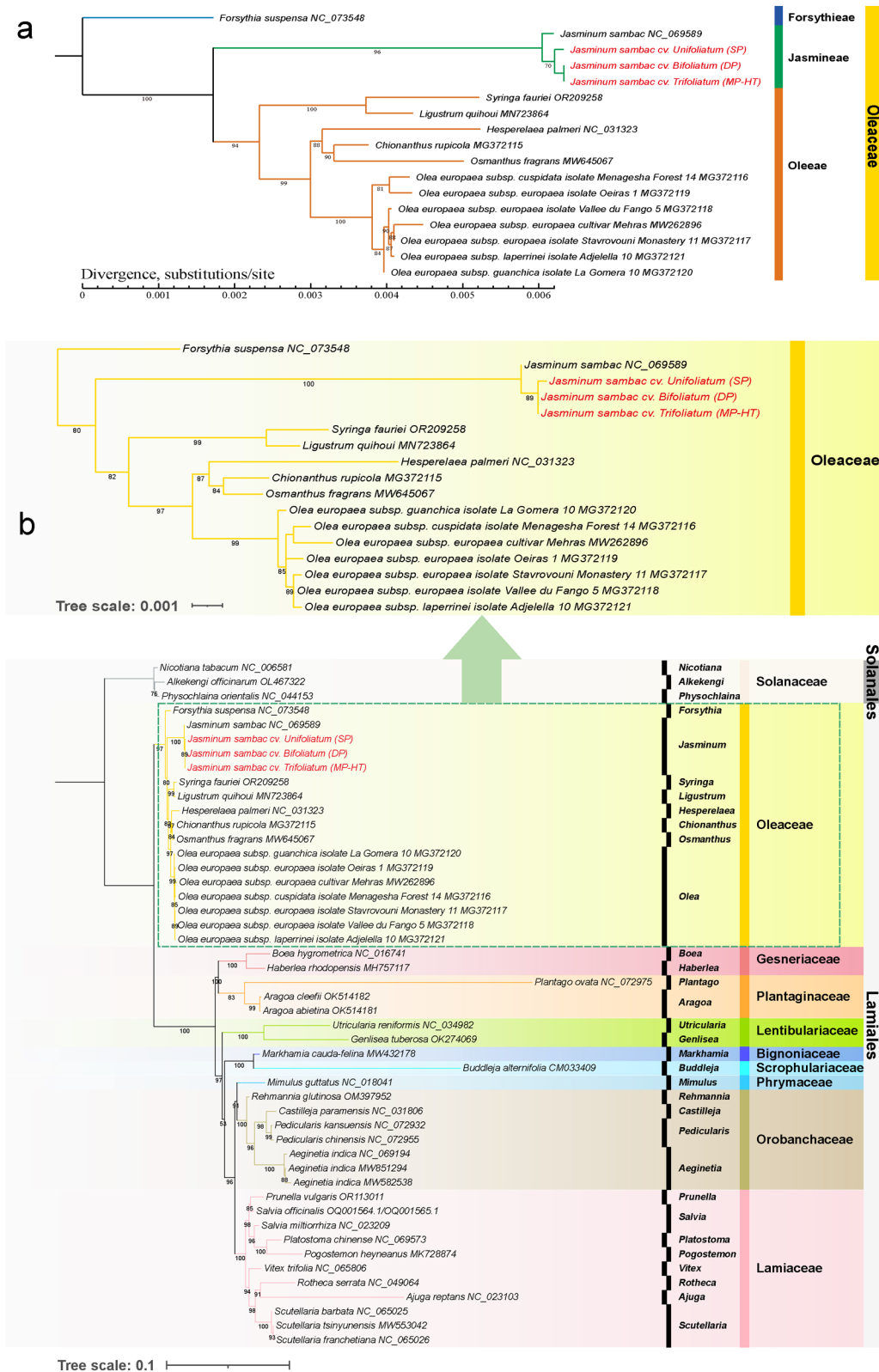
Fig. 5 Maximum-likelihood (ML) phylogenomic tree for Oleaceae and Lamiales species constructed with single-copy genes. (**a**) The ML tree of seventeen Oleaceae mitogenomes obtained by PhyML v3.0 with the best-fit substitution model "GTR + I + G". The concatenated CDS nucleotide sequences and their alignments of 37 shared single-copy protein-coding genes (PCGs) were used to reconstruct the ML tree. (**b**) The ML tree of 45 mitogenomes in Lamiales with three Solanales species as outgroup were constructed by PhyML v3.0 based on the "LG + I + G + F" model. The concatenated amino acid dataset comprising 19 common single-copy orthologous PCGs among 48 species and their amino acid alignments were used

was well supported with high bootstrap values (most MLBS≥88%). *Forsythia suspensa* from Forsythieae occupied a basal position in Oleaceae, while the others were separated into two main branches, consistent with tribe classification (Jasmineae and Oleeae) (MLBS≥94%). *J. sambac* accessions clustered into a single distinct clade in between the clade of *F. suspensa* and the clade consisting of *Syringa fauriei* and *Ligustrum quihoui*. The "HT" jasmine was nested within the same cluster as the "DP" jasmine, and both of them diverged from the common ancestor of "SP". The divergence time of the three jasmine accessions were very close to each other when calculating the *Ks* value. The branch leading to *J. sambac* is much longer than that of other Oleaceae species, implying a higher rate of evolution in the *Jasminum* genus compared to other genera within Oleaceae. All seven *O. europaea* accessions clustered together into a monophyletic clade, suggesting a common origin they shared.

The mitochondrial multiple-gene ML tree for Lamiales was reconstructed using three Solanales species as outgroup with a concatenated amino acid (aa) alignment of 19 common single-copy orthologous PCGs and 5,823 aa positions (Fig. 5b, Table S17). Our extensive taxon sampling covers all mitogenomes hitherto assembled within Lamiales, making this ML tree the most complete one ever constructed. This tree also unveils the most detailed genetic evolutionary relationships within Lamiales, with many nodes resolved for the first time. Taxonomically well-placed species were recovered with strongly supported clades. Similarly, in Oleaceae, all seventeen mitogenomes formed a monophyletic cluster with 97% bootstrap value, where *F. suspensa* was the first to diverge from the other lineages, supporting the phylogenetic tree constructed by complete cp. genomes [42]. The topologies of both ML phylogenies constructed from CDS nucleotide alignment (Fig. 5a) and amino acid alignment (Fig. 5b) were largely congruent, except for the relative relationships among *O. europaea* (olive) accessions. The relationships among olive mitogenomes appear to be interlaced and intricate, thus it is hard to be resolved despite repeated efforts had been made [12, 43]. The phylogenetic placement of Oleaceae formed a sister group to the clades of Gesneriaceae and Plantaginaceae with nearly full support, which is congruent with other phylogenies inferred from single or multiple cytoplasmic genes [34, 44–46].

DNA transfer between jasmine organelle genomes

After filtering, a total of 51.8 Gbp of Illumina high-quality data consisting of 345,331,888 paired-end (PE) clean reads were generated, representing around 106.27x genome equivalents. The complete cp. genome of *J. sambac* "HT" was assembled into a single circular double-stranded DNA molecule measuring 163,464 bp in length,

with an overall GC content of 37.58% (Fig. 6a, Table S18). The typical quadripartite structure was observed in this cp. genome, consisting of a large single copy (LSC) region of 90,739 bp and a small single copy (SSC) of 13,223 bp, separated by two identical inverted repeats (IRs) of 29,751 bp. The GC contents of IR regions (41.38%) were higher than those of LSC and SSC regions (35.84% and 32.49%), primarily due to the GC richness of ribosomal RNAs in IR regions. The "HT" cp. genome shared most gene features with "DP" and "SP" cp. genomes [47](Table S18). It encoded a total of 135 genes that were scattered singly or in groups, among which 112 genes were unique, comprising 79 PCGs, 29 tRNAs, and 4 rRNAs. Out of the 79 PCGs, 11 unique genes contained single or double introns. The IR regions harbored nineteen duplicated genes, including 8 PCGs, 4 rRNAs, and 7 tRNAs. Two copies of 16 S-trnI-trnA-23 S–4.5 S-5 S ribosomal RNA operons were identified in IRs. One *trans*-spliced gene *rps12* with two copies was found in the "HT" cp. genome, each containing three exons. Two copies of *rps12* shared the first exon being in the LSC region, while the other two duplicated exons were located in IR regions. The total length of PCGs, tRNAs and rRNAs was 83,106 bp, 2,929 bp and 9,054 bp, respectively, accounting for 50.84%, 1.79% and 5.54% of the cp. genome, whereas the non-coding regions, including intergenic spacers and introns, covered the remaining 43.62% of the genome.

With regard to nucleotide sequence and gene collinearity among the three jasmine cp. genomes, a near-perfect synteny with identical gene organization was observed between "DP" and "SP", but a large 19.9 kb inversion located at 46,285−66,200 bp was detected in "HT", encompassing 18 genes (10 PCGs and 8 tRNAs) when compared with the other two (Fig. 6b). Apart from this inversion, highly syntenic relationships of genes among three varieties were observed. It should be noted that the *accD* gene coding for a subunit of acetyl-CoA carboxyltransferase (ACCase) in "DP" and "SP" cp. genomes was truncated to non-functional fragments in the "HT" cp. genome. The *clpP* gene in "DP" and "SP" cp. genomes was replaced by *clpP1* and an additional copy of *rpl23* in "HT". Variations in the presence/absence of accD are common across different *J. sambac* cp. genomes and are linked with rearrangement hotspots, despite the generally conservative nature of cp. genomes [42]. The insertion of tandem repeats in *accD* makes it a hypervariable gene with elevated substitution rates that may easily expand or contract, and also has a high capacity to mediate the rearrangement of cp. genome [48, 49]. This chloroplast-resided *accD* was reported to be closely associated with leaf growth [50], leaf fatty-acid content [51], and embryo development [52], which could explain the low fatty-acid content, weak growth performance in "HT" leaves, and the extremely low pollen viability in jasmine. Further

Fig. 6 (See legend on next page.)
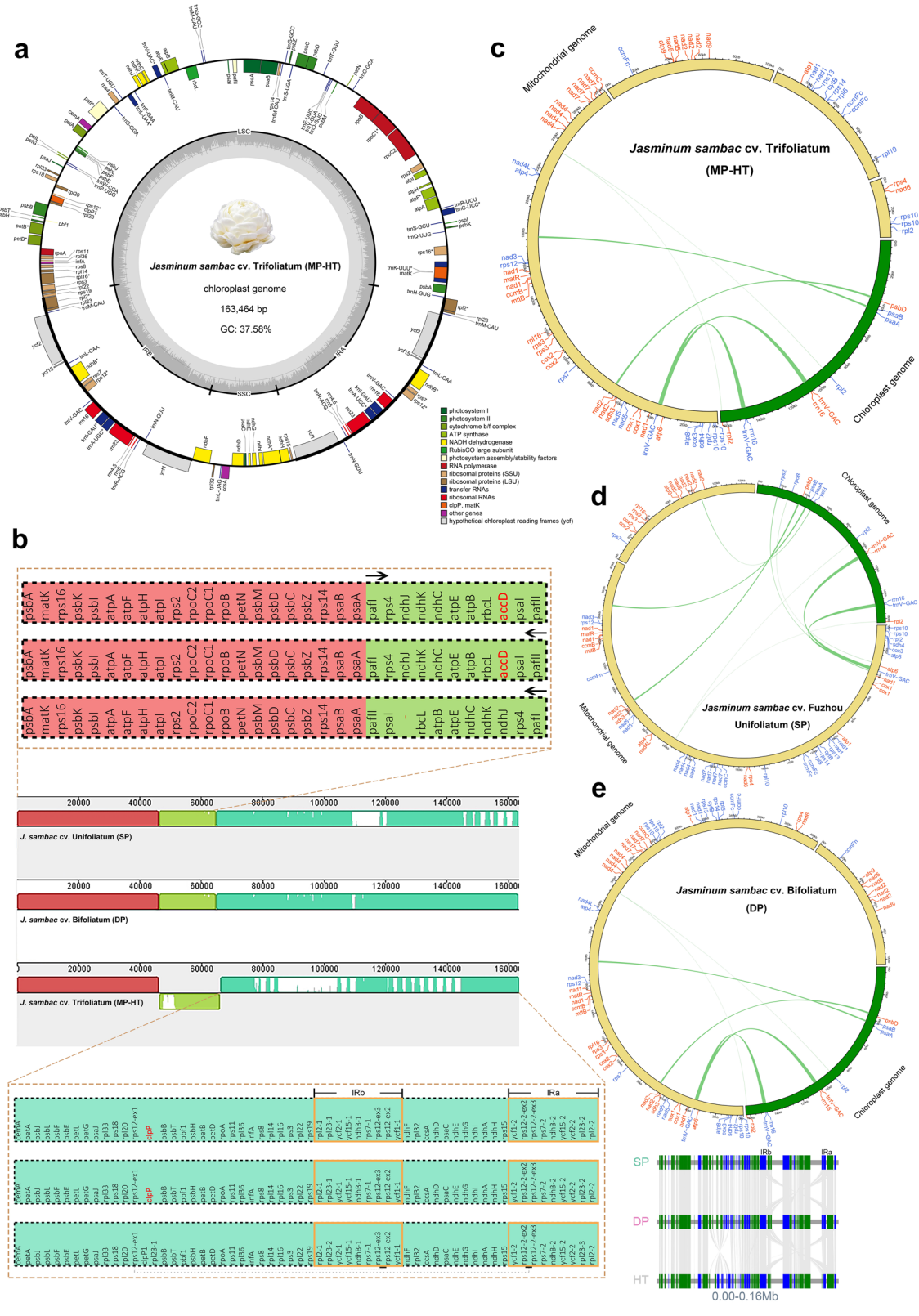
(See figure on previous page.)

Fig. 6 Syntenic comparisons of three jasmine chloroplast genomes and chloroplast-like sequences in jasmine mitogenomes. (**a**) Circular chloroplast genome map of *J. sambac cv. Trifoliatum* "HT". Genes facing inside and outside of the circle are transcribed in the clockwise and counter-clockwise directions, respectively. Genes are colored according to different functional groups. The dark-grey inner circle represents the GC content along the cp. genome. The thick lines separate the genomes into four parts: large single-copy (LSC), small single-copy (SSC) regions, and the direct ribosomal operon repeats (DRa and DRb). Intron-containing genes are indicated with asterisks (*). (**b**) Gene synteny comparison of three jasmine mitogenomes aligned using Mauve and McScanX. Red, yellow-green and emerald-green blocks represent three large locally collinear blocks (LCBs) that contain conserved syntenic clusters of protein-coding genes. A sequence identity similarity profile is shown inside each block. Gene content and arrangement of three cp. genomes are also shown in each block. A large inversion (the yellow-green block) was detected in "HT" when compared with other two. (**c-e**) Schematic representation of the distribution of homologous sequences between chloroplast genomes and mitogenomes in *J. sambac*. DNA transfer between three sets of chloroplast and mitochondrial genomes: "HT" (**c**), "SP" (**d**) and "DP" (**e**). The green and light-yellow arcs represent chloroplast and mitochondrial genomes, respectively. The lines between arcs correspond to transfers of homologous segments. Genes that overlapped with the homologous sequences are labeled next to the circles. Genes labeled in red and blue colors are transcribed in the clockwise and counterclockwise directions, respectively

investigation of cp. genomes across *Jasminum* species is needed to elucidate the patterns of *accD* presence/absence pattern and its evolutionary trace.

The number of chloroplast-derived insertions in mitogenomes were relatively constant between "HT" and "DP", with more fragments found in "SP". A total of six, six, and nine homologous DNA fragments ranging from 110 bp to 2,665 bp were identified in *J. sambac* mitogenomes of "HT", "DP" and "SP" that shared 0.91–0.99% identity with the corresponding cp genome (Fig. 6c-e, Table S19). Among these, five PCG fragments (*psaA*, *psaB*, *psbD*, *rpl2-1*, *rpl2-2*) and a 16 S-*trnV-GAC* pair were identified. The "SP" mitogenome also contained additional chloroplast gene fragments (rps2, rpoB, ycf3), making its homologous sequences at least 1,169 bp longer than those in the other two mitogenomes (6,205 bp vs. 5,026 bp) and contributing to its larger size. Chloroplast-derived genes were partially transferred into non-coding regions of mitogenomes to form non-functional components except for the intact *trnV-GAC*, which was contained in the longest homologous sequence (2,703 bp). The *trnV-GAC* gene might have migrated from the *J. sambac* mitogenome into two separate positions of the cp. genome, forming two independent *trnV* genes. It has been known that frequent gene conversion could also account for the *trnV-GAC* homologues between cp. and mt genomes [53].

A specific 14.2 kb NUPT insertion in "HT" nuclear genome
The distribution of genome-wide NUPTs and NUMTs was identified in three jasmine nuclear genomes. NUPTs and NUMTs were widespread in all thirteen chromosomes with biased distribution in the nuclear genome (Fig. 7a-c, Table S20-S21). In total, **804** (172.72 kb), **794** (153.41 kb) and **821** (158.45 kb) NUPTs were unambiguously determined in the "HT", "DP" and "SP" genomes, respectively, with insert lengths ranging from 34 to 7,219 bp (Table S20). Likewise, **972** (153.09 kb), **1,054** (159.62 kb) and **1,125** (170.14 kb) NUMTs were obtained with insert lengths of 34 to 6,360 bp (Table S21). Overall, the cumulative NUMT insert lengths and the amount in the nuclear genome are consistent with the evolutionary

timeline of the three jasmine accessions (Fig. 5a), indicating that the more anciently evolved accession contained a higher level of NUMT accumulation ("SP" > "DP" > "HT"). However, the cumulative length of NUPTs showed the opposite trend in "HT" due to a 14,252 bp NUPT insertion in chromosome 7 (Fig. 7a and Table S20). This 14.2 kb insertion, by contrast, was absent in the "DP" and "SP" nuclear genome. The 14.2 kb large insertion consisted of two 7 kb fragments that originated from different regions of the "HT" cp. genome and was integrated into "HT" Chr7: **35,015,809 – 35,030,060**. Intriguingly, two intact PCGs (*ndhH*, *rps15*) and two partial PCGs (*ycf1*, *ndhA*) from the cp. genome were included in this insertion (Fig. 7d, Table S20). The first 7,219 bp NUPT fragment,

integrated at Chr7: **35,015,809 – 35,023,027**, contained a nonfunctional 833 bp segment of *ycf1* derived from the cp. IR region. The second NUPT fragment, measuring 7,033 bp at Chr7: **35,023,028–35,030,060**, contained the intact *ndhH* and *rps15* genes, along with partial *ycf1* and *ndhA*. To validate the authenticity of this 14.2 kb insertion in "HT" nuclear genome, the entire dataset of "HT", "DP" and "SP" PacBio long reads were aligned to the "HT" nuclear genome, respectively. Three NUPT junction sites (35,015,809 – 35,023,027–35,030,060) in Chr7 of the "HT" nuclear genome were regarded as significant joint points (Fig. 7d-f, Table S20). The alignment revealed that all three junction sites were supported by multiple junction-spanning reads with coverage of 16x, 17x and 19x, respectively, when aligning "HT" reads to the "HT" refgenome, approaching the depth of "HT" whole-genome sequencing. Notably, two CCS raw reads, measuring 26,331 bp and 17,372 bp, respectively, entirely spanned all three junction sites, covering the entire insertion and its flanking sequences. By contrast, no "DP" or "SP" reads spanned these junction sites in the "HT" refgenome as expected. PCR validation of junction regions further confirmed the presence of this 14.2 kb chloroplast-derived integration in "HT" Chr7, with successful amplification Bands in "HT" but not in "DP" or "SP" (Fig. 7g). These results strongly support the accurate assembly of the

large bulk of 14.2 kb NUPT insertion along with its border regions in Chr7 of the "HT" genome.

We observed that both NUPTs and NUMTs tend to cluster in non-functional regions (gene desert regions) across all three jasmine varieties examined, like transposable elements rich (TE-rich) and the proximal centromere regions (Fig. 7a-c). Longer and younger NUPTs and NUMTs appeared to be located closer to centromere regions, aligning with previous findings in other species [54, 55]. The TE-surrounded localization of nuclear organelle DNAs (norgDNAs) in the host genome in our findings implied similar genomic roles for TEs and norgDNAs with restricted functionality, both of which may play an important role in uncovering their evolutionary footprints. Additionally, we inferred that the exogenous inserts of norgDNAs into non-functional areas could protect them from exposure under the high pressure of degradation and elimination in gene-rich regions, where the host genomes have developed a mechanism to prevent the interference or interruption in the faithful execution of functional gene expression. Still a few norgDNAs were transferred into functional regions with relatively low gene density. NUPT insertions were identified within the gene sequences of 97, 89 and 90 PCGs (0.37-0.40% of total genes) in "HT", "DP" and "SP" genomes, respectively, while NUMTs became partial sequences in 127, 116 and 123 PCGs (0.49-0.53%) in three nuclear genomes (Table S20-S21). An exception was observed for the 14.2 kb insertion, which included two intact chloroplast-derived genes and was inserted into the distal gene-rich region of the "HT" nuclear genome (Fig. 7a-c). No intact organelle-derived genes were detected in the other two varieties. The presence of more and longer norgDNAs integrated into gene-rich locations in "HT" indicates that many norgDNAs have recently immigrated into the "HT" nuclear genome.

Genome-wide distribution of common and specific NUPTs and NUMTs

Based on the pipeline for identifying common and unique NUPTs/NUMTs, we detected that a large fraction of NUPT (97.30-99.32%) and NUMT inserts (94.79-99.04%) were shared by three accessions (Table S22-S24, Fig. S10-S12). This pipeline also confirmed that this 14.2 kb insertion was "HT"-specific NUPT integration (Fig. 7h, Fig. S10 and S11, and Table S23). Both NUPT fragments located within the 14.2 kb insertion exhibited an extremely high sequence identity with the "HT" cp. genome (99.88% and 99.94%) (Table S23), much higher than the average identity of 94.55-94.79% (Table S22), suggesting that this large NUPT insertion was formed as a result of the most recent chloroplast-to-nucleus transferred events in the "HT" genome. This young large NUPT insertion, along with other newly transferred

norgDNA fragments in "HT", may tend to experience rapid recombination, fragmentation, shuffling, and elimination to maintain the equilibrium of plant nuclear genome stability, as previously reported [54]. In this case, it is more likely that the "HT" genotype has a rapidly evolving or mutable genome among the jasmine genotypes, possibly as an adaptation to changing environments. The genomic turbulence and mutability under evolutionary pressure in "HT" could also explain its reduced resistance to stressors such as pests, pathogens, and cold damage, as well as the increased occurrence of flower deformities. In addition, it is still an enigma whether the functional transfer of chloroplast genes plays an indispensable role in gene expression regulation networks and genome evolution. Although two intact open read frames of chloroplast genes (*ndhH*, *rps15*) have been horizontally transferred into nucleus, it remains unclear whether suitable transcriptional regulatory elements, such as gene promoters, enhancers and terminator sequences adjacent to these genes, exist to facilitate the proper launch of transcription within the nucleus. Additionally, suitable transit peptides are required to target the protein products back into the chloroplast [56]. This large chunk of NUPT insert, which contains intact genes along with their border sequences, offers the possibility of the co-transfer of their regulatory motifs into the nuclear genome, potentially enabling the proper functioning of these genes. Further wet experiments are needed to investigate whether these chloroplast-derived coding genes can function properly and actively within the nucleus. It also remains to be elucidated whether this insertion has any deleterious effects on adjacent genes.

Conclusion

In the present study, three complete gap-free jasmine mitogenomes with di-loop ("SP"), di-loop ("DP") and tetra-loop ("HT") conformations were assembled following a pipeline integrating PacBio long-read sequencing, bioinformatics analysis and FISH verification. The ~5 kb forward repeats and several small repeats (<500 bp) likely mediate the formation of multiple circular MtD-NAs through genomic rearrangements and homologous recombination, representing an indispensable driving force for mitogenome polymorphism and evolution in jasmine. Despite their structural divergence, these three mitogenomes showed considerable conservation in genetic content, including gene content, SSR distribution, and homologous DNAs with nuclear and cp. genomes. In Oleaceae, mitogenomes were broadly conserved in gene content but highly variable in structure and gene order, even within a given species, suggesting their high rearrangement rates. Notably, the *Jasminum* genus exhibited a higher rate of evolution compared to other genera. The most comprehensive phylogenetic
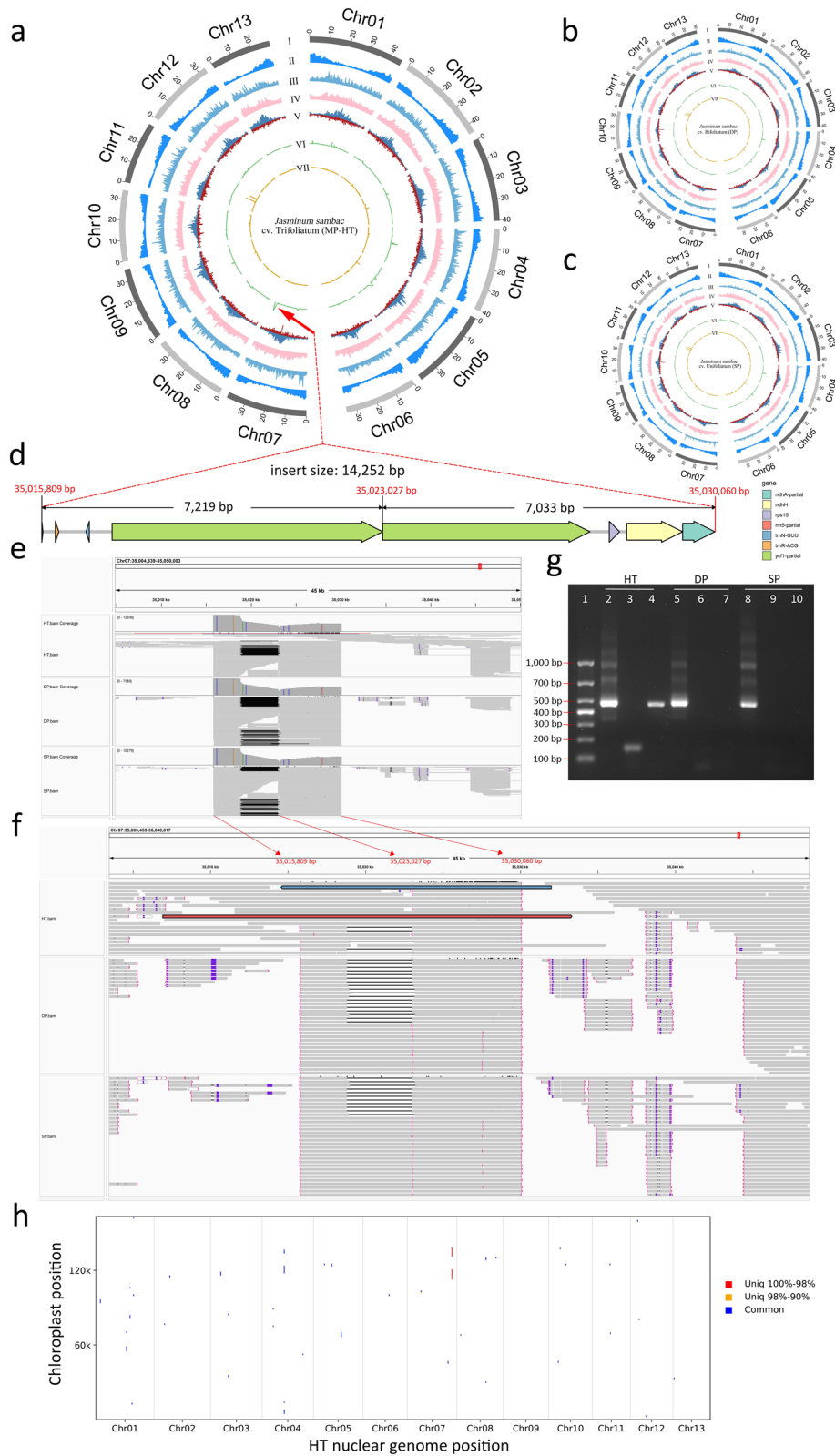
Fig. 7 (See legend on next page.)

(See figure on previous page.)
Fig. 7 The integration of a 14.2 kb chloroplast-derived fragment into "HT" nuclear genome. **(a-c)** Circos diagrams illustrating the genome-wide distribution of NUPTs and NUMTs in (a) "HT", (b) "DP", and (c) "SP". The seven rings from outermost to innermost indicate: thirteen chromosomes of *J. sambac* (I); gene density (II); DNA Transposon abundance (III); Tandem repeats (IV); LTR Gypsy and LTR Copia (V); chloroplast (VII) and mitochondrial (VIII) distribution in the *J. sambac* nuclear genomes. **(d)** Graphical alignment of the 14.2 kb chloroplast-derived insertion to Chr7 of "HT" nuclear genome. The insertion consists of two chloroplast-derived fragments. the first fragment (7,219 bp) containing a chloroplast *ycf1* partial gene was integrated at Chr7: 35,015,809 − 35,023,027. The second fragment (7,033 bp) at Chr7: 35,023,028–35,030,060 contained two intact genes (*ndhH*, *rps15*) and two partial genes (*ycf1*, *ndhA*). **(e-f)** Validation of the "HT"-specific chloroplast-derived insertion using PacBio reads. PacBio reads of "HT", "DP", and "SP" were aligned against the "HT" Chr7 and visualized in IGV. "HT" alignment showed that all three junction sites were supported by multiple junction-spanning reads. Two CCS raw reads (IDs: m64066_210130_204431/162007692/ccs and m64066_210130_204431/81723944/ccs) (highlighted in blue and red) can entirely span three junction sites, covering the whole insertion and their flanking sequences. No "DP" or "SP" reads can span over these junction sites in "HT" refgenome. **(g)** PCR validation of specific chloroplast-derived insertion in "HT". Lane 1: 1,000 bp DNA ladder; Lane 2, 5 and 8: detection of 5 S rDNA from nuclear genome as control in "HT", "DP" and "SP"; Lane 3, 6 and 9: detection of the left boundary of the specific chloroplast-derived insertion in "HT", "DP" and "SP"; Lane 4, 7 and 10: detection of the right boundary of the specific chloroplast-derived insertion in "HT", "DP" and "SP". The left and right boundaries can be amplified in "HT" but not in "DP" and "SP". **(h)** Genome-wide distribution of common and unique chloroplast fragments in the "HT" nuclear genome. The x-axis denotes the position of NUPTs on "HT" nuclear chromosomes, and the y-axis denotes the position on the "HT" chloroplast genome. Red fragments indicate "HT"-specific NUPTs, including two 7 kb insertions derived from different regions of the "HT" chloroplast genome

tree to date with well-resolved internal relationships in Lamiales offers novel insights into mitogenome genome evolution within Oleaceae and among Lamiales lineages. We also *de novo* assembled and annotated the complete "HT" cp. genome into a single circular contig based on Illumina short reads. The presence of a large 19.9 kb inversion, the absence of *accD*, and an additional copy of *rpl23* in "HT" provide intriguing clues that this variety possesses the fastest-evolving cp genome. Interestingly, evidence from PacBio sequencing raw data and PCR verification confirmed a 14.2 kb large chloroplast-derived sequence horizontally and specifically transfers to the "HT" nuclear genome. In summary, the syntenic comparison, phylogenetic inference, and norgDNA footprints support the assertion that the "SP" jasmine is the common ancestor of "DP" and "HT", with "HT" being a recently evolved genotype derived from "DP". The newly assembled jasmine organelle genomes have added to our knowledge of genetic variance and diversity behind varying traits in jasmine.

Methods
Plant materials and genome sequencing
The *J. sambac* multi-petal cultivar 'Hutou' ("HT") was conserved at The Botanical Garden of Minrong Tea Industry Co. Ltd, Fuzhou, China, under the specimen voucher number GDHTML3. Three-month-old rooted cuttings of this specimen were kindly provided by the chief scientific officer of Minrong Tea Industry and transplanted into plastic flowerpots with a top diameter of 27.5 cm and a depth of 31 cm in a greenhouse of Fujian Normal University (26°01′36.5″ N, E 119°12′33.5 ″ E), Fuzhou, Fujian Province, China. The soil moisture and plants were checked daily and watered as needed. High-quality total genomic DNA for genome sequencing was extracted from young tender healthy leaves of the best-growing "HT" individual using the Qiagen Genomic DNA extraction kit (Qiagen, Germany). Samples were collected from several tissues including roots, mature

stems, young and mature leaves, as well as six stages of flower development of "HT", for RNA isolation and RNA-seq. All samples collected were promptly frozen in liquid nitrogen for at least 20 min, followed by preservation at -80 °C in the freezer prior to DNA and RNA extraction. DNA purity and concentration (>100 ng/µl, OD 260/280 close to 1.8, OD260/230 close to 2.0) were assessed by a NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific, United States). Pacbio (Sequel II platform) and Illumina (NovaSeq platform) sequencing of the "HT" genome was carried out by Berry Genomics Company (Beijing, China).

The PacBio high-fidelity (HiFi) whole-genome sequencing data and RNA-seq data of "SP" (*J. sambac* cv. *Fuzhou Unifoliatum*) and "DP" (*J. sambac* cv. *Bifoliatum*) generated in our previous research [2] were used for mitogenome assembly and annotation, which could be downloaded from the BIG Data Center (https://bigd.big.ac.cn/) under project number PRJCA006739.

Mitogenome assembly and preliminary verification
All PacBio-generated reads from the "HT" genome were aligned to the mitogenome of the closely related species *L. quihoui* (GenBank: MN723864.1) using Minimap2 v2.10-r761 [57] with default parameters to generate the Pairwise mApping Format (PAF) file. Aligned reads with mapping quality>10 were considered as homologous reads and retained as potential jasmine mitochondrial reads. Homo-reads were extracted from the PacBio subread pool, which were then subjected to, Flye v2.8.3 [58] and Canu v2.1.1 [59] for *de novo* assembly. The draft assembled contigs were BLASTN (2.15.0+) [60] searched against the mitochondrial coding sequences (CDS) of *L. quihoui* to identify candidate "HT" mitogenome contigs. All seven candidate contigs from Flye were manually connected based on overlapping regions, ensuring complete coverage of all mitochondrial genes and resulting in the final "HT" mitogenome with 4-loop structures (Table S1). Likewise, PacBio sequencing reads for "SP" and "DP",

Fang *et al. BMC Plant Biology*   (2024) 24:861

Page 18 of 23

comprising approximately 15.35 Gb and 12.11 Gb data, were aligned against the newly assembled "HT" mitogenome using Minimap2 to mine potential mitochondrial reads. Then these potential reads were subjected to Flye and Canu programs for *de novo* assembly. The draft assemblies were BLASTN [60] searched against the "HT" mitochondrial CDS to identify candidate mitogenome contigs, yielding two Canu contigs for "DP" and five Canu contigs for "SP", respectively (Table S1). The final "SP" and "DP" mitogenome sequences were generated by identifying connection points based on overlaps among these Canu contigs. The circularity of the "SP", "DP" and "HT" assemblies were checked using the "check_circularity.pl" script from the sprai package (http://zombie.cb.k.u-tokyo.ac.jp/sprai/). We reordered and oriented the assemblies according to syntenic comparisons among "SP", "DP", and "HT" mitogenomes to ensure the same start position and orientation. The assembly workflow for the mitogenomes of "HT", "DP", and "SP" was shown in Fig. S13.

The Illumina short reads and PacBio reads of "SP", "DP", and "HT" were aligned with their respective mitogenomes as reference using BWA v0.7.18 -r1243 [61] and Minimap2, followed by the filtering of unmapped reads, multiple-mapped reads, and PCR duplicates. The sorted Binary Alignment/Map (BAM) format files were generated for downstream analysis. The accuracy of these assemblies was manually verified in Integrative Genomics Viewer (IGV v2.17.0) using the BAM output as a guide.

Fluorescence in situ hybridization (FISH)
Amplification of the mitochondrial DNA sequences was conducted by polymerase chain reaction (PCR) using the primers listed in Table S25. We designed a set of primer pairs with each pair being randomly distributed across each chromosome of the "HT" mitogenome, totaling four pairs. Using Nick Translation Mix (Roche, Mannheim, Germany), the purified PCR products of each mitochondrial chromosome were labeled with Bio-dUTP to indicate mitochondrial locations, while a plasmid with 5 S rDNA from *Oryza sativa* was labeled with Dig-dUTP to mark nuclear locations. FISH experiments were performed as previously described [62]. Briefly, plant roots were pretreated with an 8-hydroxyquinoline solution at room temperature for 3 h before fixation in 3:1 ethanol: glacial acetic acid for 24 h. The roots were then treated in an enzyme mixture at 37 °C for 2 h. The resultant cellular suspension was carefully dispensed onto the slides for further analysis. Chromosome slides were denatured for 1 min on a 70 °C hotplate for DNA strand separation. The hybridization mix containing 50% formamide, 10 mg/mL dextran sulfate, 2× SSC, and 100 ng/μL of each probe, was heated to 95 °C for 10 min, and then applied

to the denatured chromosome on the slides. A coverslip was carefully placed over the sample mixture to seal it. The slides were subsequently placed in a hybridization chamber and incubated overnight at 37 °C to facilitate hybridization. Hybridization was performed for 24 h at 37 °C in a humidified chamber. After hybridization, slides were subjected to three times washes for 5 min in 2× SSC and an additional wash in 1× PBS at room temperature for 5 min. Digoxigenin-labeled probes and biotin-labeled probes were detected using rhodamine-conjugated anti-digoxigenin (Roche Diagnostics, Mannheim, Germany) and Alexa FluorTM488 streptavidin (Thermo Fisher Scientific, Cleveland, OH, USA), respectively. Slides were air-dried and then counterstained with DAPI (Vector Laboratories, Odessa, Florida, USA). Chromosomes and FISH signals were visualized using a BX63 fluorescence microscope equipped with a DP80 CCD camera (Olympus, Tokyo, Japan). Images were adjusted with Adobe Photoshop CC.

"HT" chloroplast genome assembly
The data filtering steps were performed on the Illumina resequencing raw data for the "HT" genome with Trimmomatic v0.39 [63] prior to the chloroplast genome assembly. The resulting clean paired-end reads were utilized as input to the *de novo* assembly of the "HT" cp. genome, which was performed using GetOrganelle v1.7.5 [64] that employs "SP"Ades [65] as the core *de novo* assembler. The reference cp. genomes, "SP" and "DP" *Jasminum sambac* cp. genomes available in the NCBI GenBank (GenBank Acc. No. MN158204.1 and MN158205.1), were used as the seed. A number of potential cp. assemblies were generated in the abovementioned step. Subsequently, the sequence accuracy of final chloroplast assembly, with a particular focus on the inverted repeat (IR) order and IR continuity, was verified and, if necessary, manually corrected through BLASTN searches against the reference cp genomes with an E-value threshold of 10^{-5}. The circularity of the cp. assembly was checked by the script "check_circularity.pl" from the sprai package (http://zombie.cb.k.u-tokyo.ac.jp/sprai/) and the overlapping ends were subsequently removed from the cp. assembly. Finally, the "HT" cp. assembly was reordered and oriented according to the reference jasmine cp. genomes. In-house shell scripts were used to identify the boundaries of the LSC/IR/SSC regions of the three jasmine cp. genomes. The shell scripts were provided on GitHub (https://github.com/Datapotumas/IdentifyCpRegions).

Organelle genome annotation and physical mapping
Protein-coding genes (PCGs) were preliminarily annotated based on a combined strategy of ab initio- and homologous predictions using GeSeq online tool [66],

MITOFY web server (http://dogma.ccbb.utexas.edu/mitofy/) and BLASTN [60] searches against gene sequences of reference organelle genome with E-value of 10^{-5} and an identity threshold of 60%. We used *L. quihoui* mitogenome (GenBank: MN723864.1) and *J. sambac* cp. genomes (GenBank: MN158204.1 and MN158205.1) as reference for mitogenome and cp. genome annotation, respectively. Manual corrections of start/stop codons and exon/intron boundaries of protein-coding genes were conducted in SnapGene Viewer v7.0.2 (https://www.snapgene.com/) by referring to genes of closely related species. Four gene sequences with three atypical initiator codons were manually inspected and visualized by a self-developed python script. The code is provided on GitHub (https://github.com/HansongYan666/jasmine_genome). Transfer RNA genes (tRNAs) were predicted by tRNAscan-SE v2.0.7 [67] with default parameters, and ribosomal RNA genes (rRNAs) were identified by homologous gene evidence and transcript evidence. In the mitogenome, the three longest ncRNAs were retained by manually removing overlapping ncRNAs. Getorf attached from the EMBOSS suite (v6.6.0) [68] was employed to scan all open reading frames (ORFs) of novel genes in the entire genome with parameters: "-Tables 1 -minsize 300". Finally, the mitogenome and cp. genome maps were drawn with the online program OrganellarGenomeDRAW (OGDRAW) [69]. Mauve v0.3.0 [70] was applied to construct multiple cp. genome alignments in the presence of rearrangement. McScanX [71] with default parameters was applied to perform the syntenic comparison among three cp. genomes. Functional annotations of PCGs were carried out using sequence-similarity Blast searches with a typical cut-off E-value of 10^{-5} against five public protein databases: NCBI non-redundant (Nr) protein database, Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot, and Clusters of Orthologous Groups (COGs). The codon usage pattern and relative synonymous codon usage (RSCU) were primarily determined by cusp program in EMBOSS. The codon with an RSCU value greater than 1.0 represents preferentially used codon by amino acids, while the value equal to or less than 1.0 means randomly chosen codon or relative negative codon usage bias. The STOP codons UAA, UAG and UGA were not considered in this analysis.

The repeat structure analysis

A perl-based program MISA v2.1 (MIcroSAtellite Identification Tool) [72] was used to mine the mitogenome-wide simple sequence repeats (SSRs). Both perfect and compound repeat types were considered, with a minimum repeat length of 10 bp and a basic motif size of 2 to 6 bp. The minimum repeat length of mono-, di- tri-, tetra-, penta-, and hexanucleotides were set as 10 bp,

10 bp, 12 bp, 12 bp, 15 bp and 18 bp, respectively. In terms of long dispersed repeats, the Vmatchv2.3.1 [73] was used to determine the positions and sizes of forward (F), reverse (R), complement (C) and palindromic (P) repeats with the following criteria: a minimum repeat size of 30 bp, a seed length of 8, and a Hamming distance of 3 ("vmatch -v -l 30 -seedlength 8 -h 3 -d -p input.fa"). A Python-based pipeline for long dispersed repeat identification and statistical analysis is developed and available on GitHub: https://github.com/HansongYan666/jasmine_genome. To identify interspecific syntenic repeats, we first merged the three mitogenome sequences and then ran the pipeline. Repeat pairs originating from different varieties were classified as interspecific syntenic repeats. Interspecific BLASTN searches with an E-value of 10^{-5} were used to confirm these repeat pairs. The program nucmer from MUMmer v4.0.0 [74] was used to predict the syntenic positions and rearrangement breakpoint information among the three jasmine mitogenomes. We further determined the presence of syntenic repeats at the breakpoints based on their positions. Then we extracted breakpoint-linked syntenic repeats and visualized their positions using NGenomeSyn v1.41 [75].

Phylogenetic tree construction

The complete mitogenome sequences and their annotation files of eight Oleaceae species (14 mitogenomes in total) that are available to date were retrieved from the NCBI Genbank. These species include *J. sambac* (single-ring; GenBank: NC_069589.1), *F. suspensa* (NC_073548.1), *S. fauriei* (OR209258.1), *L. quihoui* (MN723864.1), *Hesperelaea palmeri* (NC_031323.1), *Chionanthus rupicola* (MG372115.1), *Osmanthus fragrans* (MW645067.1), and seven *O. europaea* accessions(MG372116.1-MG372121.1 and MW262896.1). Furthermore, we compiled the most comprehensive dataset ever of mitogenomes in Lamiales available as of August 15, 2023, by retrieving data from the Genbank. This dataset includes 45 mitogenomes from nine families (29 genera) within Lamiales: Oleaceae (17), Gesneriaceae (2), Plantaginaceae (3), Lentibulariaceae (2), Bignoniaceae (1), Scrophulariaceae (1), Phrymaceae (1), Orobanchaceae (7), and Lamiaceae (11) (Table S17). Single-copy orthologous protein-coding genes were identified by OrthoMCL v2.0 [76] and further aligned using MUSCLE v3.8.31 [77] with default settings. An perl script "Epal2nal.pl" was used to back-translate amino acid alignments to CDS nucleotide alignments, and is available on GitHub (https://github.com/Datapotumas/Epal2nal). Ambiguously aligned regions were then trimmed with Gblocks 0.91b [78]. The concatenated DNA codon sequence length, comprising 37 single-copy genes for each mitogenome, totaled 31,407 bp. The maximum-likelihood (ML) phylogenetic tree in Oleaceae

was further constructed by PhyML v3.0 [79], employing 1,000 bootstrap replicates and the best-fit substitution model "GTR+I+G" inferred by jModeltest v2.1.10 [80] based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Multiple sequence alignment of all 17 mitogenomes in Oleaceae was conducted on AliTV v1.0.6 [81]. The ML tree for all available Lamiales mitogenomes was also constructed by PhyML v3.0 based on the concatenated amino acid dataset comprising 19 common single-copy orthologous PCGs from 48 species. The best-fit model "LG+I+G+F" was inferred by ProtTest v3.4 [82]. Three Solanales species (NC_006581.1, NC_044153.1 and OL467322.1) served as outgroup.

Selection pressure analysis

To identify genes under positive selection in Oleaceae species, nonsynonymous (Ka) and synonymous (Ks) substitution rates of common single-copy genes were calculated. The Ka/Ks ratio (ω)>1 means the gene is subjected to positive selection, while $\omega=1$ and $\omega<1$ signify neutral selection and purifying selection, respectively. The multiple codon-based alignments of single-copy genes among all sixteen mitogenomes of seven genera in Oleaceae achieved for the ML tree construction (as described earlier) was used here to perform the positive selection analysis. The branch site-specific models were considered to perform selection pressure analysis using the subprogram CodeML from PAML v4.10.7 [83]. Two branch site models were tested: (1) Alternative model A with parameters of "model=2, NSsites=2, fix_omega=0, omega=2"; (2) Null model with parameters of "model=2, NSsites=2, fix_omega=1, omega=1". The statistical likelihood ratio test (LRT) was used to compare the alternative and null models following the formula: 2ΔL=2(L1-L0), where L1 is the alternative hypothesis and L0 is the null hypothesis. *P*-values were computed with one degree of freedom using the "chi2" program from the PAML package. If a gene has positively selected sites with a test *P*-value<0.05, we consider it as a positively selected gene.

Identification of homologous sequences transfer

The complete sequences and GenBank files for two cp. genomes of *J. sambac* "DP" (MN158205.1) and "SP" (MN158204.1) were downloaded from the NCBI Genbank. The nuclear genome assemblies for *J. sambac* "DP", "SP" and "HT" were downloaded from the National Genomics Data Center (NGDC; https://ngdc.cncb.ac.cn/) under project number PRJCA006739 and PRJCA019962, respectively. The homologous sequences from chloroplast and mitochondrial genomes to the nucleus are termed nuclear plastid sequences (**NUPTs**) and nuclear mitochondrial sequences (**NUMTs**), respectively. To predict the potential homologous DNA

transfers, BLASTN [60] searches were conducted among cp. genomes, mitogenomes and nuclear genomes with an E-value threshold of 10^{-5}. NUPTs and NUMTs were detected by alignment of chloroplast and mitochondrial sequences with the corresponding *J. sambac* nuclear genomes.

RepeatModeler v2.0.2 (http://www.repeatmasker.org/RepeatModeler/) was employed for *de novo* transposable element (TE) family identification in the three *J. sambac* nuclear genomes. This process involved integrating three *de novo* repeat-finding programs: RECON v1.08 [84], RepeatScout v1.0.6 [85], and LTR_retriever v2.9.0 [86]. The consensus results from these programs were then imported into RepeatMasker (v4.07) for the discovery and clustering of repetitive elements. Tandem repeats within nuclear genomes were identified by Tandem Repeat Finder (v4.07) [87].

To identify genotype-specific nuclear organelle junction sites, we used a bioinformatic analysis pipeline as previously described with a minor modification [88, 89]. Take "HT" vs. "DP" as an example: (1) The blast result was filtered on the basis of identity (>90%) and then was further used to generate junction regions. (2) All PacBio reads from "DP" were aligned to the assembled "HT" nuclear genome as a reference using minimap2 [57]. (3) A shared junction site between nuclear DNA (nuDNA) and nuclear organelle DNA (norgDNA) was identified when this site was spanned by at least three reads. (4) Upstream and downstream regions (+50 bp and −50 bp) of the junction sites were used as thresholds to obtain common junction sites with high confidence. This entire process was implemented using a python script called "detect_juct.py" with the blast result and the bam file as input. By following this approach, the reliable "HT"-"DP" common and "HT"-specific junction sites were determined. Similarly, this method was applied to find "DP"-specific and "SP"-specific junction sites. All these scripts were provided on GitHub (https://github.com/HansongYan666/norgDNAscripts).

Validation of the 14.2 kb NUPT insertion in "HT"

The minimap2 [57] was used to align the whole dataset of PacBio sequencing reads of "HT", "DP" and "SP" against the "HT" reference genome which contains the 14.2 kb NUPT insertion in Chr7. The mapping results were visualized using the Integrative Genomics Viewer (IGV) software [90]. We also randomly designed primer pairs that span over two border junctions of this 14.2 kb insertion in "HT" to amplify sequences on both sides of the junction site. Information regarding these primers can be found in Table S26. The PCR experiment was repeated three times to ensure the accuracy of the results. Sanger sequencing was used to confirm the PCR products.

Abbreviations

J. sambac	Jasminum sambac
SP	Single-petal Jasmine
DP	Double-petal Jasmine
MP	Multi-petal Jasmine
HT	Jasminum sambac "Hutou"
cp	Chloroplast
FISH	Fluorescence in Situ Hybridization
PCR	Polymerase Chain Reaction
mtDNA	Mitogenome/Mitochondrial Genome
IGT	Intracellular Gene Transfer
HGT	Horizontal Gene Transfer
HR	Homologous Recombination
CMS	Cytoplasmic Male Sterility
NUMT	Nuclear Mitochondrial Sequence
MTPT	Mitochondrial Plastid Transferred Fragment
PCGs	Protein-coding Genes
RSCU	Relative Synonymous Codon Usage
SD	Shine-Dalgarno
SSRs	Simple Sequence Repeats
LSU	Large Subunit
SSU	Small Subunit
LCBs	Locally Collinear Blocks
PE	Paired-End
LSC	Large Single Copy
SSC	Small Single Copy
IRs	Inverted Repeats
ACCase	Acetyl-CoA Carboxyltransferase
NUPT	Nuclear Plastid Sequence
norgDNA	Nuclear Organelle DNA
TEs	Transposable Elements
HiFi	High-Fidelity
PAF	Pairwise mApping Format
CDS	Coding Sequences
BAM	Binary Alignment/Map
IGV	Integrative Genomics Viewer
tRNAs	Transfer RNA Genes
rRNAs	Ribosomal RNA Genes
ORFs	Open Reading Frames
OGDRAW	OrganellarGenomeDRAW
Nr	NCBI Non-redundant Protein Database
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
COGs	Clusters of Orthologous Groups
ML	Maximum-likelihood
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Ka/Ks	Nonsynonymous/Synonymous Substitution Rate
LRT	Likelihood Ratio Test
nuDNA	Nuclear DNA

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05557-9.

Supplementary Material 1

Supplementary Material 2

Author contributions

J.F.: Conceptualization, Supervision, Methodology, Software, Resources, and Writing - Original draft preparation. A.L., H.Y., S.L., X.X., & L.Z.: Methodology, Software, Formal analysis, and Data curation. Y.H. & L.F.: Resources, Investigation, Validation, and Writing - Original draft preparation. R.H., P.M. & K.Z.: Conceptualization, Writing - Review and editing.

Data availability

Three newly assembled mitochondrial genome sequences with gene annotations have been deposited in the NCBI GenBank under accession numbers OR582639-OR582646. The final cp. genome sequence with gene annotation has been assigned the GenBank accession number OR588872. The PacBio HiFi whole-genome sequencing data are publicly accessible at NGDC under accession number GWHDUBQ00000000.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]College of Life Science, Fujian Normal University, Fuzhou, China
[2]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, Australia
[3]Marine and Agricultural Biotechnology Laboratory, Fuzhou Institute of Oceanography, College of Geography and Oceanography, Minjiang University, Fuzhou, China
[4]Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China

References

1. Al-Snafi AE. Pharmacological and therapeutic effects of *Jasminum sambac* - a review. Indo Am J Pharm Sci. 2018;5:1766–78.
2. Wang P, Fang J, Lin H, Yang W, Yu J, Hong Y, et al. Genomes of single-and double-petal jasmines (*Jasminum sambac*) provide insights into their divergence time and structural variations. Plant Biotechnol J. 2022;20:1232.
3. Youle RJ. Mitochondria-striking a balance between host and endosymbiont. Science. 2019;365:eaaw9855.
4. Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Rohwer FL, Mylnikov AP, Keeling PJ. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. Curr Biol. 2017;27:3717–24.
5. Burger G, Gray MW, Forget L, Lang BF. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biology Evol. 2013;5:418–38.
6. Garcia LE, Edera AA, Palmer JD, Sato H, Sanchez-Puerta MV. Horizontal gene transfers dominate the functional mitochondrial gene space of a holoparasitic plant. New Phytol. 2021;229:1701–14.
7. Wang J, Kan S, Liao X, Zhou J, Tembrock LR, Daniell H, Jin S, Wu Z. Plant organellar genomes: much done, much more to do. Trends Plant Sci. 2024;29:754–69.
8. Drouin G, Daoud H, Xia J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenetics Evol. 2008;49:827–31.
9. Christensen AC. Plant Mitochondria are a riddle wrapped in a mystery inside an Enigma. J Mol Evol. 2021;89:151–56.
10. Reboud X, Zeyl C. Organelle inheritance in plants. Heredity. 1994;72:132–40.
11. Christin P-A, Besnard G, Edwards EJ, Salamin N. Effect of genetic convergence on phylogenetic inference. Mol Phylogenetics Evol. 2012;62:921–27.
12. Van de Paer C, Bouchez O, Besnard G. Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). Mol Ecol Resour. 2018;18:407–23.

13. Wu ZQ, Liao XZ, Zhang XN, Tembrock LR, Broz A. Genomic architectural variation of plant mitochondria - a review of multichromosomal structuring. J Syst Evol. 2022;60:160–68.

14. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC. The alternative reality of plant mitochondrial DNA: one ring does not rule them all. PLoS Genet. 2019;15:e1008373.

15. Lee Y, Cho CH, Noh C, Yang JH, Park SI, Lee YM, et al. Origin of minicircular mitochondrial genomes in red algae. Nat Commun. 2023;14:3363.

16. Backert S, Nielsen BL, Börner T. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci. 1997;2:477–83.

17. 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:e1001241.

18. Gualberto JM, Newton KJ. Plant mitochondrial genomes: dynamics and mechanisms of mutation. Annu Rev Plant Biol. 2017;68:225–52.

19. Bi C, Qu Y, Hou J, Wu K, Ye N, Yin T. Deciphering the multi-chromosomal mitochondrial genome of *Populus simonii*. Front Plant Sci. 2022;13:914635.

20. Zhang F, Li W, Gao CW, Zhang D, Gao LZ. Deciphering tea tree chloroplast and mitochondrial genomes of Camellia sinensis var. assamica. Sci Data. 2019;6:209.

21. Liu H, Yu J, Yu X, Zhang D, Chang H, Li W, et al. Structural variation of mitochondrial genomes sheds light on evolutionary history of soybeans. Plant J. 2021;108:1456–72.

22. Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell. 2011;23:2499–513.

23. Sloan DB, Müller K, McCauley DE, Taylor DR, Štorchová H. Intraspecific variation in mitochondrial genome sequence, structure, and gene content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. New Phytol. 2012;196:1228–39.

24. Kmiec B, Woloszynska M, Janska H. Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Curr Genet. 2006;50:149–59.

25. Xu S, Ding Y, Sun J, Zhang Z, Wu Z, Yang T, Shen F, Xue G. A high-quality genome assembly of Jasminum sambac provides insight into floral trait formation and Oleaceae genome evolution. Mol Ecol Resour. 2022;22:724–39.

26. Zhou C, Zhu C, Tian C, Xie S, Xu K, Huang L, et al. The chromosome-scale genome assembly of *Jasminum sambac* var. Unifoliatum provides insights into the formation of floral fragrance. Hortic Plant J. 2023;9:1131–48.

27. Chen G, Mostafa S, Lu Z, Du R, Cui J, Wang Y, et al. The jasmine (*Jasminum sambac*) genome provides insight into the biosynthesis of flower fragrances and jasmonates. Genom Proteom Bioinform. 2023;21:127–49.

28. Qi X, Wang H, Liu S, Chen S, Feng J, Chen H, et al. The chromosome-level genome of double-petal phenotype jasmine (*Jasminum sambac* Aiton) provides insights into the biosynthesis of floral scent. Hortic Plant J. 2023;10:259–72.

29. Xu M, Gao Q, Jiang M, Wang W, Hu J, Chang X, et al. A novel genome sequence of *Jasminum sambac* helps uncover the molecular mechanism underlying the accumulation of jasmonates. J Exp Bot. 2023;74:1275–90.

30. Fang J, Xu X, Chen Q, Lin A, Lin S, Lei W, Zhong C, Huang Y, He Y. The complete mitochondrial genome of *Isochrysis galbana* harbors a unique repeat structure and a specific trans-spliced *cox1* gene. Front Microbiol. 2022;13:966219.

31. Clements J, Laz T, Sherman F. Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. Mol Cell Biology. 1988;8:4533–36.

32. Bock H, Brennicke A, Schuster W. *Rps3* and *rpl16* genes do not overlap in Oenothera mitochondria: GTG as a potential translation initiation codon in plant mitochondria? Plant Mol Biol. 1994;24:811–18.

33. Zitomer R, Walthall D, Rymond B, Hollenberg C. Saccharomyces cerevisiae ribosomes recognize non-AUG initiation codons. Mol Cell Biology. 1984;4:1191–97.

34. Yu X, Jiang W, Tan W, Zhang X, Tian X. Deciphering the organelle genomes and transcriptomes of a common ornamental plant *Ligustrum quihoui* reveals multiple fragments of transposable elements in the mitogenome. Int J Biol Macromol. 2020;165:1988–99.

35. Sakamoto W, Tan S-H, Murata M, Motoyoshi F. An unusual mitochondrial *atp9-rpl16* cotranscript found in the maternal distorted leaf mutant of *Arabidopsis thaliana*: implication of GUG as an initiation codon in plant mitochondria. Plant Cell Physiol. 1997;38:975–79.

36. Tran HC, Schmitt V, Lama S, Wang C, Launay-Avon A, Bernfur K, et al. An mTRAN-mRNA interaction mediates mitochondrial translation initiation in plants. Science. 2023;381:eadg0995.

37. Kang JS, Zhang HR, Wang YR, Liang SQ, Mao ZY, Zhang XC, Xiang QP. Distinctive evolutionary pattern of organelle genomes linked to the nuclear genome in Selaginellaceae. Plant J. 2020;104:1657–72.

38. 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;19:614.

39. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytol. 2010;186:299–317.

40. Woloszynska M. Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes–though this be madness, yet there's method in't. J Exp Bot. 2010;61:657–71.

41. Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proceedings of the National Academy of Sciences. 2000;97:6960-66.

42. Xu X, Huang H, Lin S, Zhou L, Yi Y, Lin E, et al. Twelve newly assembled jasmine chloroplast genomes: unveiling genomic diversity, phylogenetic relationships and evolutionary patterns among Oleaceae and *Jasminum* species. BMC Plant Biol. 2024;24:331.

43. Van de Paer C, Bouchez O, Besnard GJ. Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). Mol Ecol Resour. 2018;18:407–23.

44. Wortley AH, Rudall PJ, Harris DJ, Scotland RW. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Syst Biol. 2005;54:697–709.

45. Refulio-Rodriguez NF, Olmstead RG. Phylogeny of lamiidae. Am J Bot. 2014;101:287–99.

46. Van de Paer C, Hong-Wa C, Jeziorski C, Besnard G. Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. Gene. 2016;594:197–202.

47. Qi X, Chen S, Wang Y, Feng J, Wang H, Deng Y. Complete chloroplast genome of *Jasminum sambac* L. (Oleaceae). Brazilian J Bot. 2020;43:855–67.

48. Li J, Su Y, Wang T. The repeat sequences and elevated substitution rates of the chloroplast *accD* gene in cupressophytes. Front Plant Sci. 2018;9:533.

49. Nováková E, Zablatzká L, Brus J, Nesrstová V, Hanáček P, Kalendar R, Cvrčková F, Majeský L', Smýkal P. Allelic diversity of acetyl coenzyme a carboxylase accD/bccp genes implicated in nuclear-cytoplasmic conflict in the wild and domesticated pea (*Pisum* Sp). Int J Mol Sci. 2019;20:1773.

50. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid *accD* gene is essential and is required for leaf development. Plant J. 2005;44:237–44.

51. Madoka Y, Tomizawa K-I, Mizoi J, Nishida I, Nagano Y, Sasaki Y. Chloroplast transformation with modified *accD* operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. Plant Cell Physiol. 2002;43:1518–25.

52. Bryant N, Lloyd J, Sweeney C, Myouga F, Meinke D. Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in Arabidopsis. Plant Physiol. 2011;155:1678–89.

53. Sloan DB, Wu Z. History of plastid DNA insertions reveals weak deletion and at mutation biases in angiosperm mitochondrial genomes. Genome Biol Evol. 2014;6:3210–21.

54. Matsuo M, Ito Y, Yamauchi R, Obokata J. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. Plant Cell. 2005;17:665–75.

55. Yoshida T, Furihata HY, Kawabe A. Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species. DNA Res. 2014;21:127–40.

56. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 2004;5:123–35.

57. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

58. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–46.

59. 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;27:722–36.

60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

61. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

Fang *et al. BMC Plant Biology*   (2024) 24:861

Page 23 of 23

62. Huang Y, Ding W, Zhang M, Han J, Jing Y, Yao W, Hasterok R, Wang Z, Wang K. The formation and evolution of centromeric satellite repeats in Saccharum species. Plant J. 2021;106:616–29.

63. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

64. Jin J-J, Yu W-B, Yang J-B, Song Y, DePamphilis CW, Yi T-S, Li D-Z. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.

65. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

66. 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:W6–11.

67. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:W686–89.

68. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–77.

69. Greiner S, Lehwark P, Bock RJN. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. 2019;47:W59-W64.

70. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.

71. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49–49.

72. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017;33:2583–85.

73. Kurtz S. The Vmatch large scale sequence analysis software. Ref Type: Comput Program. 2003;412:297.

74. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

75. He W, Yang J, Jing Y, Xu L, Yu K, Fang X. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. Bioinformatics. 2023;39:btad121.

76. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

77. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–97.

78. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.

79. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.

80. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25:1253–56.

81. Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. AliTV - interactive visualization of whole genome comparisons. PeerJ Comput Sci. 2017;3:e116.

82. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21:2104–05.

83. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.

84. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.

85. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21:i351–58.

86. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176:1410–22.

87. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.

88. Fang J, Wood AM, Chen Y, Yue J, Ming R. Genomic variation between PRSV resistant transgenic SunUp and its progenitor cultivar Sunset. BMC Genomics. 2020;21:398.

89. Yue J, VanBuren R, Liu J, Fang J, Zhang X, Liao Z, et al. SunUp and Sunset genomes revealed impact of particle bombardment mediated transformation and domestication history in papaya. Nat Genet. 2022;54:715–24.

90. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

Publisher's note