The Chloroplast Genomes Comparative Analysis of Taihangia Rupestris and Taihangia Rupestris Var. Ciliate, Two Endangered and Endemic Cliff Plants in Taihang Mountain of China

Yan Zheng
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Yuan Jiang
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Yujing Miao
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Zhan Feng
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Min Zhang
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Xinke Zhang
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Tiexin Zeng
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Linfang Huang (✉ lfhuang@implad.ac.cn)
State Administration of Traditional Chinese Medicine of the People's Republic of China, Chinese Academy of Medical Sciences & Peking Union Medical College

Research Article

Keywords: Taihangia rupestris, Taihangia rupestris var. ciliate, Chloroplast genome, Bioprospecting, Comparative analysis

Posted Date: October 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-892423/v1

Abstract

The *Taihangia* is a native endangered cliff species that grows in the Taihang Mountains in China. The cp genomes with a whole length of 155,558 bp and 155,479 bp for *Taihangia rupestris* and *Taihangia rupestris* var. *rupestris*. They have 131 genes in total, covering 79 protein-coding genes, 29 tRNA, and 4 rRNA. Analyses of codon usage, RNA-editing sites, repeat sequences, and comparison of cp genomes showed a high degree of conservation. Phylogenetic analysis indicated that the *Taihangia* are closed to the *Geum*. *Taihangia* genus was inferred to have originated at 0.2057 Mya, and *Geum rupestre* was inferred to have originated at 1.4431 Mya. Overall, the gene contents, gene arrangements, the types, and frequency of codon usage, repeat sequences, and SSRs are similar and highly conserved in the species of *T. rupestris* and *T. rupestris* var. *ciliate*. It is found that based on bioprospecting, *T. rupestris* and *T. rupestris* var. *rupestris* are potential medicinal resources. This study provides a scientific basis for the conservation and sustainable use of endangered medicinal resources..

1. Background

Taihangia genus, is an endangered and endemic genus of the Taihang Mountains, China [1, 2]. According to the shape and base of the leaf, the type of serrated edge, the margin, and the presence or absence of hair on the petiole, *Taihangia* genus was divided into two subspecies, namely *Taihangia rupestris* Yu et Li and *Taihangia rupestris* var. *rupestris* Yu and Li [3, 4]. *T. rupestris* and *T. rupestris* var. *rupestris* is now at risk of extinction in the wild and has been classified as a national second-class protected plant in China. These species are often rooted in small cracks in the surface of cliffs at an altitude of 600 to 1500 m above sea level, especially on U-shaped vertical cliffs, which cannot accept direct sunlight [5–7]. *T. rupestris* is also a valuable medicinal plant whose leaves are rich in flavonoids and are commonly used to treat tinea [8]. Our previous study [9] only reported that the neighbor-joining tree based on the *psbA-trnH* sequence can discriminate the two subspecies of *T. rupestris*, and *T. rupestris* var. *rupestris*. The cp genomes of *T. rupestris* and *T. rupestris* var. *rupestris* are used to further search for molecular markers, reconstruct phylogeny, and also provide theoretical basis for bioprospecting.

Chloroplasts are the key organelles of green plants, which participate in the process of photosynthesis and provide plants with the necessary energy [10]. The chloroplast genome (plastome) is a double-stranded molecule of 115 to 165 kb in most plants [11]. The chloroplast genome is usually a tetrad in structure, containing a single large copy (LSC) and a single small copy (SSC) separated by a pair of the reverse repeat (IR) [12, 13]. Genome organization, it is content and gene structure are highly conservative [14]. Because of its conservatism, cp genome content is widely used by researchers as a tool to study phylogenetic relationships and genome research [15].

We analyzed the chloroplast genome characteristics of *Taihangia* cp genomes, conducted a comparative analysis between the 2 cp genomes. This study aimed to phylogenetically reconstruct order *Taihangia* and determine the taxonomic scheme, and further the current understanding of the evolution of the *Taihangia* genus.

2. Materials And Methods

2.1 Plant material and DNA extraction

Tender leaves were collected from adult *T. rupestris* and *T. rupestris* var. *ciliate* located in China (Guoliang Village, Xinxiang City, Henan Province, 35°43' N, 113°36'E). Two newly sequenced species were sampled with the permission of the Forestry Bureau of Wu'an City, Henan Province and the Chinese Academy of Medical Science & Peking Union Medicinal College. The samples were identified by Professor Yulin Lin, whose voucher specimens are CMPB13401 and CMPB13402 deposited in the Herbarium of the Chinese Academy of Medical Science & Peking Union Medicinal College. Tender leaves were sampled and frozen by nitrogen, transferred to the lab by drikold, and stored in an ultra-cold storage freezer (-80°C) for the usage of DNA extraction. CTAB method was performed to extract total genomic DNA from young leaves [16].

2.2 Library construction, sequencing, assembly, and annotation

The Illumina Hiseq 2500 platform (Novogene Technologies, Inc., Beijing, China) was applied to sequence the Genomic DNA. PRINSEQ lite Ver0.20.4 was performed to filter the raw data reads to get clean reads [17]. The chloroplast genomes were assembled from the highest quality clean reads by using NOVOPlasty (v.2.7.2) [18] with kmer 39 using the chloroplast genome of *Rubus amabilis* (NC_047211.1) as reference and *rbcL* as a seed.

Annotations were performed in CPGVAS (http://47.96.249.172:16019/analyzer/home) [19] using the *Rubus amabilis* (NC_047211.1) as a reference chloroplast genome. First, CPGVASAS annotation results were utilized to obtain the GFF3 format file. Apollo Genome Annotation and Curation Tool (v1.11.8) to manually correct the abnormal features based on the reference database of CpGAVAS and the tRNA genes annotated by tRNAscan-SE. Last, OrganellarGenomeDRAW (https://chlorobox.mpimp-golm.mpg.de/geseq.html) [20] was used to directly generate a corrected cp circular map [13]. The complete cp genome sequence of *T. rupestris* and *T. rupestris* var. *ciliate* and their gene annotations were submitted to GenBank (MZ151697 and MZ151698).

2.3 Codon usage, RNA editing sites, and repeat sequences

The distribution of codon usage was investigated using CodonW software with the RSCU ratio [21]. PREP suite [22] with a cutoff value of 8.0 was used to predict the RNA editing sites in the plastomes. The online REPuter database [23] was used to identify repeat sequences, including the forward (F), palindromic (P), reverse (R), and complementary repeats (C). According to the following parameters: (1) a repeat size over 30 bp; (2) more than 90% sequence identity between two replicates; and (3) Hamming distance = 3. All overlapping repeat sequences were removed. The Perl script MISA (http://pgrc.ipk-gatersleben.de/misa/) [24] was used to exploit simple sequence repeats (SSRs). The minimum number of SSRs was set to 10, 5, 4, 3, 3, and 3, for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively.

2.4 Phylogenetic analysis and estimation of divergence time

A total of 29 cp whole-genome sequences were used in cluster analysis, 27 of which belong to the Rosoideae Focke, twenty-seven plastome sequences were downloaded from the NCBI. *Ulmus parvifolia* (NC_049883) and *Barbeya oleoides* (NC_040984) genomes were included as outgroups.

The MAFFT v7.221 [25, 26] was applied to extract and aligned coding sequences and a total of 95 coding gene sequences were presented in all of the 29 species, and the RAxML v8.2.8 [27] was used to construct the Maximum-Likelihood (ML) phylogenetic tree, bootstrap probability values were calculated from 1000 replicates and GTRGAMMA model as suggested. The MEGA X (Version 10.2.6) [28] was used to estimate divergence times.

To estimate the species divergence time, the Bayesian method implemented in BEAST (version 1.10.1) with GTR+GAMMA substitution model was applied to analyze the molecular clock of *T. rupestris* and *T. rupestris* var. *ciliate*. Two fossil datasets, genus Rosa (55.8-48.6 mya) [29] and genus Oligocene fossil of Potentilla sp. date to (33.9–23.0 Ma) [30] were used to calibrate the nodes. The BEAST MCMC simulations were run for 10,000,000 generations (Whidden and Matsen, 2015). TreeAnnotator (version v1.6.1) software was used to annotate the phylogenetic results generated by BEAST and the FigTree (version v1.3.1) was used to visualize the BEAST maximum clade credibility (MCC) tree.

2.5 Genomes comparison

Nucleotide diversity 172 (π) was calculated by sliding window analysis conducted in DnaSP v.6 [31]. The step size was set to 200 bp and the window is 600 bp. Depending on the phylogenetic Analysis, four species of Trib. Colurieae, *T. rupestris*, *T. rupestris* var. *ciliate*, *Geum rupestre* (NC_037392.1), and *Geum macrophyllum* (NC_053765) were selected to perform the IR expansions and contractions analysis. The Kimura two-parameter distance model [32] was used to determine pairwise sequence divergence, which was calculated using the distmat program from the EMBOSS package [33].

2.6 Non-synonymous (Ka) and synonymous (Ks) substitution rate analysis

The non-synonymous (Ka) and synonymous (Ks) substitution ratio (Ka/Ks) of each gene was calculated in the background of different evolutionary clades by hyphy. *Ulmus parvifolia* (NC_049883) and *Barbeya oleoides* (NC_040984) genomes were set as the outgroup. For each protein-coding gene, the relative substitution rates between outgroup species and each ingroup species were assessed. To calculate the Ka/Ks ratio, protein-coding sequences for all genes of each species pair were aligned using MAFFT v7.480 [34]. The phylogenetic tree was constructed by RAxML [35]. The Hyphy was used to calculate the Ka, Ks, and the Ka/Ks ratio.

3. Results And Discussion

3.1 Chloroplast genomes features

The whole cp genome of *T. rupestris* and *T. rupestris* var. *ciliate* respectively had the length of 155,558 bp and 155,479 bp. *T. rupestris* and *T. rupestris* var. *ciliate* (Figure 2). *T. rupestris* and *T. rupestris* var. *ciliate* cp genomes display a typical quadripartite circular structure containing one large single copy (LSC), one small single copy (SSC), and two inverted repeats (IRB and IRA) regions. In *T. rupestris*, an LSC region of 18,543 bp and an SSC region of 85,857 bp were separated by a pair of IR regions of 25,579 bp. The overall GC content of the *T. rupestris* cp genome was 36.79%, and the GC content of the SSC and LSC regions was 30.80% and 34.51%, respectively. Because each IR region is relatively rich in GC-rich ribosomal RNA (rRNA) gene and transfer RNA (tRNA) gene, the GC content of the IR region was 42.80%, which was much higher than that of the LSC and SSC regions. For *T. rupestris* var. *ciliate*, the SSC region is 85,820 bp, the LSC region is 18,499 bp, and the IR region is 25,580 bp. The GC content of the above regions is 34.50%, 30.94%, and 42.79%, and the GC content of the complete cp genome sequences is 36.80% (Table 1).

Table 1

Summary of *T. rupestris* and *T. rupestris* var. *ciliata* chloroplast genome features.

Species	Regions	T(U)/%	A/%	C/%	G/%	GC/%	Length (bp)	Number of protein-coding genes	Number of tRNA genes	Number of rRNA genes
T. rupestris	Total	31.95	31.25	18.78	18.02	36.79	155,558	84	36	8
	IRA	28.57	28.63	22.16	20.64	42.80	25,579			
	IRB	28.63	28.57	20.64	22.16	42.80	25,579			
	SSC	33.38	32.11	17.79	16.72	30.80	18,543			
	LSC	34.56	34.63	16.10	14.71	34.51	85,857			
T. rupestris var. *ciliate*	Total	31.97	31.24	18.77	18.03	36.80	15,5479	84	36	8
	IRA	28.57	28.64	22.16	20.63	42.79	25,580			
	IRB	28.64	28.57	20.63	22.16	42.79	25,580			
	SSC	34.56	34.51	16.17	14.76	30.94	18,499			
	LSC	33.39	32.11	17.77	16.73	34.50	85,820			

The complete chloroplast genome of *T. rupestris* and *T. rupestris* var. *ciliate* contained 112 different genes out of which 6 are duplicated in the IRA and IRB region, for a total of 131 genes. The number of rRNA genes, tRNA genes, and protein-coding genes in the genome are 4, 29, and 79, respectively (Figure 2 and Table 2). Prediction of the *T. rupestris* and *T. rupestris var. ciliate* cp gene function was based on homology. Because these genes encode a variety of proteins, they are mainly involved in photosynthesis and other metabolic processes. Regarding photosynthesis, a subset of genes synthesize large Rubisco subunits and vesicle-like proteins. In addition, other genes encode subunits of a protein complex that mediates redox reactions to recycle electrons. Table 2 shows the gene functions and groups in the *T. rupestris* and *T. rupestris* var. *ciliate* cp genome.

Table 2
Genes present in the chloroplast genome of *T. rupestris* and *T. rupestris* var. *ciliate*

Category	**Group of genes**	**Name of genes**
rRNA	rRNA genes	*rrn16S* (×2), *rrn23S* (×2), *rrn4.5S* (×2), *rrn5S* (×2)
tRNA	tRNA genes	*trnA-UGC* (×2), *trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UCC, trnH-GUG, trnI-CAU* (×2), *trnI-GAU* (×2), *trnK-UUU, trnL-CAA* (×2), *trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU (×2), trnP-UGG, trnQ-UUG, trnR-ACG* (×2), *trnR-UCU, trnS-GCU* (×2), *trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA*
Genes for photosynthesis	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Subunits of NADH-dehydrogenase	*ndhA, ndhB* (×2), *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Subunit of rubisco	*rbcL*
Self replication	Large subunit of ribosome	*rpl14, rpl16, rpl2* (×2), *rpl20, rpl22, rpl23* (×2), *rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12* (×2), *rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7* (×2), *rps8*
	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
Other genes	Protease	*clpP*
	Translational initiation factor	*infA*
	Maturase	*matK*

Category	Group of genes	Name of genes
Unknown	Conserved open reading frames	*ycf1* (×2), *ycf2* (×2), *ycf4*

The chloroplast genome of *T. rupestris* and *T. rupestris* var. *ciliate* are found to have intron in some of the genes. Out of the 131 different genes, 15 of them contain intron (Table 3), five tRNAs (*trnK-UUU, trnG-UCC, trnL-UAA, trnI-GAU*, and *trnA-UGC*) and ten protein-coding genes (*rps16, rpoC1, ycf3, clpP, petB, petD, rpl16, rpl2, ndhB, ndhA*). Four of the genes with intron viz.: *rpl2, ndhB, trnA-UGC*, and *trnI-GAU* are situated in the inverted repeat region, the 10 genes are in the large single-copy region *(trnK-UUU, rps16, trnG-UCC, rpoC1, ycf3, trnL-UAA, clpP, petB, petD*, and *rpl16*), and 1 gene (*ndhA*) is in the short single copy region. *Ycf3* and *clpP* are the only genes with two introns, while the other 17 genes have one intron.

Table 3

Genes with intron in the *T. rupestris* and *T. rupestris* var. *ciliate* chloroplast genome and length of exons and introns

	Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII	ExonIII
T. rupestris	*trnK-UUU*	-	1634	4310	37	2605	35	/	/
	rps16	-	5444	6610	39	915	213	/	/
	trnG-UCC	+	9456	10258	32	711	60	/	/
	rpoC1	-	20984	23750	430	718	1619	/	/
	ycf3	-	43671	45654	129	721	228	753	153
	trnL-UAA	+	48468	49099	35	547	50	/	/
	clpP	-	71086	73219	71	852	291	694	226
	petB	+	76159	77590	6	784	642	/	/
	petD	+	77784	78983	9	717	474	/	/
	rpl16	-	82669	84175	9	1093	405	/	/
	rpl2	-	85922	87412	391	627	473	/	/
	ndhB	-	95935	98147	775	680	758	/	/
	trnI-GAU	+	103619	104642	32	952	40	/	/
	trnA-UGC	+	104707	105592	37	813	36	/	/
	ndhA	-	121204	123498	553	1203	539	/	/
	trnA-UGC	-	135824	136709	37	813	36	/	/
	trnI-GAU	-	136774	137797	32	952	40	/	/
	ndhB	+	143269	145481	775	680	758	/	/
	rpl2	+	154004	155494	391	627	473	/	/
T. rupestris var. *ciliata*	*trnK-UUU*	-	1628	4365	37	2666	35	/	/
	rps16	-	5494	6647	39	902	213	/	/
	trnG-UCC	+	9465	10267	32	711	60	/	/
	rpoC1	-	20991	23757	430	718	1619	/	/
	ycf3	-	43656	45639	129	721	228	753	153
	trnL-UAA	+	48446	49077	35	547	50	/	/
	clpP	-	71077	73209	71	853	291	692	226

	Gene	Strand	Start	End	ExonI	IntronI	ExonII	IntronII	ExonIII
	petB	+	76141	77573	6	785	642	/	/
	petD	+	77768	78967	9	717	474	/	/
	rpl16	-	82656	84137	9	1068	405	/	/
	rpl2	-	85885	87375	391	627	473	/	/
	ndhB	-	95898	98110	775	680	758	/	/
	trnI-GAU	+	103592	104615	32	952	40	/	/
	trnA-UGC	+	104680	105565	37	813	36	/	/
	ndhA	-	121137	123436	553	1208	539	/	/
	ycf1	-	125368	131164	1058	151	4588	/	/
	trnA-UGC	-	135901	136786	37	813	36	/	/
	trnI-GAU	-	136851	137874	32	952	40	/	/
	ndhB	+	143356	145568	775	680	758	/	/
	rpl2	+	154091	155581	391	627	473	/	/

3.2 Codon usage, RNA editing sites, and repeat sequences

3.2.1 Codon usage

The codon usage frequency was calculated from the sequence of protein-coding genes, where the RSCU (relative frequency of occurrence of synonymous codon usage for a specific amino acid) values are shown in Figure 3 and Table S1. The *T. rupestris* and *T. rupestris* var. *ciliate* plastomes showed very similar frequencies of codon usage despite morphological and evolutionary divergence among them. We found that all possible codons of amino acids are used in their plasmids as specified in Table S1. The protein-coding genes present a total of 26,628 codons in *T. rupestris* to 26,632 in *T. rupestris* var. *ciliate* plastome (Figure 3, Table S1). Leucine (10.55 in *T. rupestris*, 9.98 in *T. rupestris* var. *ciliate*), Serine (7.65 in, 9.26 in *T. rupestris* var. *ciliate*), and Arginine (5.98 in *T. rupestris*, 6.41 in *T. rupestris* var. *ciliate*) are the most abundant in *T. rupestris* and *T. rupestris* var. *ciliate*. Methionine (2.34 in *T. rupestris*, 1.73 in *T. rupestris* var. *ciliate*) and Tryptophan (1.71 in *T. rupestris*, 1.92 in *T. rupestris* var. *ciliate*) are the least abundant in *T. rupestris* and *T. rupestris* var. *ciliate*. Moreover, we found that the distribution of codon types was consistent in *T. rupestris* and *T. rupestris* var. *ciliate*, which are also consistent with the patterns detected in *Rubus* [36] and other angiosperms [37] and algal lineages [38].

3.2.2 RNA editing analysis

After transcription of chloroplast mRNA molecules, RNA editing, a site-specific C to U conversion process, usually regulates gene expression and translation in the chloroplast [39]. The types and amounts of RNA editing are the same in *T. rupestris* and *T. rupestris* var. *ciliate*. In 27 protein-coding genes, the total number of 124 possible RNA editing sites were predicted among *T. rupestris* and *T. rupestris* var. *ciliate* plastomes (Table S2). These genes include photosynthesis-related genes (*atpA*, *atpB*, *atpF*, *atpI*, *ndHA*, *ndhB*, *ndhD*, *ndhF*, *ndhG*, *petB*, *petD*, *petG*, and *psbE*), self-replication genes (*rpl2*, *rpl23*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, and *rps14*), and others (*matK* and *ycf3*). The highest number of potential editing sites were found in *ndhB* gene (14 sites), followed by the *psbB* gene (10 sites). Like mt-genome, no correlation was observed between the length of the gene and the predicted RNA-editing sites in the protein-coding genes (Table S2).

3.2.3 Long-Repeat and SSR Analysis

Using the default settings of the REPuter program to screen for repetitive sequences in the *T. rupestris* and *T. rupestris* var. *ciliate* chloroplast genomes, the program showed that only three types of repeats were present in the genome, viz. palindromic, forward and reverse, and no complement repeats were detected in the cp genome (Table S3). Table S3 demonstrated that 17 palindromic repeats, 28 forward repeat, and 4 reverse repeats were shown in the *T. rupestris*, and 17 palindromic repeats, 29 forward repeats, and 3 reverse repeats were exhibited in the *T. rupestris* var. *ciliate*. Most of the size of the repeats are between 20 and 29 bp (77.55%), followed by 30–39 bp (12.24%) whereas 40-49 bp (8.16%) and 50-59 bp (2.04%) are the least. In all, there are 49 number repeats in *T. rupestris* cp genome. In the *T. rupestris* var. *ciliate*, 49 repeats were shown in the *T. rupestris* var. *ciliate*, which are between 20 and 29 bp (77.55%), followed by 30-39 bp (12.24%), and 40-49 bp (8.16%) and 50-59 bp (2.04%) are the least.

There are 61 and 65 simple sequence repeats in *T. rupestris* and *T. rupestris* var. *ciliate*, respectively. Single nucleotide simple sequence repeats (SSRs) are the most abundant. Among all SSR types, A and T were the most commonly used bases, and A and T of *T. rupestris* are 28 and 31. There are 30 and 31 of A and T in the *T. rupestris* var. *ciliate* (Table S4). Our results showed the intraspecific variation in repeat number, repeats distribution, and repeat motifs, the highly similar morphological characteristics of *T. rupestris* and *T. rupestris* var. *ciliate* presented minor SSRs changes.

3.3 Phylogenetic analysis and time estimation

The cp genomes of the 29 species were applied to to reconstruct the phylogenetic tree determine the phylogenetic relationship and tribal positions of the nine species of *Taihangia*. Using 95 protein-coding genes and the complete plastome sequences, we performed phylogenetic analyses of the 27 Rosoideae Focke plastomes (Figure 4A). *G. macrophyllum*, *G. rupestre*, *T. rupestris*, and *T. rupestris* var. *ciliate* were clustered in one clade with strong support, and were divided into two major subclades. Sub clade 1 which is monophyletic includes *G. macrophyllum* and clade 2 containing *G. rupestre*, *T. rupestris*, and *T. rupestris* var. *ciliate*. Clade 2 has two major subclades, One branch is *G. rupestre* and the other is *T. rupestris*, and *T. rupestris* var. *ciliate.* Figure 4A indicated that *G. rupestre* is closely related to *T. rupestris*, and *T. rupestris* var. *ciliate*. *G. macrophyllum*, *G. rupestre*, *T. rupestris*, and *T. rupestris* var. *ciliate* belong to the Trib. Colurieae, the result is the same as it. The result indicated that four of them are closely related.

Divergence time was estimated for each internal node of the phylogenetic tree (Figure 4B). *Taihangia* genus was inferred to have originated at 0.2057 Mya, *G. rupestre* was inferred to have originated at 1.4431 Mya, and *G. macrophyllum* was inferred to have originated at 9.8532 Mya. The detected divergence time of *T. rupestris*, and *T. rupestris* var. *ciliate* may contribute to future studies on genus *Taihangia*.

3.4 Comparative analysis and sequence divergence analysis

3.4.1 Sliding window analysis

Sliding window analysis using the DnaSP program reveals highly variable regions in the cp genomes of two *Taihangia*. The sliding windows analysis (Figure 5) highlights two plastome regions as hotspots of nucleotide divergence among *T. rupestris* and *T. rupestris* var. *ciliate*. These hotspots correspond to three intergenic regions (*petA-psbJ, psbJ-psbL, and trnR*^{UCU}*-atpA*) and four genes (*psbA* and *ndhF*).

3.4.2 IR expansion and contraction

IR expansions and contractions are common in cp genomes, which results in the change in cp genome size [40]. The differences in IRs may also reflect phylogenetic history. Here, we selected four species of Trib. Colurieae and compared their sizes and the junctions of their LSC, SSC, and IR regions. Although the lengths of the IR regions, ranging from 25,579 bp to 26,152 bp, varied little among the four species, some differences in the IR expansions and contractions were observed.

As shown in Figure 6, the *rps19* is mostly located in the LSC region and the LSC-IRB boundary at bases 1-8 bp. The gene *ycf1* was found to have 1107 bp, 1098 bp, 1107 bp, and 1188 in the IRB region in *T. rupestris*, *T. rupestris* var. *ciliate*, *G. rupestre*, and *G. macrophyllum* respectively. Whereas, the gene *ycf*1 was found to have 4,523 bp, 4,532 bp, 4,523 bp, and 4,424 in the SSC region in *T. rupestris*, *T. rupestris* var. *ciliate*, *G. rupestre*, and *G. macrophyllum.* Furthermore, the *ycf*1 gene extended into the SSC region excepted the *G. rupestre* genome, and the longest overlap between the SSC region and *ycf*1 genes was also observed in *T. rupestris* var. *ciliate*. The *trnH* is mostly located in the LSC region and the IRB-LSC boundary at bases 4-65 bp. In summary, the structure of the cp genomes is conservative among *T. rupestris* and *T. rupestris* var. *ciliate*.

3.3.3 Genome comparison

We analyzed the sequence differences of *T. rupestris*, *T. rupestris* var. *ciliate*, *G. macrophyllum*, and *G. rupestre* cp genomes using mVISTA (Figure 7), the *G. rupestre* cp genome sequence was set as the reference cp genome. Sequence comparison of 4 whole plastomes generated multiply aligned sequences of 155,558 bp in length. The analysis shows overall sequence identity and divergent regions in *Taihangia* and *Geum*. *T. rupestris* and *T. rupestris* var. *ciliate* cp genome sequences showed very high sequence similarities. A high degree of synteny and gene order conservation indicates evolutionary conservation at the plastome level (Figure 7). Notably, the LSC and SSC regions have greater divergence than the IRs, while the non-coding regions show higher sequence divergence than the coding regions, while the exons, introns, and ncRNA generally had little variation between genomes, which is similar to the results of previous studies [41].

3.3.4 Kimura’s two-parameter (K2P) analysis

To discover the hypervariable regions among *T. rupestris*, *T. rupestris* var. *ciliate*, *G. rupestre*, and *G. macrophyllum.* 95 intergenic regions were retracted from the chloroplast genomes of 4 species, and the genetic distance of the intergenic regions were calculated by the K2p (Kimura 2-parameter) model. A total of 29 intergenic regions has K2p values ranging from 2.35 to 25.745. Among them, *rps16-trnQ-UUG*, *trnH-GUG-psbA*, *trnF-GAA-ndhJ* have higher K2p values, which are 5.745, 4.752, and 4.607, respectively. It can be seen that these regions vary greatly among the chloroplast genomes of the four species, and can be used as potential molecular marker development regions (Figure 8).

3.5 Selective pressures analysis

Nonsynonymous (amino acid-replacing, Ka) and synonymous (Ks) substitutions and their ratio (Ka/Ks) are applied to reveal the intensity of natural selection on DNA sequence evolution [42, 43]. Ka/Ks value > 1 indicates positive selection, Ka/ Ks < 1 indicates purification or negative selection, and Ka/ Ks value 1 indicates neutral selection. The Ka/Ks ratio was calculated and compared for 96 protein-coding genes in *T. rupestris*, *T. rupestris* var. *ciliate*, *G. rupestre*, and *G. macrophyllum* chloroplast genomes to investigate genome evolution (Table 4). The Ka/Ks of

Taihangia. *T. rupestris*, *T. rupestris* var. *ciliate*, *G. macrophyllum* and *G. rupestre* are 1.24455, 0.814338, 0.652981, and 0.732876, respectively. This result indicated that *T. rupestris* was subjected to positive pressure selection, *T. rupestris* var. *ciliate*, *G. macrophyllum*, and *G. rupestre* were subjected to negative selection.

Table 4. Substitution rates of 96 protein-coding genes in four chloroplast genomes.

Species	Ka	Ks	Ka/Ks
Taihangia rupestris	0.001144	0.000919	1.24455
Taihangia rupestris var*. ciliata*	0.002096	0.002574	0.814338
Geum rupestre	0.00066	0.00101	0.652981
Geum macrophyllum	0.012739	0.017382	0.732876

4. Conclusion

To further search for molecular markers, reconstruct phylogeny and provide a theoretical basis for bioprospecting, we characterized for the first time the complete chloroplast sequences of two species of *Taihangia* in the Taihang Mountains of China. Our study presents the most comprehensive chloroplast phylogenomic and biogeographical analyses of the taxonomically complex genus *Taihangia*. *T. rupestris* (155,558 bp) and *T. rupestris* var. *ciliate* (155,479 bp) chloroplast genome is fully characterized and compared to their closely related species of *Geum* genus. Both *T. rupestris* and *T. rupestris* var. *ciliate* have 4 rRNA genes, 29 tRNA genes, and 79 protein-coding genes. Overall, the gene contents, gene arrangements, The types, and frequency of codon usage, repeat sequences, and SSRs are similar and highly conserved in the species of *T. rupestris* and *T. rupestris* var. *ciliate*. In addition, high sequence variation in protein-coding and intergenic regions, whole cp genomes, nucleotide substitution are similar in the chloroplast genomes of species of the *Taihangia* and *Geum*. *T. rupestris*, *T. rupestris* var. *ciliate*, *G. macrophyllum*, and *G. rupestre* are closely related, which indicates that *Taihangia* is closely related to *Geum*. *Taihangia* genus was inferred to have originated at 0.2057 Mya, *Geum rupestre* was inferred to have originated at 1.4431 Mya, and *Geum macrophyllum* was inferred to have originated at 9.8532 Mya. The Kimura's two-parameter value and patterns and amino acid sites under potentially positive selection in the chloroplast genomes of species of the *Taihangia* and *Geum* may be useful for the development of genealogy-specific markers for genetic diversity and genetic evolution studies. *Taihangia* genus might be the potential medicinal resource according to bioprospecting, we exploring its chemical composition and efficacy are under way. This study is useful for the protection and rational use are of the great significance of *Taihangia* resources.

Declarations

Acknowledgment

This work was supported by the National Natural Science Foundation of China (U1812403-1 and 82073960), Chengdu University of Traditional Chinese Medicine (003109034001), Beijing Natural Scientific Foundation (7202135) and National Science & Technology Fundamental Resources Investigation Program of China (2018FY100701), which are gratefully acknowledged.

Author Contributions

LFH contributed to the concept and design of the study and supervised this research. YZ analyzed the data and wrote the manuscript. YJ and YJM contributed to the chloroplast genome analysis. ZF, MZ, XKZ, and TXZ reviewed and edited the manuscript. All authors have read and approved the final manuscript.

Availability of data and materials

The complete chloroplast genomes were generated in the NCBI database (https://www.ncbi.nlm.nih.gov/) with accession numbers MZ151697 and MZ151698. Supplemental tables are available in Supplemental table.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Declaration of Competing Interest

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

Author details

[1]Key Laboratory of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. [2]Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi, China. [3]Dali University, Da Li, Yun Nan, China.

References

1. Li W, Zhang L, Ding Z, Wang G, Zhang Y, Gong H, Chang T, Zhang Y: **De novo sequencing and comparative transcriptome analysis of the male and hermaphroditic flowers provide insights into the regulation of flower formation in andromonoecious taihangia rupestris**. *BMC Plant Biol* 2017, **17**(1):54.
2. Yü TT, Li CL: **Taihangia Yü et Li - a new genus of Rosaceae from China**. *Chih wu fen lei hsueh pao = Acta phytotaxonomica Sinica* 1980.
3. Yu TC, Li CL: **Taihangia Yu et Li–a new genus of Rosaceae from China**. *Chih wu fen lei hsueh pao = Acta phytotaxonomica Sinica* 1980.
4. Wang S, Xie Y: **China species red list**. 2004.
5. Yü T, Li CL: **The Systematic Position of Genus Taihangia in Rosaceae**. *Computer Technology & Development* 1983.

6. Shen SH, Lu WL, Wang FH: **Studies on the reproductive biology of Taihangia rupestris: I Analysis on the habitat of T. repestris**. *Biodiversity Science* 1994, **2**(4):210–212.
7. Lu W, Shen S, Wang F: **studies on reproductive biology or Taihangia rupes2 Investigation and study of sexual and asexual reproduction**. *Biodiversity Science* 1995, **03**(1):8–14.
8. Xutian S, Tai S, Yuan CS: **Handbook of traditional Chinese medicine**. *WORLD SCIENTIFIC* 2014.
9. Sun X, Wang YP, Liu C, Huang LF: **Molecular identification of Taihangia rupestris Yu et Li, an endangered species endemic to China**. *South African Journal of Botany* 2019, **124**:173–177.
10. Raven JA, Allen JF: **Genomics and chloroplast evolution: what did cyanobacteria do for plants?** *Genome biology* 2003, **4**(3):209.
11. Ravi V, Khurana JP, Tyagi AK, Khurana P: **An update on chloroplast genomes**. *Plant Systematics & Evolution* 2008, **271**(1-2):101–122.
12. Bendich A: **Circular Chloroplast Chromosomes: The Grand Illusion**. *The Plant cell* 2004, **16**:1661–1666.
13. Alzahrani DA, Yaradua SS, Albokhari EJ, Abba A: **Complete chloroplast genome sequence of Barleria prionitis, comparative chloroplast genomics and phylogenetic relationships among Acanthoideae**. *BMC Genomics* 2020, **21**(1).
14. Asaf S, Khan A, Khan M, Imran Q, Kang S-M, Al-Hosni K, Jeong E, Lee K, Lee I-J: **Comparative analysis of complete plastid genomes from wild soybean (Glycine soja) and nine other Glycine species**. *PLOS ONE* 2017, **12**:e0182281.
15. Cho KS, Yun BK, Yoon YH, Hong SY, Yang TJ: **Complete Chloroplast Genome Sequence of Tartary Buckwheat (Fagopyrum tataricum) and Comparative Analysis with Common Buckwheat (F. esculentum)**. *Plos One* 2015, **10**(5):e0125332.
16. Porebski S, Bailey LG, Baum BR: **Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components**. *Plant Molecular Biology Reporter* 1997, **15**(1):8–15.
17. Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011.
18. Nicolas D, Patrick M, Guillaume S: **NOVOPlasty: de novo assembly of organelle genomes from whole genome data**. *Nucleic Acids Research* (4):4.
19. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C: **CPGAVAS2, an integrated plastome sequence annotator and analyzer**. *Nucleic Acids Research* 2019, **47**(W1):W65-W73.
20. Michael T, Pascal L, Tommaso P, Ulbricht-Jones ES, Axel F, Ralph B, Stephan G: **GeSeq – versatile and accurate annotation of organelle genomes**. *Nucleic Acids Research* 2017(W1):W1.
21. Sharp PM, Li WH: **The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications**. *Nucl Acids Res* 1987, **15**(3):1281–1295.
22. Mower JP: **The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments**. *Nucleic Acids Research* 2009, **37**(Web Server issue):W253-259.
23. Kurtz, Choudhuri, Ohlebusch, Schleiermacher: **REPuter: the manifold applications of repeat analysis on a genomic scale**. *Nucleic Acids Research* 2001.
24. Beier S, Thiel T, Münch T, Scholz U, Mascher M: **MISA-web: a web server for microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583–2585.
25. Zhang D, Gao F, Jakovlić I, Zou H, Zhang J, Li WX, Wang GT: **PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies**. *Mol Ecol Resour* 2020, **20**(1):348–355.

26. Katoh K, Rozewicki J, Yamada KD: **MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization**. *Brief Bioinform* 2019, **20**(4):1160–1166.
27. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312–1313.
28. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547–1549.
29. Dobe C, Paule J: **A comprehensive chloroplast DNA-based phylogeny of the genus Potentilla (Rosaceae): implications for its geographic origin, phylogeography and generic circumscription**. *Molecular Phylogenetics & Evolution* 2010, **56**(1):156–175.
30. Becker, Herman, F.: **The York Ranch Flora of the Upper Ruby Basin, Southwestern Montana**. *Palaeontographica Abteilung B* 1973, **143**(1-4):18–93.
31. Rozas J, Sánchezdelbarrio J, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods**. *Bioinformatics* 2003, **19**(18):2496–2497.
32. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences**. *J Mol Evol* 1980, **16**(2):111–120.
33. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276–277.
34. Katoh, K.: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Research* 2002, **30**(14):3059–3066.
35. Alexandros S: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014(9):1312–1313.
36. Yang JY, Takayama K, Pak JH, Kim SC: **Comparison of the Whole-Plastome Sequence between the Bonin Islands Endemic Rubus boninensis and Its Close Relative, Rubus trifidus (Rosaceae), in the Southern Korean Peninsula**. *Genes* 2019, **10**(10):774-.
37. Ravi, V., Khurana, J., P., Tyagi, A., K., P.: **An update on chloroplast genomes**. *Plant Systematics & Evolution* 2008.
38. Morton BR: **Selection on the codon bias of chloroplast and cyanelle genes in di3erent plant and algal lineages**. *Journal of Molecular Evolution* 1998, **46**(4):449–459.
39. Zhang Y, An D, Li C, Zhao Z, Wang W: **The complete chloroplast genome of greater duckweed (Spirodela polyrhiza 7498) using PacBio long reads: insights into the chloroplast evolution and transcription regulation**. *BMC Genomics* 2020, **21**(1).
40. Cui JL, Zhang YY, Vijayakumar V, Zhang G, Wang ML, Wang JH: **Secondary Metabolite Accumulation Associates with Ecological Succession of Endophytic Fungi in Cynomorium songaricum Rupr**. *J Agric Food Chem* 2018, **66**(22):5499–5509.
41. Hong Z, Wu Z, Zhao K, Yang Z, Xu D: **Comparative Analyses of Five Complete Chloroplast Genomes from the Genus Pterocarpus (Fabacaeae)**. *International Journal of Molecular Sciences* 2020, **21**(11):3758.
42. Yang, Nielsen: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models**. *Molecular biology and evolution* 2000.
43. Fang J, Lin A, Yuan X, Chen Y, Xue T: **The complete chloroplast genome of Isochrysis galbana and comparison with related haptophyte species**. *Algal Research* 2020, **50**:101989.

Figures

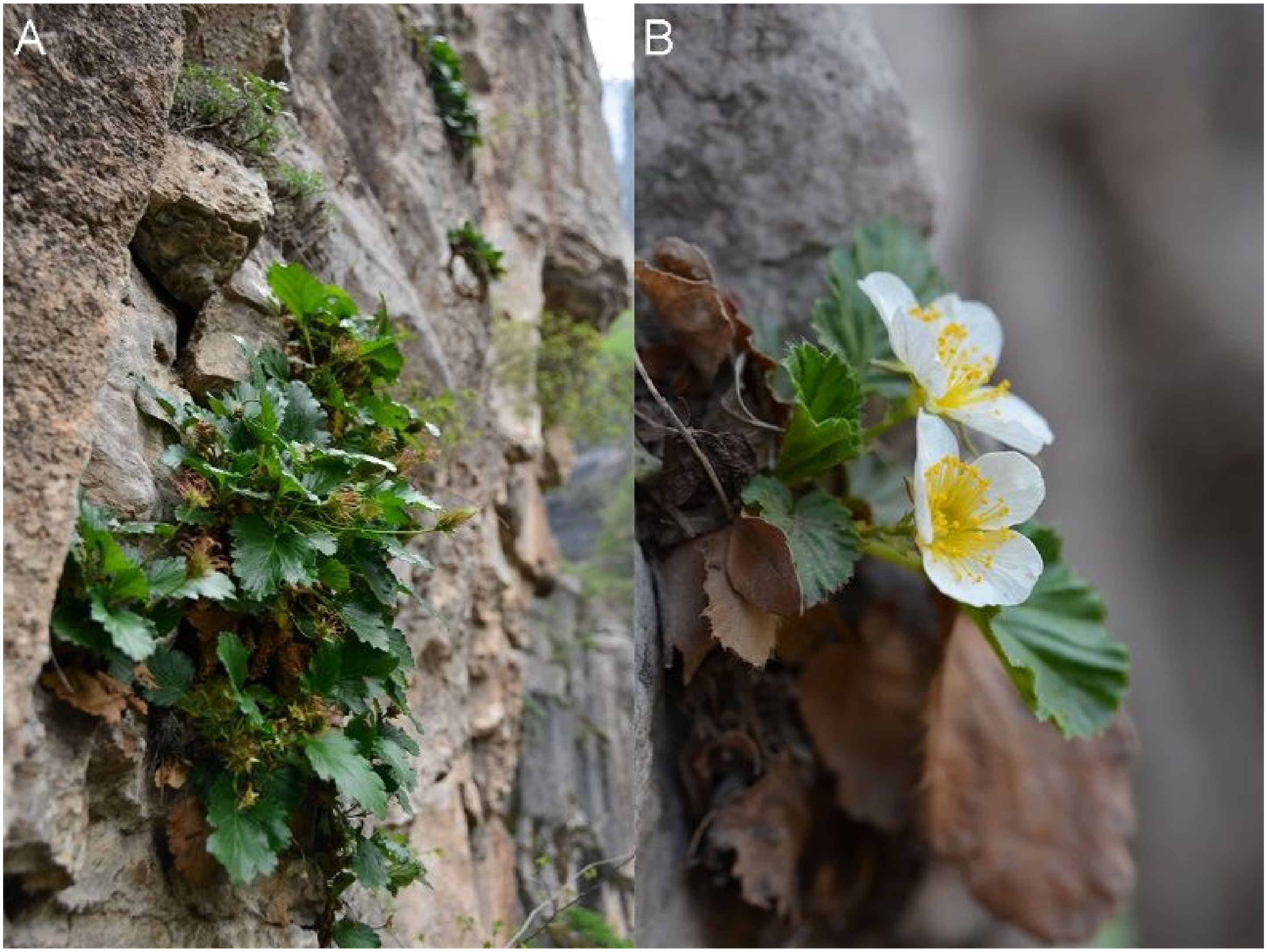

Figure 1

The habitat of T. rupestris and T. rupestris var. ciliate. (a) T. rupestris. (b) T. rupestris var. ciliate.

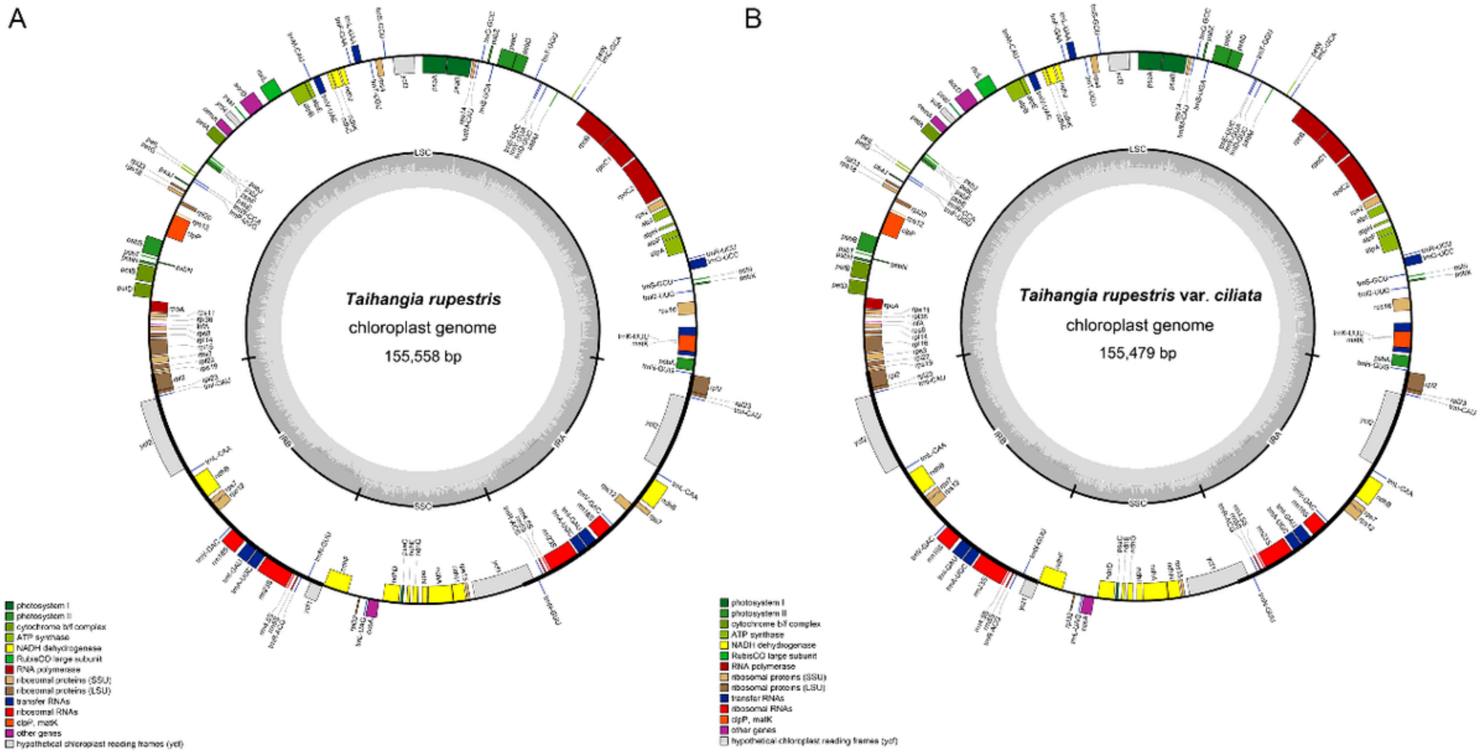

Figure 2

Chloroplast genome maps of T. rupestris and T. rupestris var. ciliate. A. The Chloroplast genome map of T. rupestris. B. The Chloroplast genome map of T. rupestris var. Genes inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise. Genes of different functions are color-coded. The darker gray in the inner circle shows the GC content, while the lighter gray shows the AT content.

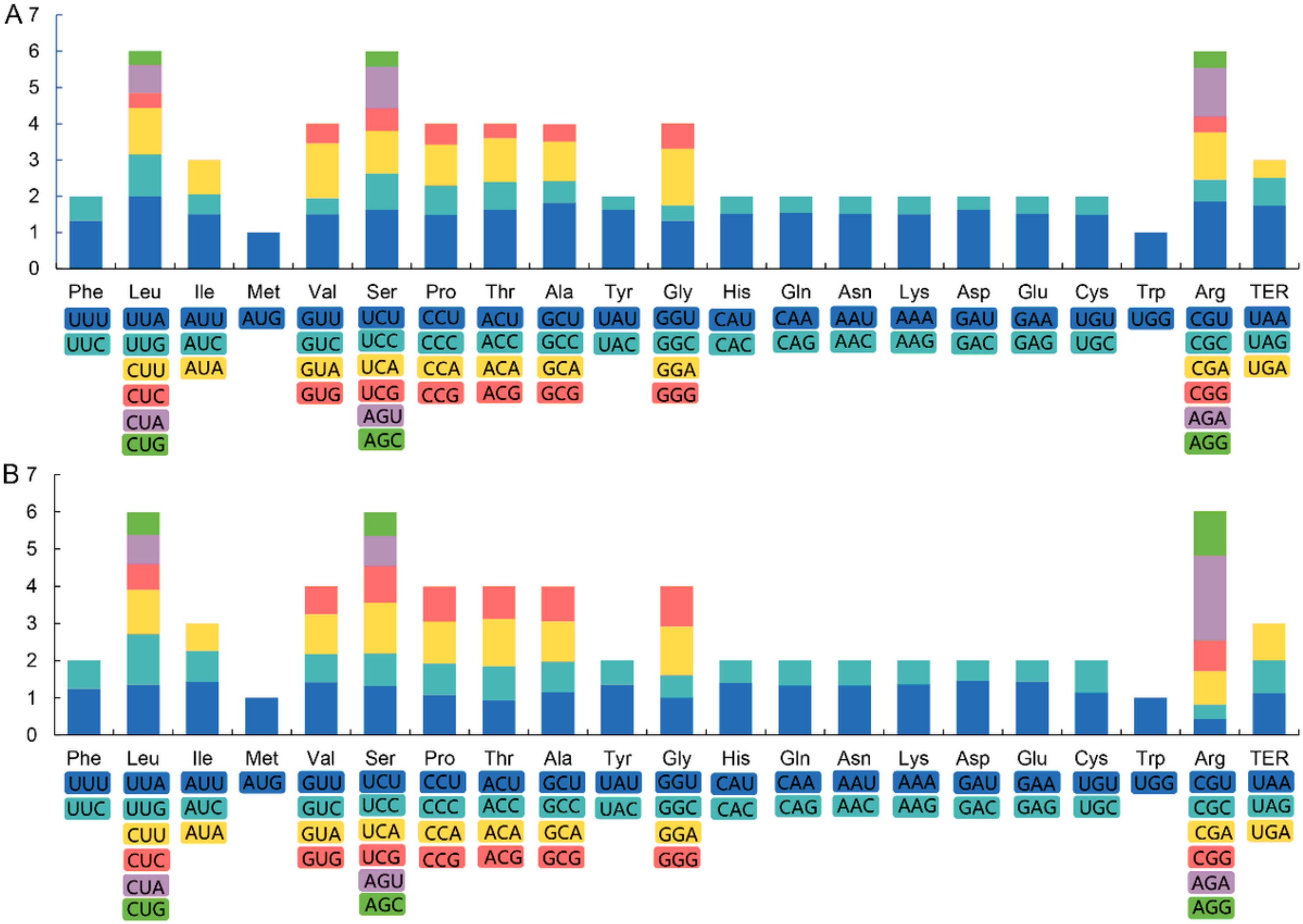

Figure 3

Codon content of 20 amino acids and stop codons in protein-coding genes. A. T. rupestris, B. T. rupestris var. ciliate. Note: The histogram on the left-hand side of each amino acid denotes codon usage within T. rupestris and T. rupestris var. ciliate, and the right-hand side denotes the codon RSCU values. Colors correspond to codons listed underneath the columns.

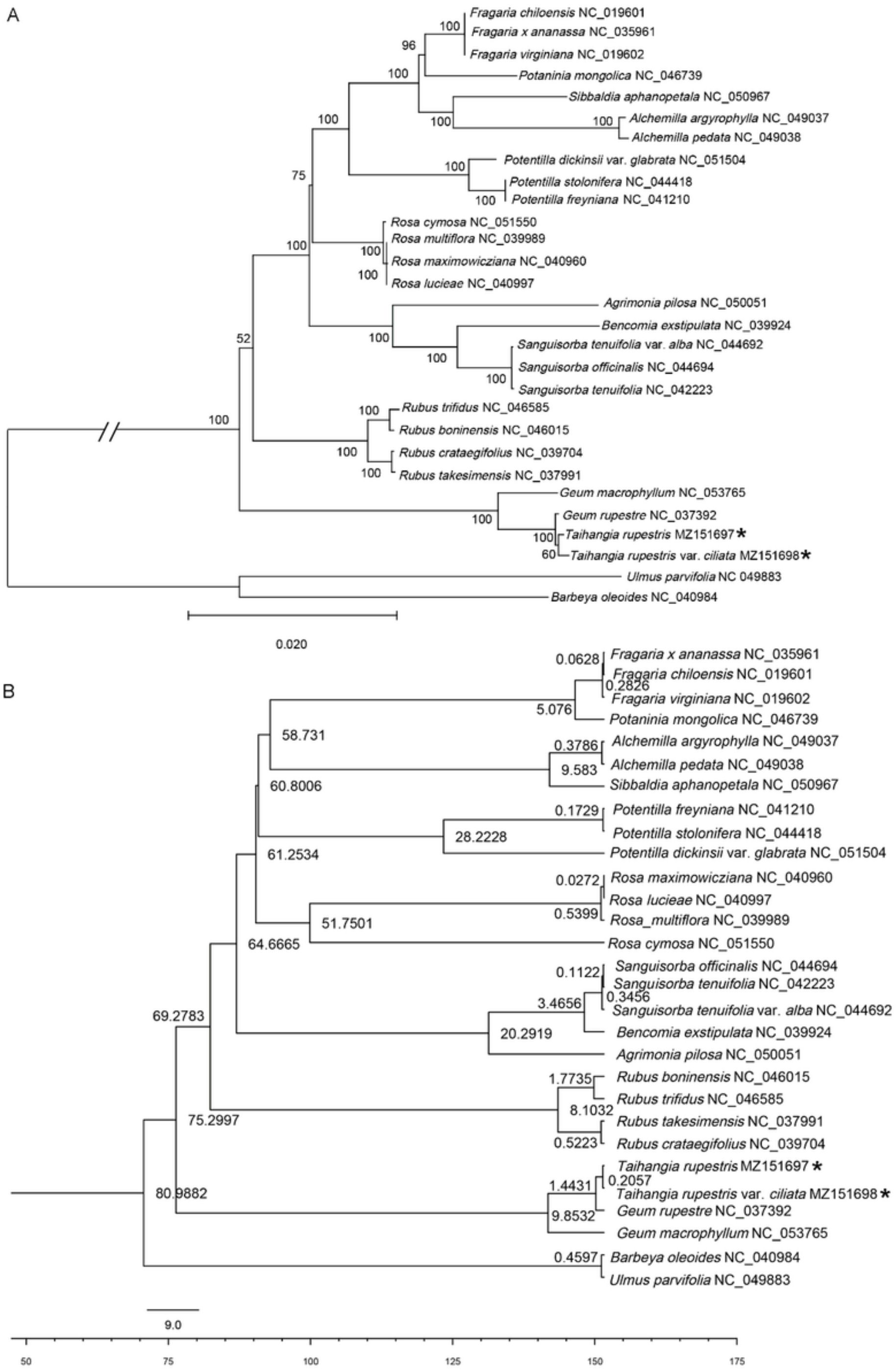
A
Fragaria chiloensis NC_019601
Fragaria x ananassa NC_035961
Fragaria virginiana NC_019602
Potaninia mongolica NC_046739
Sibbaldia aphanopetala NC_050967
Alchemilla argyrophylla NC_049037
Alchemilla pedata NC_049038
Potentilla dickinsii var. glabrata NC_051504
Potentilla stolonifera NC_044418
Potentilla freyniana NC_041210
Rosa cymosa NC_051550
Rosa multiflora NC_039989
Rosa maximowicziana NC_040960
Rosa lucieae NC_040997
Agrimonia pilosa NC_050051
Bencomia exstipulata NC_039924
Sanguisorba tenuifolia var. alba NC_044692
Sanguisorba officinalis NC_044694
Sanguisorba tenuifolia NC_042223
Rubus trifidus NC_046585
Rubus boninensis NC_046015
Rubus crataegifolius NC_039704
Rubus takesimensis NC_037991
Geum macrophyllum NC_053765
Geum rupestre NC_037392
Taihangia rupestris MZ151697 *
Taihangia rupestris var. ciliata MZ151698 *
Ulmus parvifolia NC 049883
Barbeya oleoides NC_040984
100
96
100
100
100
100
100
75
100
100
100
100
100
100
100
52
100
100
100
100
100
100
100
60
0.020
B
Fragaria x ananassa NC_035961
Fragaria chiloensis NC_019601
Fragaria virginiana NC_019602
Potaninia mongolica NC_046739
Alchemilla argyrophylla NC_049037
Alchemilla pedata NC_049038
Sibbaldia aphanopetala NC_050967
Potentilla freyniana NC_041210
Potentilla stolonifera NC_044418
Potentilla dickinsii var. glabrata NC_051504
Rosa maximowicziana NC_040960
Rosa lucieae NC_040997
Rosa_multiflora NC_039989
Rosa cymosa NC_051550
Sanguisorba officinalis NC_044694
Sanguisorba tenuifolia NC_042223
Sanguisorba tenuifolia var. alba NC_044692
Bencomia exstipulata NC_039924
Agrimonia pilosa NC_050051
Rubus boninensis NC_046015
Rubus trifidus NC_046585
Rubus takesimensis NC_037991
Rubus crataegifolius NC_039704
Taihangia rupestris MZ151697 *
Taihangia rupestris var. ciliata MZ151698 *
Geum rupestre NC_037392
Geum macrophyllum NC_053765
Barbeya oleoides NC_040984
Ulmus parvifolia NC_049883
0.0628
0.2826
5.076
58.731
0.3786
9.583
60.8006
0.1729
28.2228
61.2534
0.0272
0.5399
51.7501
64.6665
0.1122
0.3456
3.4656
20.2919
69.2783
1.7735
8.1032
0.5223
75.2997
1.4431
0.2057
9.8532
80.9882
0.4597
9.0
50
75
100
125
150
175

Figure 4

Phylogenetic analysis and divergence time estimation. A. Maximum likelihood tree inferred from 29 representative taxa of Rosoideae Focke. Bootstrap values based on 1000 replicates are shown on each node. * represents the newly assembled plastomes of T. rupestris and T. rupestris var. ciliate in this study. B. Divergence time estimation.

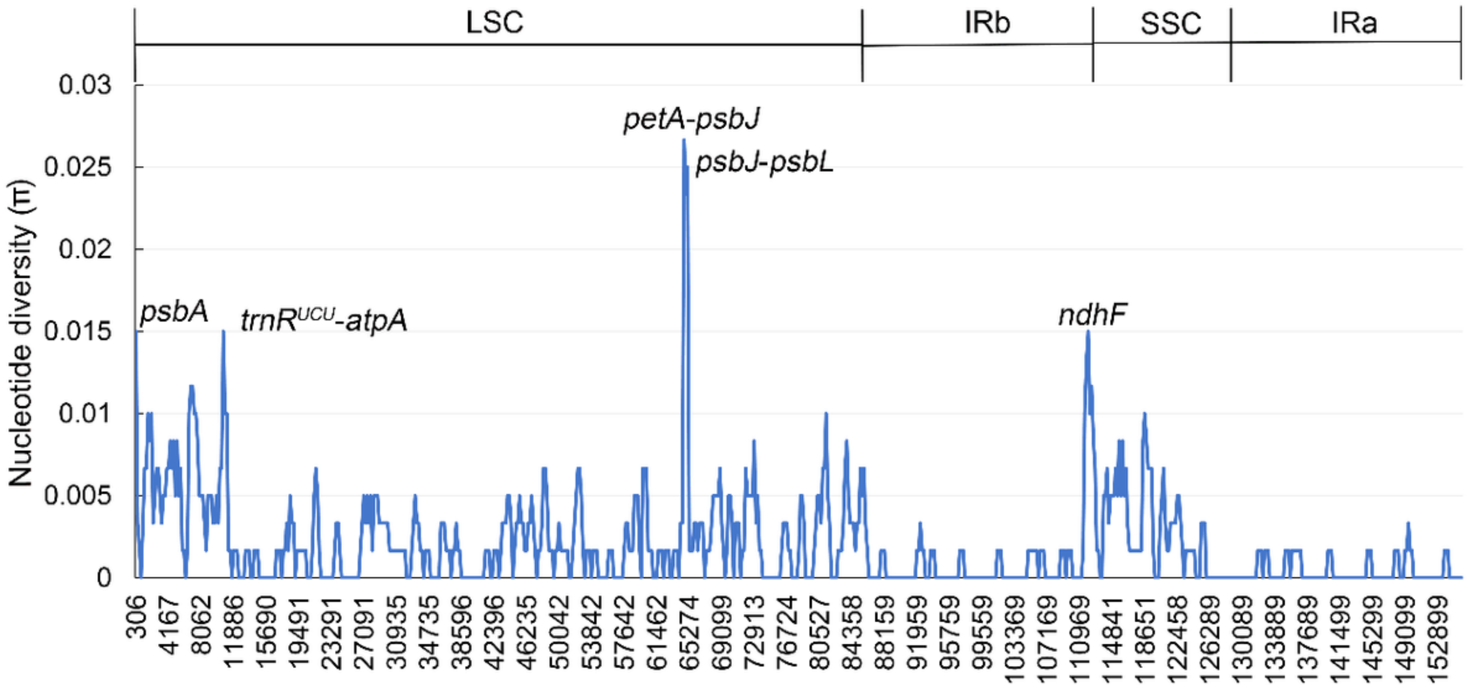

Figure 5

Sliding window test of nucleotide diversity (π) in the multiple alignments of T. rupestris and T. rupestris var. ciliate plastomes. Note: Peak regions with a π value of >0.015 were labeled with loci tags of genic or intergenic region names. π values were calculated in 1 kb sliding windows with 100 bp steps. LSC, large single-copy region; IRA, inverted repeat region a; SSC, small single-copy region; IRB, inverted repeat region b.

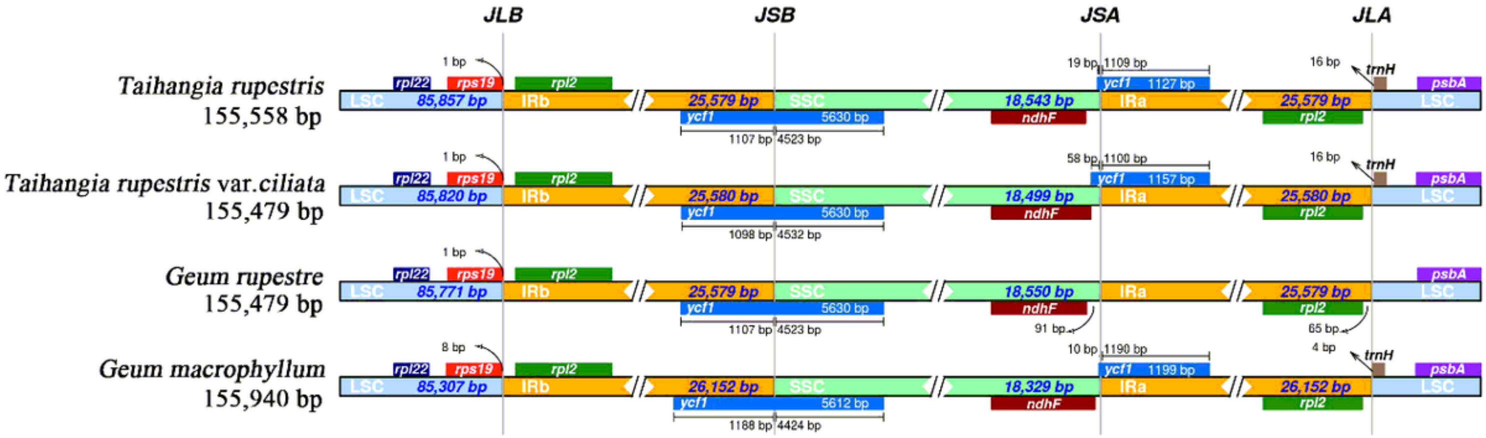

Figure 6

Comparison of the borders of the IR, SSC, and LSC regions among four chloroplast genomes of T. rupestris, T. rupestris var. ciliate, G. rupestre, and G. macrophyllum.

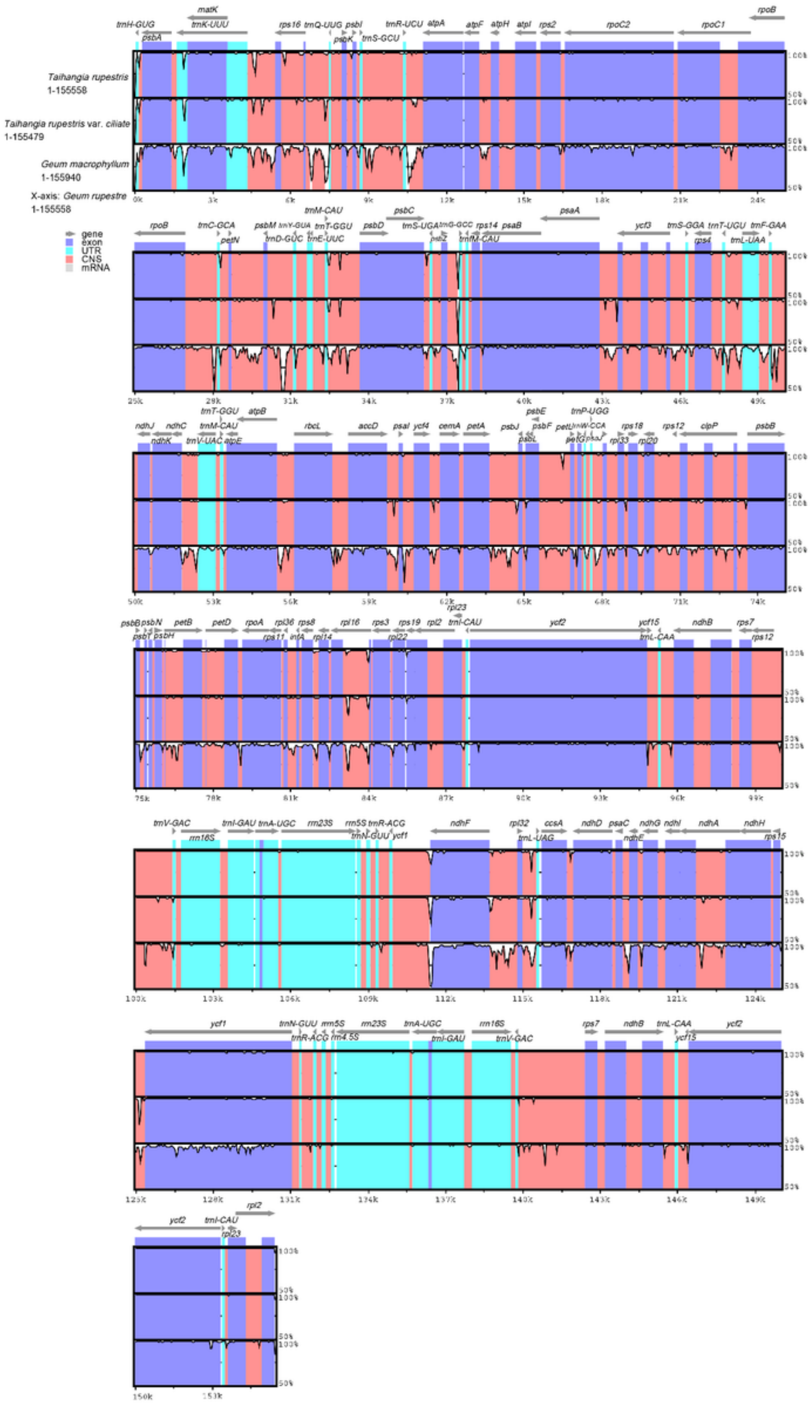

Figure 7

Global alignment of four (T. rupestris, T. rupestris var. ciliate, G. macrophyllum, and G. rupestre) chloroplast genomes of Pterocarpus using mVISTA.Y-axis indicates the range of identity (50–100%). Alignment was performed using G. rupestre as a reference.

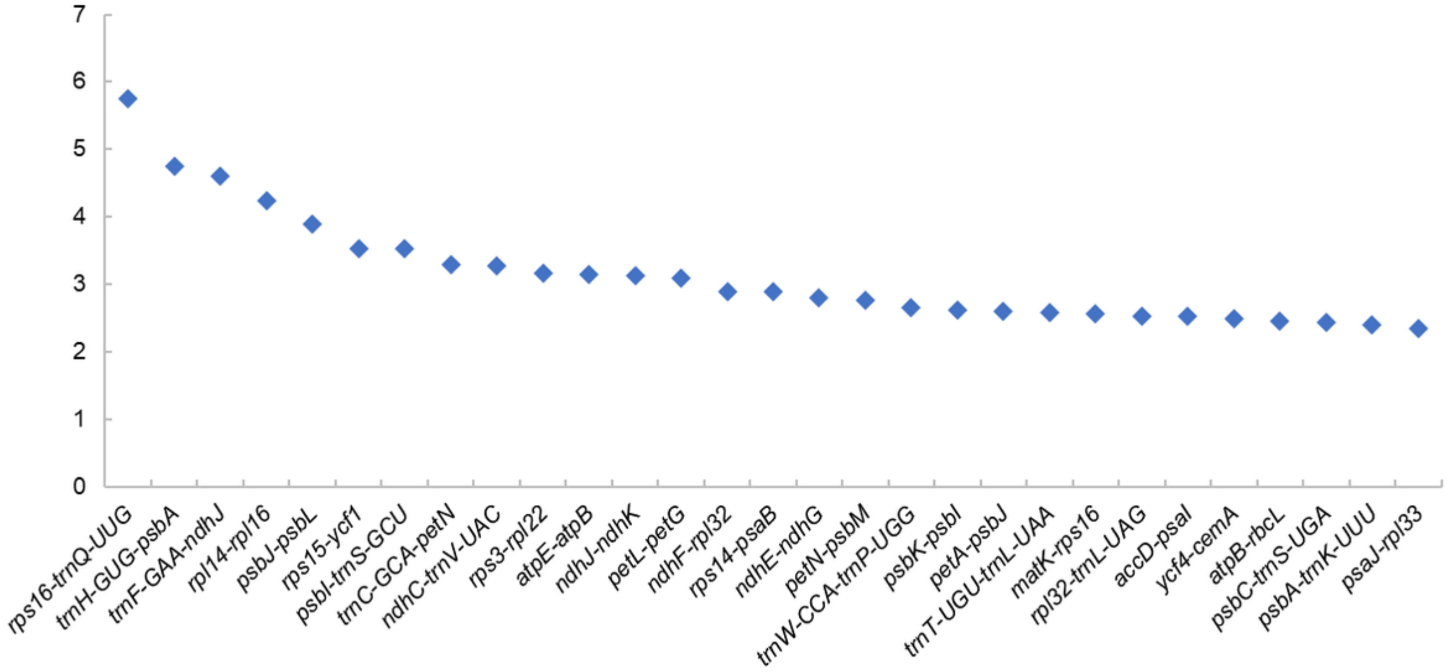

Figure 8

The genetic distance analysis of the genetic distance between T. rupestris and T. rupestris var. ciliate.

Supplementary Files

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

- Supplementarytable.xlsx