Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Characterization of The Complete Chloroplast Genome and Comparative Genomic Analysis of Trema orientalis L.

Amira Ibrahim (amiranasreldeen@sci.aru.edu.eg)
Arish University https://orcid.org/0000-0002-5904-9736
Khairiah M. Alwutayd
Princess Noura Bint AbdulRahman University: Princess Nourah bint Abdulrahman University
Fatma Ahmed Safhi
Princess Nora bint Abdul Rahman University: Princess Nourah bint Abdulrahman University
Rana M. Alshegaihi
Jeddah University: University of Jeddah
Mohammed Alqurashi
Taif University College of Science
Amal Alyamani
Taif University College of Science
Salman Aloufi
Taif University
Badr Alharthi
Taif University College of Science
Eman Fayad
Taif University College of Science
Diaa bd El-Moneim
Arish University

Research Article

Keywords: Phylogenetic Analysis, Genetic diversity, Cannabaceae, Inverted repeats (IRs), Saudi Arabia

Posted Date: June 26th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3078238/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Genetic Resources and Crop Evolution on August 1st, 2023. See the published version at https://doi.org/10.1007/s10722-023-01678-6.

Abstract

This study, *T. orientalis*was used as an example to explore chloroplast genome evolution and determine Cannabaceae's phylogenetic relationship using phylogenetic analysis. Comparing six *Trema* species chloroplast genomes shows that gene order, gene content, and length are highly conserved yet dynamically evolve among species. The whole *T. orientalis* chloroplast genome is 157,134 bp long with accession number OQ871457 and includes a pair of inverted repeats (IRs) of 25,493 bp separated by a small single-copy region of 19,320 bp and a large single-copy region of 86,822 bp. The total content of GC is 36.3%. The chloroplast genome was annotated to include 129 genes, 84 of which code for proteins, 37 for tRNA, and 8 for rRNA. 127 SSRs were found, with the highest concentration in p1 (60), whose length varied from 10 to 16 bp; these areas could serve as foundational molecular markers for the *Trema* genus. The IRS repeats were found: 17 were forward repeats (F), 25 were palindromic repeats (P), and five were reverse repetitions (R). *T. orientalis* and *T. orientalis* (NC _039734.1), with 99 percent similarity, were found in the same group in a phylogenetic analysis of *Trema* species. IR scope expansion and contraction were also determined and compared with 17 related species in this family. It is the first report of the chloroplast genome of *T. orientalis* collected from Western Desert, Saudi Arabia, providing an important data reference for future investigations into genetic diversity and plant evolution. Such information based on the complete chloroplast genomes facilitates the evolution of species-specific molecular tools to discriminate *T. orientalis.*

Highlight

1. Study a detailed analysis of the entire chloroplast genome of *Trema orientalis*, offering a comprehensive understanding of its genetic structure.

2. The study compares the chloroplast genome of *Trema orientalis* with those of other plant species, facilitating a comparative analysis that sheds light on the evolutionary relationships and divergence patterns in this plant.

3. Identify the genes present in the chloroplast genome of *Trema orientalis* and analyze their arrangement. This information helps in deciphering the functional roles of different genes and their potential contributions to the plant's physiology.

4. Provides insights into the evolutionary history of *Trema orientalis*. It contributes to our understanding of the adaptive changes that have occurred within this plant species over time.

5. The characterization and comparative analysis of the *Trema orientalis* chloroplast genome have implications for conservation efforts and phylogenetic research. This information can aid in the development of conservation strategies and contribute to the taxonomy and classification of related plant species.

Introduction

The Cannabaceae family has ten genera and about 117 species (Sytsma et al. 2002; Bell et al. 2010; Byng et al. 2016). The Cannabaceae family consists primarily of woody plants. However, it does include at least one herb (*Cannabis* L.) and a few vines (*Humulus* L.). The family is found worldwide in both tropical and temperate climates. Species such as *Aphananthe* (Thunb.) Planch., *Celtis* L., and *Trema* Lour (Yang et al., 2013). *Trema orientalis* L. is a tree in the Cannabaceae family with leaves that stay green all year. The height of this tree changes depending on the weather and where it grows. There is a tendency for the leaf base to be different lengths and widths. The length can range from 2 to 20 cm, and the width from 1.2 to 7.2 cm. Even though the flowers are small, green, and not very noticeable, they are carried in dense bunches that are short and close together. The small fruits are round, dark green or purple, and turn black when ready; they are carried on very short stalks (Farzana et al., 2022). This plant has strong roots that help it stay alive during long periods of drought (Adinortey et al., 2013). Some common names for the plant are pigeon wood, hop out, charcoal tree, Indian charcoal tree, Indian nettle tree, and gunpowder tree. This tree species can be found worldwide (Orwa et al., 2009), and it can grow in different climate zones and soil types, from heavy clay to light sand (Smith, 1966). *T. orientalis* is a potentially versatile animal feed. Nonetheless, the adequate seed remains a significant obstacle for most fodder promotion attempts (Franzel et al. 2014). The seeds of *T. orientalis* are gathered from the wild, where populations have diminished in part due to the destruction of natural habitats and may also have a significant role in determining the distribution of pioneer species like *T. orientalis* (Goodale et al., 2014). Hence, stochastic alterations in the genetic integrity of the seeds of this promising fodder species in the wild are expected to occur (Schippmann et al., 2002; Nantongo and Gwali, 2018). Determining the genetic structure of *T. orientalis* can aid in developing conservation, management, and sustainable use strategies (Frankham et al., 2002; Nantongo et al., 2016; Coates et al., 2018; Nantongo et al., 2020).

In addition to being used to make paper and poles, it has also been used in traditional medicine. Almost every part of the plant is used as medicine to treat infections in tropical areas, diseases caused by worms, and lung inflammation (Nkansa-Kyerematen, 1992; Adinortey et al., 2013). Even though they are important for medicine, not much has been written about them recently (Al-Robai et al., 2022), and there aren't many genomic resources that can help improve and domesticate them. *T. orientalis* is often employed as a natural pioneer in conventional medicine to treat illnesses (Adinortey et al., 2013). Fever reduction and infection prevention are two common uses for this species. Tremetol, simiarenol, and simiarenone are important phytochemical ingredients of *T. orientalis* leaves; tremetol, swertianin, scopoletin, and numerous fatty acids and glycosides are found in the stem bark; sterols and fatty acids are found in the roots (Parvez et al., 2019). Plants, algae, and cyanobacteria use chloroplast organelles to perform photosynthesis. Chloroplasts also perform several crucial metabolic roles. Many amino acids, lipids, pigments, and vitamins are among these. Starch is stored, and sugar is biosynthesized as well. Plants can't grow or develop without the energy provided by the nitrogen cycle and sulfate reduction (Neuhaus and Emes, 2000; Bausher et al., 2006; Richardson and Schnell, 2020). The chloroplast DNA is a typical double-stranded circular genome found in higher plants (Sugiura, 1995; Odintsova and Yurina, 2006; Ruhlman and Jansen, 2014; Iram et al., 2019). One large single-copy (LSC) region, one short single-copy (SSC) region, and two inverted repeats (IR) sections make up the normal chloroplast genome (Zhou

et al., 2016). Because of their maternal inheritance, small genome size, and low mutation rate, chloroplasts' genomic information has been widely used to produce molecular markers for use in population genetics, genome evolution, phylogenetics, and constructing DNA barcoding markers (Sun et al., 2020; Guan et al. 2022; Feng et al., 2023).

As a result of their low nucleotide substitution rates, structural simplicity, and uniparental inheritance (Yang et al., 2019), chloroplast genomes are often used for species identification (Yu et al., 2021) and are excellent resources for phylogenetic investigations (Yang et al., 2019). Because of its consistent structure and wealth of genetic data may be used to investigate intricate evolutionary connections (Oldenburg and Bendich, 2016). Nonetheless, DNA barcoding uses some genes, such as rbcL and matK, to identify species positively; this gives molecular marker research hope (Hollingsworth, 2011). As chloroplast genomes carry more genetic information than gene fragments, they are used in studies of plant genetic diversity and conservation (Wariss et al., 2018). Chloroplast genome data is augmented by next-generation sequencing (NGS) technology, which assembles it swiftly and affordably (Tangphatsornruang et al. 2010). Thus, the current study sought to use next-generation sequencing technology to sequence, assemble, and describe the complete chloroplast genome sequences of a medicinal *T. orientalis* wild variant found in the Western Desert area and study the genomic relationships among *T. orientalis* and its related species. This information from the chloroplast genome paves the way for investigations into the phylogenetic evaluation, practical application, and conservation genetics of *T. orientalis*.

Materials and methods

Sample collection and DNA extraction

Leaves of *T. orientalis* were collected fresh from the ground in the Jazan region of Saudi Arabia (17°15'N 43°06'E) and then air-dried before being analyzed. The Faculty of Science at Arish University has a herbarium with preserved voucher specimens. Total genomic DNA was isolated from 2 g of dried leaves using a WizPrepTM gDNA Mini Kit (Cell/Tissue; Korea) according to the manufacturer's instructions. Once DNA damage was analyzed by electrophoresis on a 1.0% agarose gel, the quality of the DNA was assessed using a QuantusTM Fluorometer (Promega, USA).2.2. Library construction and mitogenome assembly. Using the standard protocol of the TruSeq library preparation kit, high-quality DNA extracts were broken up so that a 300-bp short-insert library could be built (Illumina, San Diego, California, USA). On the Illumina HiSeq 4000 platform, the library was sequenced in pair-end mode with 150-bp reads (Novogene, China). They were using the default option, a plant sequence from the chloroplast genome of *Trema orientalis* (Accession: NC _039734.1). The chloroplast genome was checked by running BLASTN against the non-redundant nucleotide database at NCBI (accessed November 18, 2022). The entire chloroplast genome was put together using the Novoplasty assembler. After removing the low-quality data, the high-quality clean reads were checked, and de novo assembly was done using the single-contig method (Magdy et al., 2019; Magdy and Ouyang, 2020).

Gene annotation

The online annotation tool GeSeq (Tillich et al., 2017) was used to describe the chloroplast genome as a circular molecule. The tRNA scan-SE 2.0 search server (Lowe and Chan, 2016) and their anticodon sequences and typical cloverleaf secondary structures were used to confirm that all the tRNAs were correct. The coding sequences were checked and fixed by translating them using Geneious Prime (Kearse et al., 2012). OGDRAW (version v1.2) was used to construct a map of the cpDNA for the *T. orientalis* strain (Al-Robai et al., 2022). The relative synonymous codon usage (RSCU) was analyzed with the help of the application CodonW (version 1.4.4) (http://codonw.sourceforge.net) (accessed on 22 November 20202).

Repeats identification

The REPuter tool4 (Kurtz et al., 2001) was used to identify repeat sequences, including forward, reverse, palindrome, and complement sequences. If the Hamming distance equals three, the maximum length of repeats is 30 base pairs, and the identity is greater than 90%. MISA's basic repeat setting was utilized to analyze the simple sequence repeats (SSRs), as stated by Beier et al. (2017).

Comparative analysis of the chloroplast genome

Eighteen species of *T. orientalis* were analyzed using IRscope (Amiryousefi et al., 2018) to compare and contrast their SSC, IRs, and LSC margins visually using *T. orientalis* sequence annotation data obtained from *T. orientalis* (NC _039734.1) as a reference. Ka/Ks, or the rate of nonsynonymous (Ka) to synonymous (Ks) nucleotide substitution, was also computed from the alignments using the KaKs Calculator.

Phylogeny analysis

MAFFT, version 7.427, aligned all 36 cpDNA sequences (Kuraku et al., 2013). RAxML (version 8.2.10) was used to estimate the phylogenetic trees with a maximum likelihood method, utilizing the 'GTRGAMMA' evolutionary model and a quick boot-strap algorithm with 1000 replications (Kuraku et al., 2013).

Results

Chloroplast genome characteristics and features

There are 157,134 bp in the assembled chloroplast genomes of *T. orientalis*. The Chloroplast genome includes a large-single copy (LSC) section that is 86,822 bp long, a small-single copy (SSC) region that is 19,320 bp long, and a pair of inverted repeat regions (IRa and IRb) that are 25,493 bp long (Fig. 1). Total GC concentration in the chloroplast genome was 36.30 percent, whereas GC percentages in the LSC, SSC, and IR regions were 34.0, 298.4, and 42.8 percent, respectively. This region's high GC concentration (55.3%) is due to the four rRNAs' high GC content. About 129 coding genes were found in the genome; 84 encoded proteins, 37 encoded tRNAs, and 8 encoded ribosomal RNAs. Six PCGs (rps12, rpl2, rpl23, ycf2, ndhB, rps7), seven tRNAs (trnIGAU/CAU, trnLCAA, trnVGAC, trnAUGC, trnNGUU), and four rRNAs (rrn16, rrn23, rrn4.5, rrn5) were replicated. The SSC area had 12 protein-coding genes and 1 tRNA gene, whereas the LSC had 62 protein-coding genes and 22 tRNA genes. Twenty intron-containing genes (12 PCGs and

six tRNA genes) and two intron-containing genes (clpP and ycf3) were found in the *T. orientalis* chloroplast genome. The 5' ends of rps12 were found in the LSC area, while the duplicated 3' ends were found in the IR region, indicating that trans-splicing occurred during its production. Of all introns containing the matK gene, the one in trnKUUU is the longest, with 2,603 base pairs (Table 1). There were just 11 genes in the SSC area, whereas the LSC region had 62. Not only did IRa and IRb share four rRNA genes, but the location of those genes was also interesting (Table 2).

Table 1
Annotated genes in the *T. orientalis* Chloroplast genome.

Group of Genes	Name of Gene	Category
Ribosomal RNA	rrn16[3], rrn 23[3], rrn4.5[3], rrn5[3]	Self-replication
Transfer RNA	trnH-GUG, trnK-UUU[1], trnQ-UUG, trnS-GCU, trnG-UCC[1], trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-GCC, trnfM-CAU, trnS-GGA, trnT-UGU, trnL-UAA[1], trnF-GAA, trnV-UAC[1], trnM-CAU, trnW-CCA, trnP-UGG, trnL- CAU, trnL-CAA[3], trnV-GAC[3], trnI-GAU[1,3], trnA-UGC[1,3], trnR-ACG[3], trnN-GUU[3], trnL-UAG, trnI-CAU	
Large subunit of ribosome	rpl33, rpl20, rpl36, rpl14, rpl16[1], rpl22, rpl2[1,3], rpl23[3], rpl32	
Small subunit of ribosome	rps16[1], rps2, rps14, rps4, rps18, rps12[1,3], rps11, rps8, rps3, rps19, rps7[3], rps15	
Subunits of photosystem I	psaB, psaA, psaI, psaJ, psaC	
Subunits of photosystem II	psbA, psbK, psbI, psbM, psbD, psbC, psbZ, psbJ, psbL, psbF, psbE, psbB, psbT, psbN, psbH,	
RNA polymerase subunits	rpoC2, rpoC1, rpoB, rpoA	
Subunits of ATP synthase	atpA, atpF[1], atpH, atpI, atpE, atpB	Photosynthesis
Subunits of NADH	ndhJ, ndhK, ndhC, ndhB[1,3], ndhF, ndhD, ndhE, ndhG, ndhI, ndhA[1], ndhH	
Subunits of cytochrome	petN, petA, petL, petG, petB[1], petD[1]	
Large subunit of RuBisCO	rbcL	
Subunit of acetyl-CoA	accD	Other genes
Envelope membrane protein	cemA	

Group of Genes	Name of Gene	Category
C-type cytochrome synthesis gene	ccsA	
Maturase	matK	
Protease	clpP	
Conserved open reading frames	ycf3, ycf4, ycf2[2], ycf1	Unknown Function

Table 2
The gene number of the chloroplast genome in four regions of *T. orientalis.*

Region	Number of CDS	Number of tRNA	Number of rRNA	Total
LSC	62	22	0	84
SSC	11	1	0	12
IRA	6	11	4	21
IRB	5	12	4	21

The cpDNA genome structure of its closely related species was compared and analyzed using the CGVIEW program and the annotated cpDNA genome sequence of *T. orientalis* (Fig. 2). The rRNA and tRNA coding regions were shown to have significant similarities between different *Trema* species. Also, there were minor variations in the protein-coding areas. It was noticed that the GC content of the two IR regions is noticeably higher than that of the LSC and SSC regions (Fig. 2).

Tandem repeats

In the present study, 735 tandem repeats were identified in the *T. orientalis* chloroplast genome, mostly in noncoding regions, including intergenic and introns (Table 3). The lengths of the tandem repeats ranged from 5 to 47 bp. Based on the quadripartite structure of the chloroplast genome, the most repeat sites were detected in the LSC region (549, 74.5%), followed by IR (104, 14.82%) and SSC (82, 11.15%). These repeats included 238 mononucleotides (32.38%), 45 dinucleotides (6.12%), 39 trinucleotides (5.31%), 90 tetranucleotides (12.24%), 111 pentanucleotide (15.10%), 116 hexanucleotides (15.78), 47 hepta-nucleotide (6.39%), 24 octa nucleotide (3.27%), 14 nona-nucleotide (1.90%), and 11 deca-nucleotide (1.50%). The A/T (227) mononucleotide repeats profusely existed in the T. orientalis chloroplast genome and less frequently C/G (11), and the longest repeat was one T type of 27 bp. The AT/TA motif contributed to 37 dinucleotides (82.22%), and the longest type of dinucleotides was AT type of 24 bp. The highest abundance motifs in trinucleotide repeats were AAT and TAA, AAAT motif in tetranucleotide, and

AATAA in pentanucleotide repeats. The rest motifs in other repeat classes had a similar abundance ratio which ranged from 0.27–0.14%.

Table 3
Repeated sequences in the *T. orientalis* chloroplast genome including repeat Class, repeat abundances, and percentage abundance.

Repeat Class	Repeat abundances	Percentage abundance (%)
Mononucleotide	238	32.38
Dinucleotide	45	6.12
Trinucleotide	39	5.31
Tetranucleotide	90	12.24
Pentanucleotide	111	15.10
Hexanucleotide	116	15.78
7-nucleotide	47	6.39
8-nucleotide	24	3.27
9-nucleotide	14	1.90
10-nucleotide	11	1.50
Total	735	100.00

SNPs and indels

The pair-wise sequence alignment of *T. orientalis* chloroplast genomes revealed 147 variants, including 75 SNPs and 72 InDels in protein-coding and non-coding areas (introns and intergenic regions; Tables 4,5). 32 Single nucleotide polymorphism SNPs and zero indels were found in chloroplast genome protein-coding genes, with ycf1 and rpoB genes having the most SNPs. The LSC and SSC regions have more substitutions than the IRs regions, and transversion (65.33%) outnumbers transition (34.6%), as shown in Table 4. The LSC region had the most SNPs (58), followed by the SSC (13), and each IR region was the fewest (4). This study found 72 intergenic indels, 40 deletions and 32 insertions (Table 5). T. orientalis chloroplast genomes are rich in short indels, especially 1 bp indels. Deletions favored the LSC region (29 LSC, 7 in IR, and 4 in SSc). The LSC had 25 insertions, the SSC 7, and the IR area had no insertion.

Table 4
Comparison of substitution types of SNPs in the *T. orientalis* chloroplast genome.

Region	Transition		Transversion			
	A/G	C/T	A/T	A/C	C/G	G/T
IRa	1	0	0	0	0	1
IRb	0	1	0	1	0	0
LSC	8	10	9	10	7	14
SSC	3	3	1	1	0	5
Total	12	14	10	12	7	20

Table 5
Indel markers of the *T. orientalis* chloroplast genome.

Polymorphism Type	Reference Nucleotide(s)	Variant Nucleotide(s)	Location	IGS
Insertion		ACTAT	LSC	psbI-trnS
Deletion	GGC		LSC	psbZ-trnG
Insertion		CAGGATA	LSC	trnG-trnfM
Deletion	TCTAATTGT		LSC	trnF-ndhJ
Deletion	TAAATATATAGAAAT		LSC	psaJ-rpl33
Deletion	GTAAGAA		LSC	rps12-clpP
Deletion	AAAT		LSC	clpP-psbB
Deletion	TT		IRa	trnL-ndhB
Deletion	AAAAGT		IRa	ndhF CDS
Insertion		CCT	SSC	ndhF-rpl32
Insertion		ATTTTT	SSC	rps15-ycf1
Deletion	AA		IRb	ndhB-trnL

Codon usage pattern analysis

The results from sequencing the chloroplast genome of *T. orientalis* revealed 52378 codons (Fig. 3). One codon, either AUG or UGG, encoding the amino acids methionine and tryptophan. Two to six codons were used to encode the remaining amino acids, including the sic codons for Arginine, leucine, and Serine. Codon usage was illustrated in Fig. 3. As an amino acid, serine was the most common among *T.*

orientalis and appeared 4872 times. Tryptophan, on the other hand, was the rarest of the amino acids (647). Meanwhile, AGA had the highest RSCU value (1.94) of the six codons encoding Arginine, indicating that it was the most preferred and widely used. Also, 32 codons had RSCU values above 1, 25 of them ended in A or U. Most codons with RSCU values larger than 1 had A/U as the terminal codon, while those with C/G as the terminal codon often had RSCU values less than 1. In general, this suggests that codons ending in A or U are preferred by the cpDNA gene of *T. orientalis*. The RSCU values of *T. orientalis* and five closely related species were compared. The total RSCU of all the codons used to encode a single amino acid was nearly identical. Furthermore, the RSCU values of identical codons were nearly equal in these species, suggesting that their codon usage habits were more stable and rarely changed (Tables S1, S2).

Analysis of the repeats

Tandem repeat sequences of tens of nucleotides are known as simple sequence repeat (SSR) markers. Each repeat unit of an SSR marker consists of a small number of nucleotides (often between one and six). The chloroplast gene of *T. orientalis* was analyzed, and 127 SSR sites were discovered. Ninety of the mononucleotide repetitions were A/T pairs, 33 were T repeats, and 34 were A repeats; just one was G/C. The base composition of SSRs favors AT, which is in keeping with the fact that AT is present in relatively high concentrations in the chloroplast genome. The dinucleotide sequence consisted of AT/TA repeats, appearing nine times, followed by the trinucleotide sequence ATT/TTA repeats three times, as shown in Table S3. The 105 IRSs were found; 17 were forward repeats (F), 25 were palindromic repeats (P), 5 were reverse repetitions (R), and 16 were complement repeats (C) (Table S4). IRS length was between 30 and 51 bp (Fig. 4 and Table S5). The P sequence was the longest at 149825 base pairs. Seven repeats were found in the P-type (with a length of 31 & 32 bp), while five were found in the F-type (with a length of 35 bp). There are six types of SSRs from P1 to P6; type P1 was the highest with a value of 60, followed by P4 with 14, and compared to its related species, it showed that light difference between them as shown in Table S6. Repeat type and position 1 and two for each one with their E- value were illustrated in Table S7.

IR regions characteristics

Using the chloroplast genomes of *T. orientalis* and 17 related genus from the Cannabaceae family (8 species from *Trema* and 9 *Cetlis*), we compared the junction structure to observe the change of IR borders (Fig. 3). The node genes were mostly rpl22, rps19, rpl2, ndhF, ycf1, trnh, and PsbA for all nine *Trema* species, except *T. domingense*, where ycf 1 was replicated with 1103 bp before ndhf and *T. orientalis* (157,174 bp), and rpl19, which was duplicated with 90 bp. In contrast to the *Trema* species, where the rps19 and rpl2 genes have vanished and the rps3 gene is positioned on the left (IRB/SSC) with 650 bp, Cetlis species showed variation in their IR borders. The results revealed that the genomic structure, such as gene order and number, was conserved between the four chloroplast genomes. However, some differences in the IR expansions and contractions still existed. The *T. orientale* had shorter IR regions compared with *T. orientalis*. Additionally, the length of ndhf gene in *T. orientale* was similar to *T. orientale* (157,174 bp) and 4 bp shorter than that of *T. orientalis* and *T. orientale* (157,192 bp), whereas and ycf1 gene was similar to *T. orientale* (157,174 bp) and 2 bp longer than that of *T. orientalis* and *T. orientale* (157,192 bp) as shown in Fig. 5.

KaKs analysis

The correlation was calculated using KaKs analysis. KaKs analysis was estimated for the *T. orientalis* cp genome in this research. To look for evidence of adaptive mutation, we calculated synonymous (Ks) and nonsynonymous (Ka) substitution rates (Table S8). These findings indicate that nine genes in *T. orientalis* cpDNA, psbK, petN, psbC, psaI, petG, rpl20, psbT, rpl16 and psaC, with Ka/ks ratios of > 1, were subjected to positive selection in comparisons of this species. However, the remaining genes, with Ka/Ks 1, were subjected to negative selection, which indicates a slower rate of evolution.

Phylogenetic analysis

To investigate the family tree of the investigated plant, we sampled chloroplast DNA from *T. orientalis* and ten closely related species. Alignments of all 11 cpDNAs were calculated with MAFFT, and an ML tree was established with RAxML, which implemented the GTR-model with *Arabidopsis thaliana* as an out-group. The high bootstrap values (between 90 and 100) substantially support all associations in Fig. 6. A phylogenetic tree with strong support for most branches reveals two separate clades. *T. orientalis* YXing886 forms one clade, whereas other closely related species constitute the other. With a 157,192 bp sequence, the *T. orientalis* under study is geographically near to its closest relatives (reference).

Discussion

Plant cpDNA typically contains around 120 genes, several involved in gene expression or photosynthesis (Jansen et al., 2005). There are 84 putative protein-coding genes, 46 transfer RNA (tRNA), and eight ribosomal RNA (rRNA) genes in *T. orientalis's* exact cpDNAs, which are a circular molecule of 157,134 bp and exhibit a peculiar quadripartite structure (Fig. 1, Table 2). The average GC concentration of cpDNA is 36. 3%; however, the GC content of IR regions is higher than that of LSC and SSC regions (Fig. 2). *T. orientalis* species were similar to the GC levels found in the chloroplast genomes of other angiosperm species, ranging from 36.1 to 36.9% (Zhang et al., 2018). *T. orientalis* and other members of the Cannabaceae family experienced the same phenomenon (Zhang et al., 2018). The amount of GCs in a molecule is used as a proxy for its secondary structures' stability and the local recombination rate (Meunier and Duret, 2004).

Chloroplast genomes provide rich sources of phylogenetic information, and numerous studies using chloroplast DNA sequences have been carried out during the past two decades, greatly enhancing our understanding of the evolutionary relationships among angiosperms (Jansen et al., 2007; Moore et al., 2007; Liu et al., 2020). Here we sequenced and assembled the complete chloroplast genome of *Trema orientalis* from the Western desert in Saudi Arabia, using Illumina sequencing reads derived from the whole genome. It was possible to obtain the chloroplast genome without first separating the chloroplast DNA (Eguiluz et al., 2017). It provided sufficient genetic resources for discriminating species and phylogenetic analysis of *Trema* species through a comprehensive comparison of chloroplast genome sequences from Cannabaceae. The cpDNA rearrangements may proceed further, and the species' genetic diversity may increase due to the presence of repetitive sequences. Because of their high polymorphism,

low substitution rate, and codominant nature, cpSSRs are invaluable genetic tools for answering fundamental and practical concerns in plant biology (Deng et al., 2021).

Whereas previous studies have shown between 118 and 140 SSRs and 30 and 101 long repeats in the cpDNAs of six *Trema* species (*T. orientalis*, *T. tomentosum*, *T. levigatum*, *T. sdomingense*, *T. cannabinum*, and *T. angustifolia*) (Meunier and Duret, 2004), the present study found just 127 SSRs. Most of the SSRs and long repeats found in the seven species of Euonymus were mononucleotide SSRs, whereas the complement repeats made up a smaller percentage. SSR motifs may be useful as molecular identifiers for determining species, examining population genetics, and distinguishing between individuals (Pereira et al., 2013; Pezoa et al., 2021).

Interestingly, the complete chloroplast genome of *Trema orientale* species has been previously published in the GenBank database (Accession number: MT165918.1) but recorded under the Ulmaceae family. The percentage of pairwise identity between the newly sequenced chloroplast genomes and the *T. orientale* was 99.8%. Although the high similarity percentage, there was a difference in the number of tRNA genes, 37 tRNAs in newly sequenced chloroplast genomes. In contrast, 36 tRNAs in the published chloroplast genome lacked trnK^{UUU}. According to previous studies, repetitive sequences play an important role in stabilizing and rearranging chloroplast genome sequences (Weng et al., 2014; Wang et al., 2018). In addition, it is very important to note that the majority of repeats were found in regions that weren't coding, including introns and intergenic regions, which can be taken as an indication that non-coding regions evolved faster than coding regions (Hong et al., 2017; Skuza et al., 2019). This study confirmed this result in the LSC region as this contains a large amount of intergenic sequence. Because of their analytical and highly polymorphic nature, long repetitive sequences have been used as suitable molecular markers for authentication (Choi et al., 2016), plant evolution, phylogenetics, and polymorphism studies (Williams and Miller, 2016; Park et al., 2017). A phylogenetic tree was constructed using cpDNA sequence data from 11 species in the *Trema* species. The cluster o the studied Charcoal tree was closely related to the sister species of *T. orientalis* to the rest of the family, with a bootstrap support of 99 percent.

Significant evolutionary events, such as the frequent expansion and contraction of the IR region, may be responsible for the cpDNA size change (He et al., 2017). These events cause fluxes in the LSC/IR junctions, which initiate pseudogenes, gene duplication, or the reversion of duplicated genes to a single copy. Earlier research demonstrated that the expansion and contraction of IRs alter the evolution of protein-coding genes in the Cannabaceae family (Zhang et al., 2018). The cpDNAs of 18 different species from the family Cannabaceae were compared in this study. When comparing cpDNAs from different angiosperm species, there is a high level of conservation at the LSC/SSC and IR region boundaries (Palmer 1985). Trema species differed in IR contraction and expansion from *Celtis* species. The findings add to our existing understanding of evolutionary trends in angiosperms. Plants rely on genetic variation to maintain their evolution potential to adapt to ever-changing environmental conditions (Livingston, 1996). It is believed that nucleotide substitutions and microstructural mutations, such as insertions and deletion inversions, are a major driving force in sequence evolution despite the remarkable conservation of chloroplast genomes relative to gene content (Britten et al. 2003). For natural mutations, point

mutations were more common than frameshifts (Raes and Van de Peer, 2005). As expected, more SNPs were found in the *T. orientalis* chloroplast genomes than in Indel. Most occurred in intergenic regions, consistent with the hypothesis that non-coding sequences evolved more slowly than CDS (Wu et al., 2023).

Conclusion

Finally, we evaluated repeat sequencing, codon preferences, and nucleic acid diversity after assembling the chloroplast genome of *T. orientalis* (charcoal tree), providing information for the cpDNA genome, evolution and phylogenetic relationship of related species of *Trema*. The chloroplast genome of *T. orientalis* is 157,134 bp long, and it differs from the chloroplast genomes of other *T. orientalis* species in a few base positions. The phylogenetic study shows that *T. orientalis* is closely related to *T. orientalis* reference (NC _039734.1). Intriguingly, we identified four candidate target sites in the IR, LSC, and SSC that may serve as molecular markers: rpl22, rps19, ycf1, ndhF, psbT, and rpl2. Nearly 735 tandem repeats have been identified, which can be used for population genetics studies within *Trema* species. Our results enrich the data on the chloroplast genomes of the *T. orientalis* species, lay an essential foundation for accurate molecular identification, and give insight into the evolutionary pattern of these species. These molecular markers can differentiate across *Trema* species and its related genera and provide a theoretical foundation for future research into germplasm resources and genetic breeding methods. They have also been used to examine DNA sequence variations among plant species.

Abbreviations

IRs Inverted repeats

LSC Large single-copy

SSC Short single-copy

NGS Next-generation sequencing

cpDNA Chloroplast DNA

RSCU Relative synonymous codon usage

SSRs Simple sequence repeats

SNP Single nucleotide polymorphism

Declarations

Funding

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R402), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Competing Interests: All authors declare they have no relevant financial or non-financial interests to disclose.

Author Contributions

A.A.I. and D.A. conceived and designed the experiments; A..AI. Analysed the data; A..A.I summarized the data; A.A.I. and DA wrote the manuscript; All authors revised the manuscript.

Data Availability Statement: The complete chloroplast genome sequence of *T. orientalis* was deposited at NCBI GenBank (accession number OQ871457). All data generated or analysed during this study are included in this published article [and its supplementary information files].

Acknowledgments

All authors thank Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R402), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adinortey MB, Galyuon IK, Asamoah NO (2013) *Trema orientalis* Linn. Blume: A Potential for Prospecting for Drugs for Various Uses. Pharmacogn Rev *7*, 67.
2. Al-Robai SA, Zabin SA, Ahmed AA et al (2022) Phenolic Contents, Anticancer, Antioxidant, and Antimicrobial Capacities of MeOH Extract from the Aerial Parts of Trema Orientalis Plant. Open Che 20: 666–678.
3. Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: An Online Program to Visualize the Junction Sites of Chloroplast Genomes. Bioinformatics 34:3030–3031.
4. Bausher MG, Singh ND, Lee SB (2006) The Complete Chloroplast Genome Sequence of *Citrus Sinensis* (L.) Osbeck Var'Ridge Pineapple': Organization and Phylogenetic Relationships to Other Angiosperms. BMC Plant Biol 6:1–11.
5. Beier S, Thiel T, Münch T et al (2017) MISA-Web: A Web Server for Microsatellite Prediction. Bioinformatics 33: 2583–2585.
6. Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296–1303.
7. Britten RJ, Rowen L, Williams J (2003) Majority of Divergence between Closely Related DNA Samples Is Due to Indels. PNAS 100:4661–4665.

8. Byng JW, Chase MW, Christenhusz MJM (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20.
9. Choi KS, Chung MG, Park S (2016) The Complete Chloroplast Genome Sequences of Three *Veroniceae* Species (Plantaginaceae): Comparative Analysis and Highly Divergent Regions. Front Plant Sci 7: 355.
10. Coates DJ, Byrne M, Moritz C (2018) Genetic Diversity and Conservation Units: Dealing With the Species Population Continuum in the Age of Genomics. Front Ecol Evol 6(165).
11. Deng G, Yang M, Zhao K, Yang Y, Huang X, Cheng X (2021) The complete chloroplast genome of *Cannabis sativa* variety Yunma 7. Mitochondrial DNA B Resour. 6(2):531-532.
12. Eguiluz M, Rodrigues NF, Guzman F et al (2017) The Chloroplast Genome Sequence from Eugenia Uniflora, a Myrtaceae from Neotropics. Plant Syst Evol 303:1199–1212.
13. Farzana M, Rahman MM, Ferdous T (2022) Review on *Trema orientalis* as a Potential Bioresource in Tropical Countries. Trees 36: 1169–1177.
14. Feng J, Xiong Y, Su X et al (2023) Analysis of Complete Chloroplast Genome: Structure, Phylogenetic Relationships of *Galega orientalis* and Evolutionary Inference of Galegeae. Genes 14: 176.
15. Frankham R, Ballou, SEJD, Briscoe DA et al (2002) Introduction to conservation genetics. Cambridge university press
16. Franzel S, Carsan S, Lukuyu B et al (2014) Fodder trees for improving livestock productivity and smallholder livelihoods in Africa. Curr Opin Environ Sustain 6:98–1.
17. Goodale UM, Berlyn GP, Gregoire TG et al (2014) Differences in survival and growth among tropical rain forest pioneer tree seedlings in relation to canopy openness and herbivory. Biotro pica 46(2):183–93.
18. Guan YH, Liu WW, Duan BZ (2022) The first complete chloroplast genome of *Vicatia thibetica* de Boiss.: genome features, comparative analysis, and phylogenetic relationships. Physiol Mol Biol Plants 28:439–454.
19. He L, Qian J, Li X et al (2017) Complete Chloroplast Genome of Medicinal Plant Lonicera Japonica: Genome Rearrangement, Intron Gain and Loss, and Implications for Phylogenetic Studies. Molecules 22: 249.
20. Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad SciUSA 108:19451–19452.
21. Hong SY, Cheon KS, Yoo KO (2017) Complete Chloroplast Genome Sequences and Comparative Analysis of *Chenopodium Quinoa* and C. *Album*. Front Plant Sci 8:1696.
22. Iram S, Hayat MQ, Tahir M (2019) Chloroplast Genome Sequence of Artemisia Scoparia: Comparative Analyses and Screening of Mutational Hotspots. Plants 8: 476.
23. Jansen RK, Cai Z, Raubeson LA et al (2007) Analysis of 81 Genes from 64 Plastid Genomes Resolves Relationships in Angio-sperms and Identifies Genome-Scale Evolutionary Patterns. PNAS 104: 19369–19374.

24. Jansen RK, Raubeson LA, Boore JL et al (2005) Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. Methods Enzymol 395: 348–384.
25. Richardson LGL, Schnell DJ (2020) Origins, function, and regulation of the TOC-TIC general protein import machinery of plastids. J Exp Bot 71(4):1226-1238.
26. Kearse M, Moir= R, Wilson A et al (2012) Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. Bioinformatics 28: 1647–1649.
27. Kuraku S, Zmasek CM, Nishimura Oet al (2013) A Leaves Facilitates On-Demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity. Nucleic Acids Res. 41:W22–W28.
28. Kurtz S, Choudhuri JV, Ohlebusch E et al (2001) REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale. Nucleic Acids Res 29: 4633–4642.
29. Liu H, Hu H, Zhang S et al (2020) The complete chloroplast genome of the rare species *Epimedium tianmenshanensis* and comparative analysis with related species. Physiol Mol Biol Plants 26:2075–2083.
30. Livingston K (1996) Conservation Genetics. Case Histories from Nature. John C. Avise and James L. Hamrick, Eds. Chapman and Hall, New York, Xx, 512 Pp., Illus. 39.
31. Lowe TM, Chan PP (2016) TRNAscan-SE On-Line: Integrating Search and Context for Analysis of Transfer RNA Genes. Nucleic Acids Res 44:W54–W57.
32. Magdy M, Ou L, Yu H et al (2019) Pan-Plastome Approach Empowers the Assessment of Genetic Variation in Cultivated Capsicum Species. Hortic Res6:108.
33. Magdy M, Ouyang B (2020) The Complete Mitochondrial Genome of the Chiltepin Pepper (*Capsicum Annuum* Var. Glabriusculum), the Wild Progenitor of *Capsicum Annuum* L. Mitochondrial DNA Part B 5:683–684.
34. Meunier J, Duret L(2004) Recombination Drives the Evolution of GC-Content in the Human Genome. Mol Biol Evol 21:984–990.
35. Moore MJ, Bel CD, Soltis PS et al (2007) Using Plastid Genome-Scale Data to Resolve Enigmatic Relationships among Basal Angiosperms. PNAS 104: 19363–19368.
36. Nantongo JS, Eilu G, Geburek T et al (2016) Detection of self-incompatibility genotypes in *Prunus africana*: Characterization, evolution and spatial analysis. Plos one 11(6):e0155638.
37. Nantongo JS, Gwali S (2018) Long-term viability of populations of *Prunus africana* ((hook. f.) kalm.) in Mabira forest: implications for in situ conservation. Afr J Ecol 56(1):136–9.
38. Nantongo JS, Potts BM, Hugh F et al (2020) Quantitative Genetic Variation in Bark Stripping of *Pinus radiata*. Forests 11(12):1356.
39. Neuhaus H, Emes M (2000) Nonphotosynthetic Metabolism in Plastids. Annu Rev Plant Biol 51: 111.
40. Nkansa-Kyerematen K (1992) Ghana Herbal Pharmacopoeia.; Policy Research and Strategic Planning Institute (PORSPI): Accra.

41. Odintsova MS, Yurina NP (2006) Chloroplast Genomics of Land Plants and Algae. In Biotechnological Applications of Photosynthetic Proteins: Biochips, Biosensors and Biodevices; Biotechnology Intelligence Unit; Springer US: Boston, MA, pp. 57–72.
42. Oldenburg DJ, Bendich AJ (2016) The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. Curr Genet 62:431–442.
43. Orwa C, Mutua A, Kindt R (2009) Agroforestree Database: A Tree Reference and Selection Guide. Version 4. Agroforestree Database: a tree reference and selection guide.
44. Palmer JD (1985) Comparative Organization of Chloroplast Genomes. Annu Rev Genet 19: 325–354.
45. Park I, Yang S, Choi et al (2017) The Complete Chloroplast Genome Sequences of *Aconitum pseudolaeve* and *Aconitum longecassidatum*, and Development of Molecular Markers for Distinguishing Species in the *Aconitum* Subgenus *Lycoctonum*. Molecules 22:2012.
46. Parvez A, Azad AK, Islam MZ et al (2019) A Phytochemical and Pharmacological Review on *Trema oreintalis*: A Potential Medicinal Plant. Pharmacologyonline 3:103-119.
47. Pereira GS, Nunes ES, Laperuta LDC et al (2013) Molecular Polymorphism and Linkage Analysis in Sweet Passion Fruit, an Outcrossing Species. Ann Appl Biol 162: 347–361.
48. Pezoa I, Villacreses J, Rubilar M et al (2021) Generation of Chloroplast Molecular Markers to Differentiate Sophora toromiro and Its Hybrids as a First Approach to Its Reintroduction in Rapa Nui (Easter Island). Plants 10: 342.
49. Raes J, Van de Peer Y (2005) Functional Divergence of Proteins through Frameshift Mutations. Trends Genet 21:428–431.
50. Ruhlman TA, Jansen RK (2014) The Plastid Genomes of Flowering Plants. In Chloroplast Biotechnology; Maliga, P., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJVol. 1132, pp. 3–38.
51. Schippmann U, Leaman DJ Cunningham A (2002) Impact of cultivation and gathering of medicinal plants on biodiversity: global trends and issues. Biodiveristy and the ecosystem approach in agriculture, forestry and fishries Satellite event on the occasion of the Ninth regular session of the commission on genetic resources for food and agriculture Rome. Inter departmental working group on bio logical diversity for food and agriculture, Rome. FAO2002. 12–13 October 2002.
52. Skuza L, Szućko I, Filip E (2019) Genetic Diversity and Relationship between Cultivated, Weedy and Wild Rye Species as Revealed by Chloroplast and Mitochondrial DNA Non-Coding Regions Analysis. PloS one 14:e0213023.
53. Smith C (1966) Common Names of South African Plants. Pretoria: Department of Agricultural Technical Services. Botanical Research Institute. Botanical Survey Memoir 35.
54. Sugiura M (1995) The Chloroplast Genome. Essays Biochem 30: 49–57.
55. Sun JL, Han Y, Cui XM (2020) Development and Application of Chloroplast Molecular Markers in Panax Notoginseng. Zhong Yao Cai45*:* 1342–1349.

56. Sytsma KJ, Morawetz J, Pires JC (2002) Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on *rbcL, trnL-trnF*, and *ndhF* sequences. Am J Bot 89:1531–1546.
57. Tangphatsornruang S, Sangsrakru D, Chanprasert J et al (2010) The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17:11–22.
58. Tillich M, Lehwark P, Pellizzer T et al (2017) GeSeq – Versatile and Accurate Annotation of Organelle Genomes. Nucleic Acids Res 45:W6–W11.
59. Wang X, Zhou T, Bai G (2018) Complete Chloroplast Genome Sequence of Fagopyrum *Dibotrys*: Genome Features, Comparative Analysis and Phylogenetic Relationships. Sci Rep 8: 1–12.
60. Wariss HM, Yi TS, Wang H et al (2018) The chloroplast genome of a rare and an endangered species *Salweenia bouffordiana* (Leguminosae) in China. Conserv Genet Resour 10:405–407
61. Weng ML, Blazier JC, Govindu M (2014) Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates. Mol Biol Evol 31: 645–659.
62. Williams AV, Miller JT, Small I (2016) Integration of Complete Chloroplast Genome Sequences with Small Amplicon Datasets Improves Phylogenetic Resolution in *Acacia*. Mol Phylogenet Evol 96: 1–8.
63. Wu ML, Yan RR, Xu X et al (2023) Characterization of the Plastid Genome of the Vulnerable Endemic *Indosasa lipoensis* and Phylogenetic Analysis. Diversity 15: 197.
64. Yang MQ, van Velzen R, Bakker TF (2013) Molecular phylogenetics and character evolution of Cannabaceae. Taxon 62:473–485.
65. Yang Z, Wang G, Ma Q et al (2019) The complete chloroplast genomes of three Betulaceae species: implications for molecular phylogeny and historical biogeography. PeerJ7:e6320.
66. Yu X, Wang W, Yang H et al (2021). Transcriptome and comparative chloroplast genome analysis of *Vincetoxicum versicolor*: insights into molecular evolution and phylogenetic implication. Front Genet 12:602528.
67. Zhang H, Jin J, Moore MJ et al (2018) Plastome characteristics of Cannabaceae. Plant Divers 23 40(3):127-137.
68. Zhou T, Zhao J, Chen C (2016) Characterization of the Complete Chloroplast Genome Sequence of *Primula Veris* (Ericales: Primulaceae). Conserv Genet Resour 8:455–458.

Figures

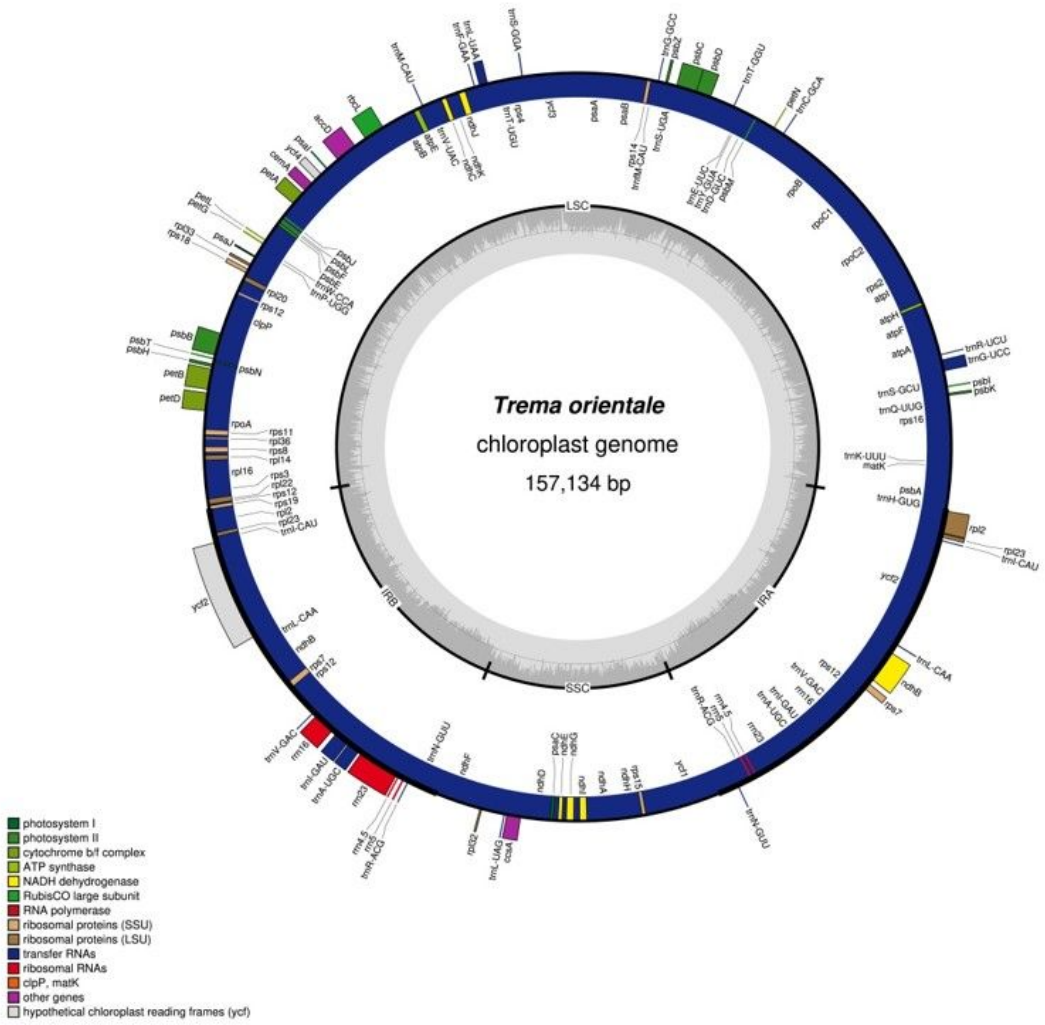


Figure 1

*T. orientalis*chloroplast gene map; the genes inside and outside the outer circle are transcribed clockwise and anticlockwise. . The thick lines show the inverted repeats (IRa and IRb) that divide the genomes into LSC and SSC regions. Grey bars inside the circle represent GC content, while lighter grey indicates AT content.

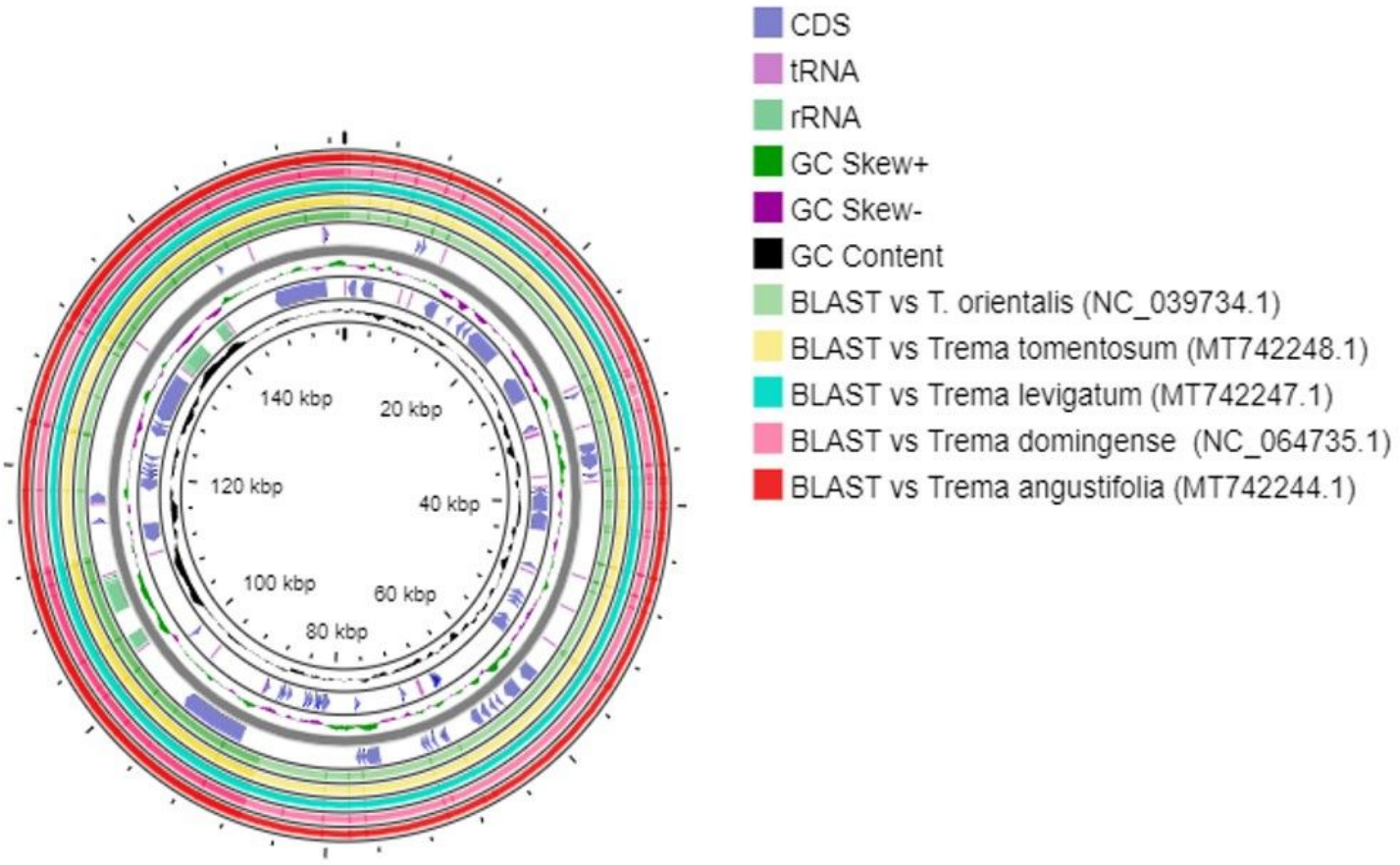


Figure 2

GC content of *T. orientalis* Chloroplast DNA; the outer five circles demonstrate similarities to related species. The inner circles indicate the CDSs of protein-coding genes (blue), rRNA operon (light purple), and tRNA information (reddish brown); the third circle shows GC contents; the fourth circle represents GC skew (green, >0; purple, <0); and the fifth circle represents cpDNA locations.

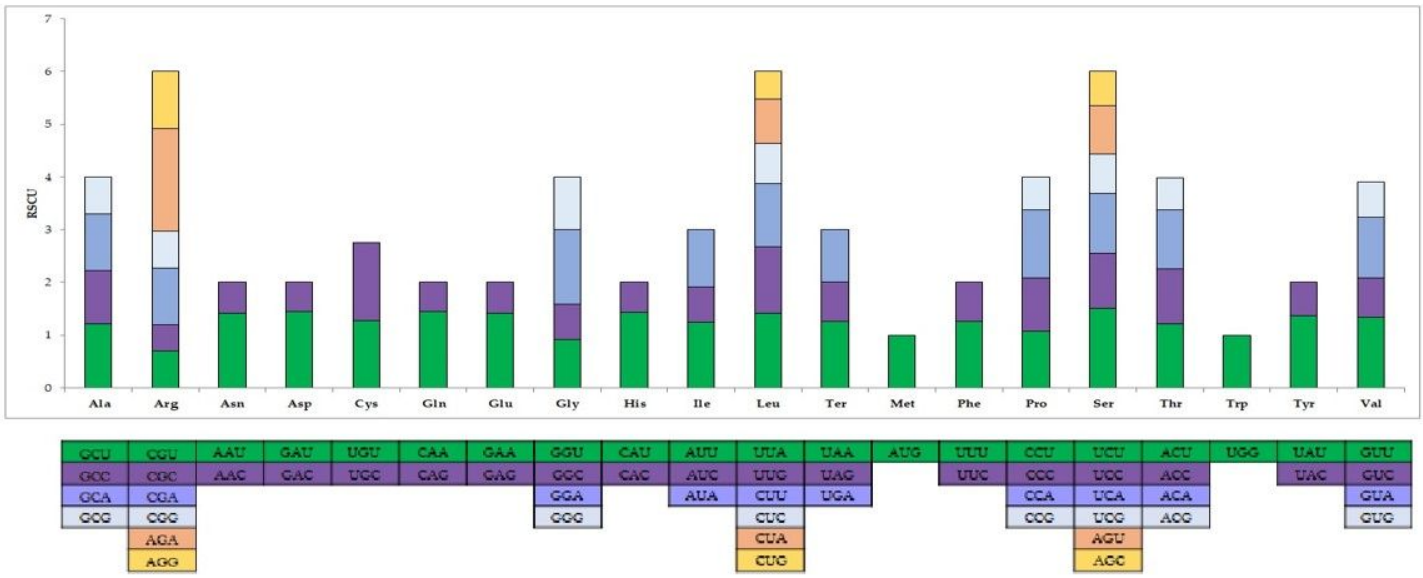


Figure 3

Relative synonymous codon usage (RSCU) values for amino acids of *T. orientalis.* The colors of the histograms correspond to the colors of the codons.

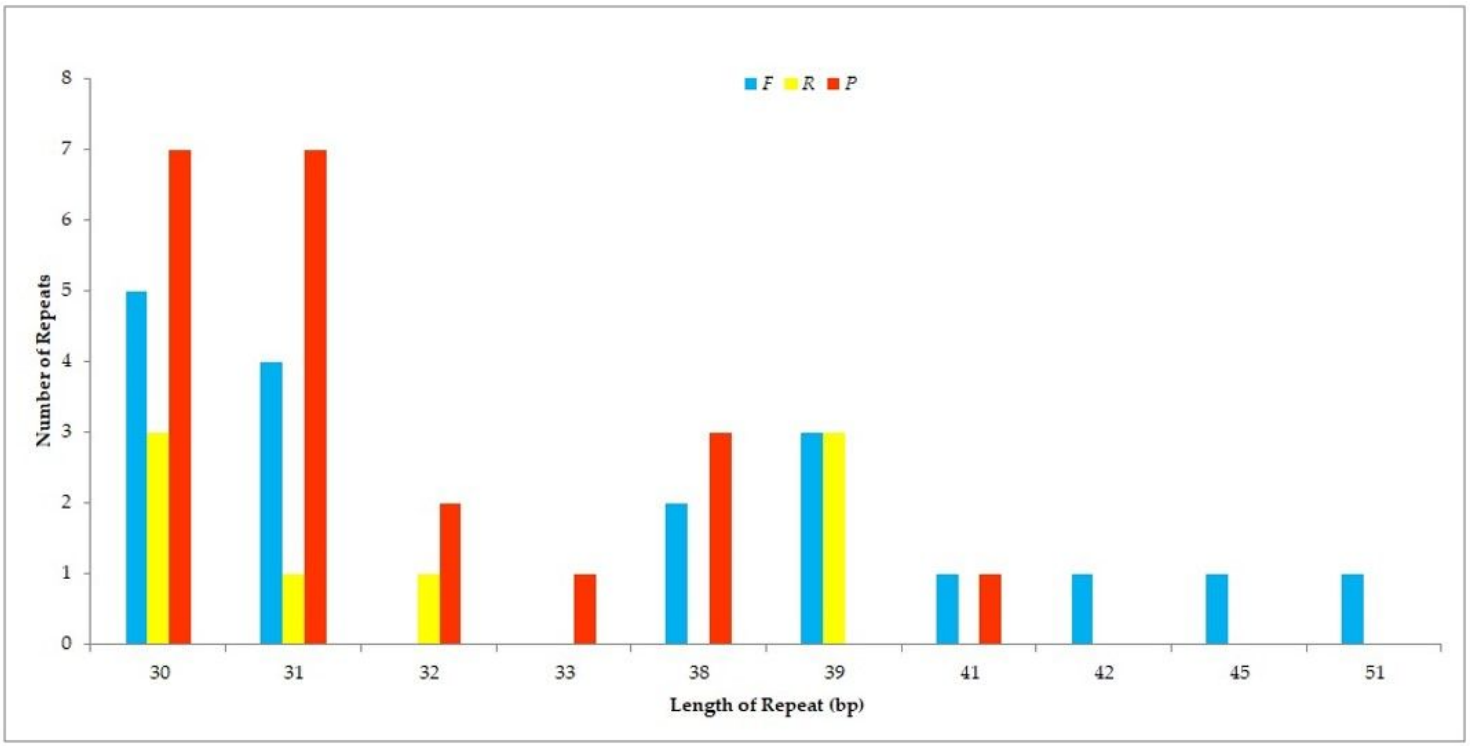


Figure 4

Analysis of repeat sequences in chloroplast genome of *T. orientalis*; Size and dispersion of the IRS. Length is an indicator of how many times this pattern will be repeated. The repetition sequence type is indicated by the variable type. In this case, the number of each type accurately reflects the total number

of types. The X-axis depicts the various categories of distributed repeats, while the Y-axis indicates the total number of such elements. F for forward, P for palindrome,and R for reverse

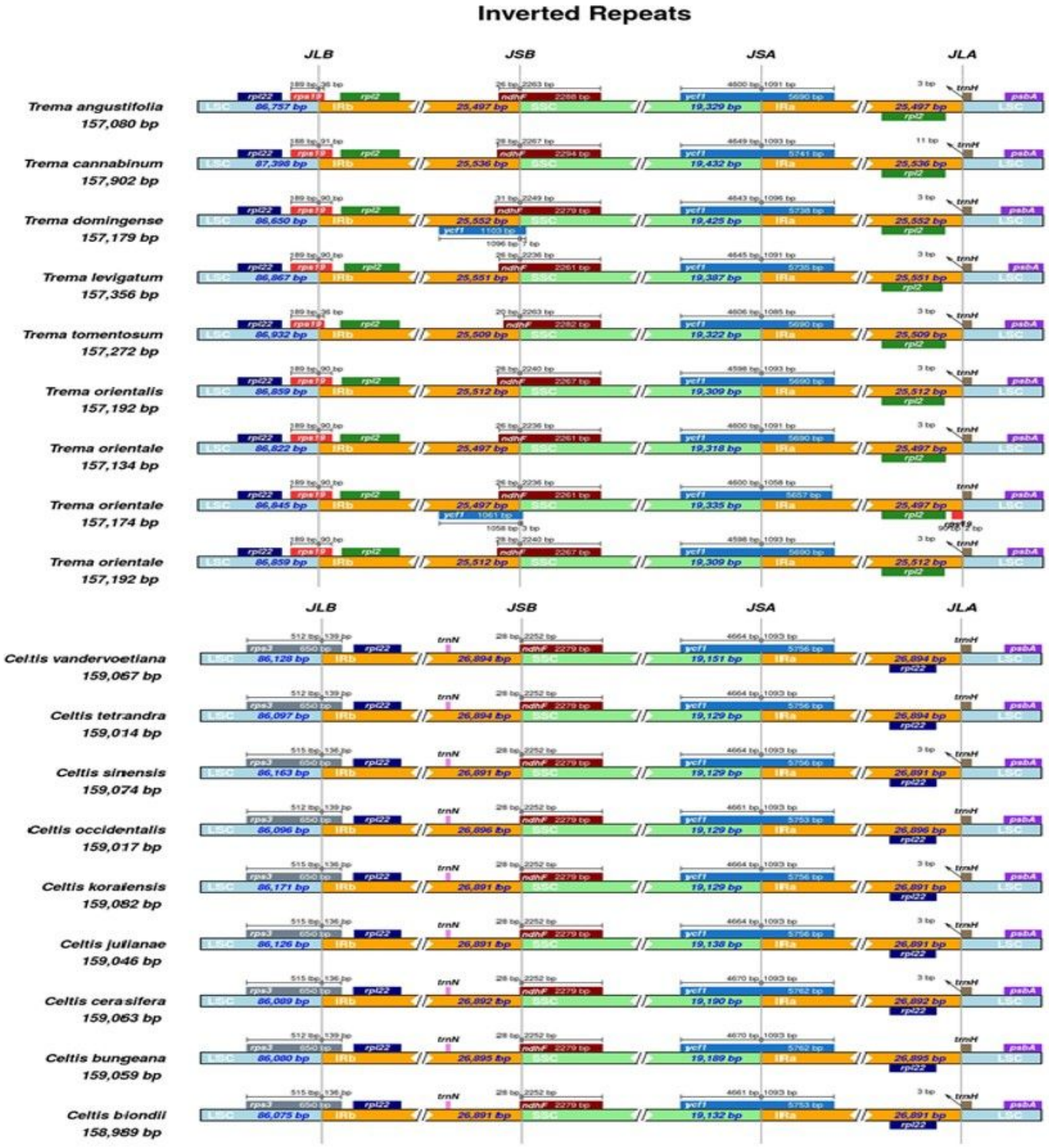


Figure 5

Boundary analysis of LSC, SSC, IRa-IRb regions of chloroplast genomes of eighteen species of *T. orientalis* and its related species from family Cannabaceae.

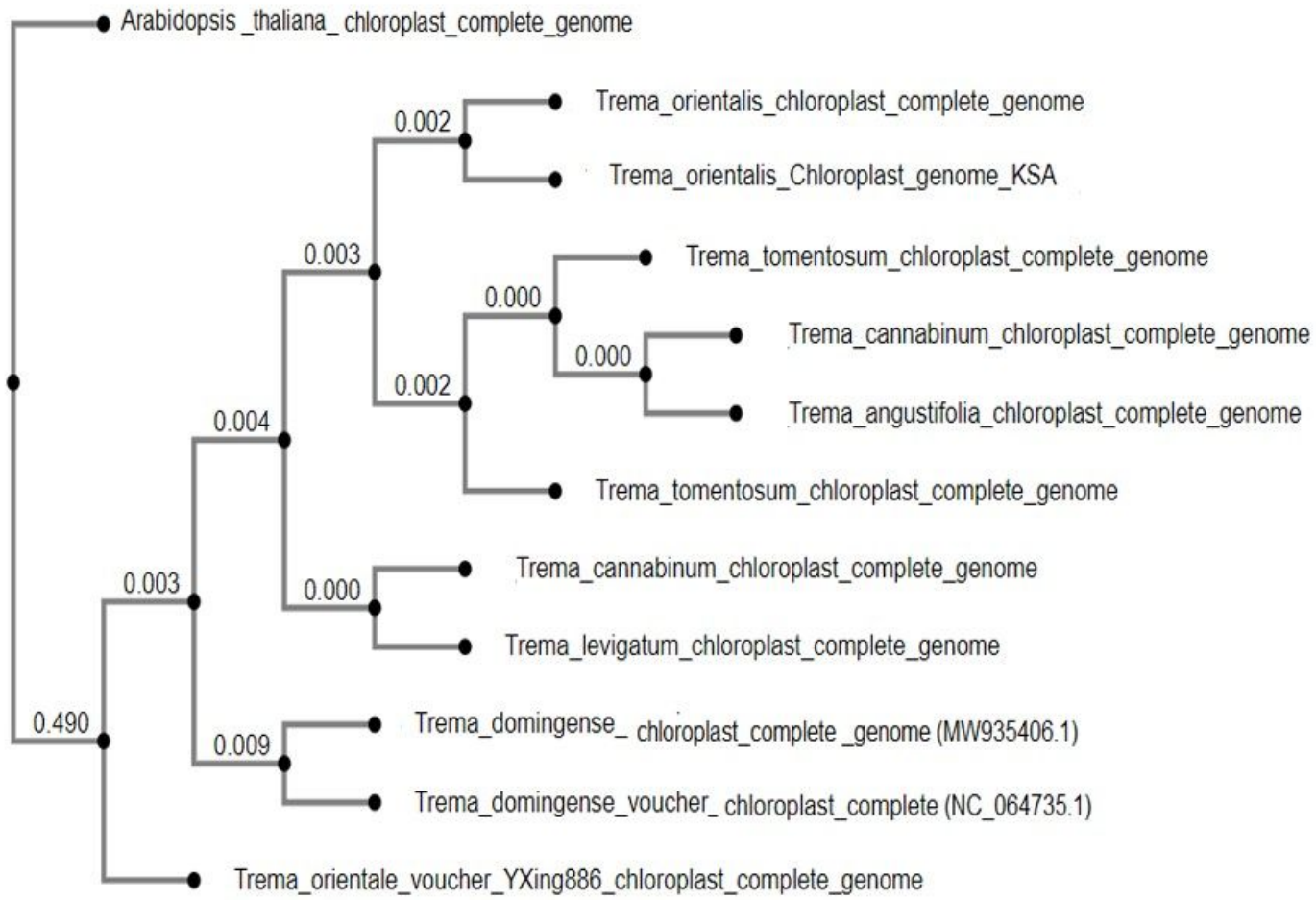


Figure 6

Phylogenetic tree of 12 complete cpDNAs constructed using the maximum likelihood (ML).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary1.docx