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.

Camellia sinensis var. Assamica cv. Duntsa (Theaceae) chloroplast genome and comparative analysis: mutational hotspots and phylogenetic relationships

Jin Li
759904539@qq.com
Shaoyang Academy of Agricultural Sciences
Xiao-Yan Qiu
HuaHuai University Huahuai
Han Tang
Shaoyang Academy of Agricultural Sciences
Jun Tang
Shaoyang Academy of Agricultural Sciences
Ni Zhong
National Research Center of Engineering and Technology for Utilization of Botanical Functional Ingredients, Agricultural University Of Hunan
Tian-Tian Liu
Shaoyang Academy of Agricultural Sciences
Li-Zheng Xiao
National Research Center of Engineering and Technology for Utilization of Botanical Functional Ingredients, Agricultural University Of Hunan
Hua Luo
Shaoyang Academy of Agricultural Sciences

Research Article

Keywords: Camellia sinensis var. Assamica cv. Duntsa, Chloroplast genomes, Comparative analysis, Phylogenetic tree

Posted Date: March 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3971373/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background: Camellia sinensis *var.* assamica *cv.* Duntsa *(*C. duntsa*) is an ancient tea accession in Hunan China. In order to understand the genetic background information of* C. duntsa*, clarify the relationship between* C. duntsa *and other tea trees. we sequenced the complete chloroplast genome of* C. duntsa *using the Illumina NovaSeq platform and compared it to other published chloroplast genomes from tea plants.*

Results: *The* C. duntsa *chloroplast genome is 157,025 bp in length with a GC content of 37.30%. It consists of a short single copy (SSC) region (18,277 bp), a large single copy (LSC) region (86,586 bp), and two inverted repeat regions (IRs) (26,081 bp). A total of 135 genes were identified, including 87 protein-coding genes (PCGs), 8 ribosomal RNA genes (rRNAs), 37 transfer RNA genes (tRNAs), and 3 pseudogene genes (2* ycf15 *and 1* ycf1*). In addition, a total of 968 long repetitive sequences were detected by comparative analysis with other tea tree chloroplast genes, of which 409 were forward, 557 were palindromic, and 2 were reverse. Among the 241-249 SSRs loci analyzed for comparison, most of them were single nucleic acid loci composed of A/T. Besides, 6 mutation hotspots (*rpoC1, ycf1, petB, ndhD, rpl16, rpoC2*) were identified.*

Conclusion: *Phylogenetic analysis showed that* C. duntsa *shows a relatively close evolutionary relationship with* Camellia sinensis *var.* sinensis *cv. Anhua,* Camellia sinensis *var.* sinensis *cv. Fudingdabaicha, and* Camellia ptilophylla*. The results can provide valuable information for better understanding* Camellia s*pecies chloroplast evolution.*

Background

Camellia sinensis var. *Assamica* cv. *Duntsa* (*C. duntsa*) is a tea plant called *"Cheng Bu Dong Cha"* in Chinese [1], belonging to the *Camellia* genus of the *Theaceae* family. It is one of the five major local specialty tea germplasm resources in Hunan Province, China. The main production area of this variety is located in Chengbu Miao Autonomous County, situated at the southern end of the Xuefeng Mountains. The leaves of *C. duntsa* exhibit a high polyphenol content, ranging from 27.26 to 35.63%, and an amino acid content ranging from 1.78 to 4.45%. Such chemical profiles make this variety suitable for producing various teas, including black and green tea. Notably, black tea is made from fresh *C. duntsa* leaves possess unique attributes, with Kung fu black tea presenting a strong floral aroma and a pungent, mellow, and astringent taste and crushed red tea shares similarities with large leaf species of Yunnan and Assam species of India. The high utilization value of *C. duntsa* suggests it is a promising candidate for further development. Besides, numerous techniques, including morphological, cytological, biochemical, and molecular markers, have been applied to *C.duntsa* with the aim of elucidating the species relationships within this genus. The results of these studies indicate that *C. duntsa* is a tea germplasm resource with both primitive and evolutionary characteristics, suggesting that it is a transitional species. However, the lack of detailed molecular phylogenetic investigations in these studies has hindered the understanding of its interspecific relationships.

Chloroplasts are organelles essential to plant life's continued existence, adaption, and development [2]. Compared to the genome of the nucleus, the chloroplast genome is considerably more small. Despite this, they are responsible for encoding essential proteins involved in processes such as photosynthesis, the fixation of carbon and nitrogen, and the formation of starch, pigments, fatty acids, and amino acids [3]. The chloroplast genomes of many angiosperms have multibranched linear structures, and their sizes average between 115 and 165 kb [2]. The typical chloroplast genome structure is known as a tetrameric structure, which consists of a long single-copy region (LSC), a short single-copy region (SSC), and a pair of inverted repeat regions (IRs) separating the SSC from the LSC. IRs are regions that undergo expansion and contraction during the evolution of the chloroplast genome, affecting the rearrangement of the linear order of homologous sequences in the chloroplast genome [4], it is also the main reason for the inconsistent size of the chloroplast genome [5]. Thus, the length of the IRs plays an important role in maintaining the stability of the chloroplast genome [6].

Currently, the study of chloroplasts at the genomic level is mainly focused on DNA barcoding and species tracing based on chloroplast genome sequences and chloroplast genome-wide analysis [7]. A comparative and phylogenetic analysis of six species of *Ligularia* revealed that *L. fischeri* and *L. jauensis* are closely related [8]. The *ycf1b* region in the chloroplast genome can be used as a molecular marker to distinguish different types of *Pterocarpus* wood [9]. The researchers [10]analyzed the ability of traditional chloroplast genome barcodes, species-specific barcodes, and chloroplast whole-genome barcodes to identify *Camphoridae* species using chloroplast whole-genome and nuclear gene data from 191 individuals of 25 genera and 133 species of *Camphoridae*. The chloroplast genomes of many tea trees have been deciphered [11], but the chloroplast genome of *C. duntsa* has not been reported.

In our current study, we (1) displayed the complete sequence of the chloroplast genome from *C.duntsa* and (2) compared it to other published chloroplast genomes from tea plants to learn about preferred codon usage, repeat structure characteristics, and selection pressure; (3) identified high nucleotide diversity hotspots and (4) created potential DNA markers, and (5) built phylogenetic trees connecting the various species and (6) ascertained the taxonomic position of *C.duntsa*. The results can provide a theoretical foundation for the classification study of *C.duntsa*. Our research also reported vital genetic data for marker development of tea tree plants and species recognition, which provides valuable information for better understanding *Camellia* species chloroplast evolution.

Results

Chloroplast Genome Organization and Features

The sequencing of the entire chloroplast genome of *C.duntsa* has been completed, and the data has been deposited in GenBank under the accession number OL450397. The Illumina NovaSeq 6000 read yielded 25,069,581 Illumina reads (SRA accession number: SRR16934628) and 7,520,874,300 bases for *C.duntsa*. The genome alignment contains 481,746 pair-end reads, the average coverage of loci is 992.6106, and the insert size is 377.62 ± 58.58. The clean data had a GC content of 38.58%, with a Q20 value of 96.29% and a Q30 value of 90.48%. *C.duntsa*'s 157,025 bp chloroplast genome was created by cutting, selecting, and assembling reads. The percentage of GC found in the genome is 37.30%, there are no gaps in the assembly sequence, and there is complete genome coverage by sequencing. According to these findings, the chloroplast DNA of *C.duntsa* sequencing and assembly results were high quality.

The circular chloroplast genome of *C.duntsa* is composed of four conserved sections: a short single copy (SSC) area (18,277 base pairs), a large single copy (LSC) region (86,586 base pairs), and two inverted repeats areas (IRs) (26,081 base pairs) that split the SSC and LSC (Fig. 1). According to the information provided in Table 1, the chloroplast genome of *C.duntsa* contains 135 genes that have been discovered. There are eighty-seven protein-coding genes (PCGs), eight ribosomal RNA genes (rRNAs), thirty-seven transfer RNA genes (tRNAs), and three pseudogene genes (two *ycf15* and one *ycf1*). There are twenty-three genes had introns, eleven of these PCGs (*petB*, *petD*, *atpF*, *ndhA*, *ndhB* (2), *rpoC1*, *rpl16* and *rpl2* (2), *rps16*) contain only one intron, and four PCGs (*rps12*(2), *clpP*, and *ycf*3) have two introns (Table 1).

Table 1
C.duntsa chloroplast genome annotated gene list.

Category	Gene group	Gene name
Photosynthesis	Subunits of photosystem I	psaA,psaB,psaC,psaI,psaJ
	Subunits of photosystem II	psbA,psbB,psbC,psbD,psbE,psbF,psbH,psbI,psbJ,psbK,psbL,psbM,psbN,psbT,psbZ
	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 ATP synthase	atpA,atpB,atpE,atpF*,atpH,atpI
	Large subunit of rubisco	rbcL
Self-replication	Proteins of large ribosomal subunit	rpl14,rpl16*,rpl2*(2),rpl20,rpl22,rpl23(2),rpl32,rpl33,rpl36
	Proteins of small ribosomal subunit	rps11,rps12**(2),rps14,rps15,rps16*,rps18,rps19,rps2,rps3,rps4,rps7(2),rps8
	Subunits of RNA polymerase	rpoA,rpoB,rpoC1*,rpoC2
	Ribosomal RNAs	rrn16(2),rrn23(2),rrn4.5(2),rrn5(2)
	Transfer RNAs	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,trnS-GGA,trnS-UGA,trnT-GGU,trnT-UGU,trnV-GAC(2),trnV-UAC*,trnW-CCA,trnY-GUA,trnfM-CAU
Other genes	Maturase	matK
	Protease	clpP**
	Envelope membrane protein	cemA
	Acetyl-CoA carboxylase	accD
	c-type cytochrome synthesis gene	ccsA
	Translation initiation factor	infA
Genes of unknown function	Conserved hypothetical chloroplast ORF	#ycf1,#ycf15(2),orf42(2),ycf1,ycf2(2),ycf3**,ycf4
Notes:Gene*:Gene with one introns;Gene:Gene with two introns;#Gene:Pseudo gene;Gene(2):Number of copies of multi-copy genes;**		

Chloroplast Genome Comparison

The chloroplast genomes in all tea tree samples were highly conserved, with no alterations in inversion or translocation found. The lengths of the whole chloroplast DNA sequences ranged from 72,756 bp (*Camellia sinensis* var. *Dehungensis)* to 81,408 bp (*Camellia sinensis* var. *sinensis* cv. Fudingdabaicha), the rRNAs lengths ranged from 9,044 bp (*Camellia sinensis* var. *sinensis* cv. Baiye 1) to 9,048 bp (*Camellia sinensis* cv. *Kuntza*, *Camellia sinensis* var. *sinensis*

cv. Huangjinyacha, *Camellia sinensis* var. *sinensis* cv. Zijuanhua, *Camellia sinensis* var. *assamica* cv. Yunkang10). The intergene sections measured 46,269 bp (*Camellia sinensis* var. *sinensis* cv. Wuyi Narcissus) to 55,266 bp (*Camellia sinensis* var. *Dehungensis*), while the intron regions measured 14,532 bp (*Camellia ptilophylla)* to 20,382 bp (*Camellia sinensis* var. *sinensis* cv. Xilian 1) (Table 2).

Table 2
Summary of complete chloroplast genomes for twenty-one species in *Camellia*.

		Total		LSC		SSC		IRa/IRb		CDS		Cis_spliced intron		tRNA
species	Accession number	Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)	Leng (bp)
Camellia sinensis var. *Assamica* cv. *Duntsa*	OL450397	157025	37.30	86586	35.33	18277	30.55	26081	42.95	79413	37.58	20381	37.36	280
Camellia sinensis var. *Assamica*	JQ975030	157162	31.07	86607	35.33	18285	30.51	26135	42.91	79176	30.87	19309	32.61	280
Camellia sinensis cv.*Kuntza*	KC143082	157103	37.31	86645	35.34	18276	30.54	26091	42.95	79698	37.53	20365	37.40	281
Camellia sinensis var. Dehungensis	KJ806279	157110	37.30	86645	35.34	18276	30.54	26091	42.95	72756	37.59	19305	37.28	280
Camellia sinensis var. *Pubilimba*	KJ806280	157086	37.32	86678	35.32	18266	30.6	26071	42.98	80622	37.58	15210	36.42	280
Camellia sinensis var. *sinensis*	KJ806281	157117	37.30	86662	35.31	18275	30.56	26090	42.95	80542	37.58	15192	36.41	280
Camellia sinensis var. *assamica* cv. Yunkang10	MH019307	157100	37.29	86649	35.31	18285	30.51	26083	42.95	79092	37.47	17902	37.91	279
Camellia sinensis var. *sinensis* cv. Anhua	MH042531	157025	37.30	86586	35.33	18277	30.55	26081	42.95	80620	37.57	15196	36.38	280
Camellia sinensis var. *sinensis* cv. Baiye 1	MN086819	156691	37.32	86246	35.35	18279	30.55	26083	42.95	80208	37.52	15377	38.66	266
Camellia sinensis var. *sinensis* cv. Wuyi Narcissus	MT612435	156762	37.30	86301	35.32	18281	30.55	26090	42.94	81318	37.61	19269	37.29	280
Camellia sinensis var. *sinensis* cv. Baijiguan	MT773373	157099	37.30	86643	35.32	18276	30.56	26090	42.95	81234	37.62	19306	37.28	280
Camellia sinensis var. *sinensis* cv. Shuijingui	MT773376	157126	37.30	86654	35.32	18282	30.56	26095	42.94	81357	37.63	19299	37.28	280
Camellia sinensis var. *sinensis* cv. Fudingdabaicha	MZ817088	157025	37.30	86586	35.33	18277	30.55	26081	42.95	81408	37.62	19308	37.26	280
Camellia taliensis	NC022264	156974	37.32	86672	35.32	18436	30.59	25933	43.05	79577	37.57	16947	37.25	281
Camellia grandibracteata	NC024659	157127	37.29	86656	35.32	18285	30.52	26093	42.94	80656	37.56	15205	36.39	280
Camellia ptilophylla	NC038198	157097	37.30	86631	35.32	18286	30.54	26090	42.96	81177	37.58	14532	38.5	267
Camellia sinensis var. sinensis cv. Songzhong	OL690350	157100	37.29	86658	35.31	18282	30.54	26080	42.96	80928	37.58	19315	37.24	280
Camellia sinensis var. sinensis cv. Huangjinye	OL690351	157060	37.30	86634	35.32	18266	30.54	26080	42.96	81093	37.57	19302	37.26	280

		Total		LSC		SSC		IRa/IRb		CDS		Cis_spliced intron		tRNA
Camellia sinensis var. sinensis cv. Huangjinyacha	OL840043	157096	37.29	86653	35.30	18283	30.54	26080	42.96	79743	37.52	20370	37.34	2813
Camellia sinensis var. sinensis cv. Zijuanhua	OL840044	157169	37.29	86662	35.32	18285	30.52	26111	42.94	79785	37.52	20376	37.36	2813
Camellia sinensis var. sinensis cv. Xilian 1	ON060661	157038	37.30	86612	35.32	18282	30.55	26072	42.95	79467	37.55	20382	37.36	2802

Codon Biases

The chloroplast genome of *C.duntsa* has 22,969 codons (65 kinds) that code for 20 amino acids. According to Supplemental Table 1, the amino acid with the highest frequency is leucine (Leu; 10.32%; 2,371 instances), whereas the amino acid with the lowest frequency is cysteine (Cys; 1.10%; 252 instances). The chloroplast genome of *C.duntsa* shows a preference for synonymous codons that end in A or U and have an RSCU greater than one (Supplemental Table 1 and Supplemental Fig. 1).

The frequency of codon usage in the twenty-one tea tree chloroplast genomes, which encode 64 proteins with three stop codons (UAA, UAG, and UGA), is shown in Supplemental Table 1. Thirty of these twenty-one species contain RSCU > 1 codons that end in A or U. The RSCU for the UUA encoding leucine ranged from 1.9805 to 2.0017, the highest among the twenty-one tea species. In addition to *C.duntsa*, the other twenty tea plants expressing arginine CGC had the lowest RSCU of 0.3236. The lowest RSCU encoding serine AGC in *C.duntsa* was 0.3326.

Twenty-one Camellia Species' Inverted Repetitive Sections Contract and Expand

Twenty-one different tea plants' LSC/IRb/SSC/IRa boundary areas were analyzed and contrasted to evaluate the potential effect of IRs on their chloroplast genomes (Fig. 2). The extent of the IR area in those tea species was comparable, ranging from 25,996(*Camellia taliensis*) to 26,135(*Camellia sinensis* var. *Assamica*) base pairs (bp) (Fig. 2).In the present study, the distance of *rpl2* gene from the LSC-IRb region was 106 bp for most of the tea trees. However, the distance of *rpl2* gene from the LSC-IRb region was 149 bp, 129 bp and 112 bp for *Camellia sinensis* var. *Assamica*, *Camellia sinensis* cv. *Kuntza* and *Camellia sinensis* var. *Pubilimba*, respectively. The distances between the *trnN* gene and the SSC-IRa region were 1,381 bp for most of the tea trees, with the shortest distance of 1,346 bp for *Camellia taliensis*. The distances between the *ndhF* gene and the SSC-IRb ranged from 5bp (*Camellia sinensis* var. *pubilimba*) to 164bp (*Camellia sinensis* var. *sinensis*). The distance of *ndhF* gene from SSC-IRb for most of the tea trees was 56 bp. The genes *rps19* and *ycf 1* span the LSC-IRb region and the SSC-IRa region, respectively. In Fig. 2, it can be seen that *Camellia sinensis* var. *Assamica* cv. *Duntsa*, *Camellia sinensis* var. *sinensis* cv. Anhua and *Camellia sinensis* var. *sinensis* cv. Fudingdabaicha have similar IRs boundaries.

Repeat Sequences and Microsatellite Analyses of Chloroplast Genome

The twenty-one tea tree chloroplast genomes in this study had a total of 968 long repetitive sequences involving three types, of which 409 were Forward type, 557 were Palindromic type, and two were Reverse type (Fig. 3A). However, the complement type was not found (Supplemental Table 3). Among them, the Reverse type repeats were found in *Camellia sinensis* cv. *Kuntza* and *Camellia sinensis* var. *sinensis* cv. Baijiguan, respectively. The lengths of these dispersed repeat sequences ranged from 30 bp to 97 bp, with the most common lengths of 30, 38, 42 and 60 bp (Fig. 3B).

Using MISA analysis, a total of 241–249 SSRs were detected in twenty-one tea tree chloroplast genomes (Fig. 4A). Among them, the percentages of A and T single nucleotide repeat sequences were 40.03% and 56.97%, respectively; and the percentages of C and G single nucleotide repeat sequences were 1.73% and 1.27%, respectively (Fig. 4B). Figure 4C showed that the LSC region had the highest SSR frequency (57.61%), followed by the SSC region (20.30%) and then the IR region (22.09%). Figure 4D shows that the spacer region had the highest number of SSRs, followed by coding gene region and finally intron region.

Highly Variable Regions

The chloroplast genomes of *C. duntsa* and its relatives were analyzed for nucleotide polymorphisms using DnaSP software v6.0, and a total of 1,168 variants were found, of which 577 were in the LSC, 375 in the SSC, and 216 in the IRs genome (Table 3). Their nucleic acid variability ranged from 0-0.00729, and the average Pi values of LSC, SSC and IR distributions were 0.00010, 0.0018 and 0.00039, respectively (Table 3). Sliding window analysis showed that six intergenic regions: *psbI-atpA*, *rsp4-ndhJ*, *ycf4-cemA*, *psbH-petB*, *ycf15-ndhB*, and *ndhF-rpl32*, and three genes, *rps19*, *rps15*, and *ycf1*, exhibited a high level of variation (Fig. 5). Table 4 shows that the size of these nine regions varied greatly from 124 base pairs (*psbH-petB*) to 5,628 base pairs (*ycf1*), with Pi values ranging from 0.00316 (*rps19*) to 0.00728 (*ndhF-rpl32*).

Table 3
Analyses of variable sites in chloroplast genomes of *Camellia*.

	Number of sites	Number of variable sites	Number of informative sites	Nucleotide diversity
LSC region	87103	577	177	0.00101
SSC region	18326	375	71	0.00183
IRs region	52284	216	64	0.00078
Complete cp genome	157713	1168	312	0.00091

Table 4
Nine hypervariable regions of chloroplast genomes of *Camellia*.

No	region	Length (bp)	Variable sites	Parsimony informative sites	Nucleotide diversity (Pi)
1	psbI-atpA	2,118	52	10	0.00441
2	rps4-ndhJ	3,139	40	9	0.00372
3	ycf4-cemA	900	15	3	0.00378
4	psbH-petB	124	16	1	0.00508
5	rps19	285	6	5	0.00316
6	ycf15-ndhB	1,007	8	6	0.00351
7	ndhF-rpl32	683	21	10	0.00728
8	rps15	273	4	3	0.00355
9	ycf1	5,628	73	24	0.00430
	total	14,157	235	71	0.03879

Estimation of Genomic Selection Pressure in Tea Plants Chloroplast

In this study, using *C.duntsa* as a reference, the values of non-synonymous to synonymous substitution (Ka/Ks) ratio of 20 tea chloroplast genes was analyzed (Supplemental Table 5). Most of the PCGs exhibited a Ka/Ks ratio value much smaller than 1.00, indicating a purifying selection pattern. However, the Ka/Ks values of *rpoC1* in *Camellia sinensis* var. *Assamica*, *Camellia sinensis* var. *sinensis* cv. Anhua, *Camellia sinensis* var. *sinensis* cv. Songzhong, *and Camellia sinensis* var. *sinensis* cv. Xilian 1; *ycf1* in *Camellia sinensis* var. *sinensis*, *Camellia sinensis* var. *sinensis* cv. *Baijiguan, Camellia sinensis* var. *sinensis* cv. Songzhong, and *Camellia sinensis* var. *sinensis* cv. Huangjinyacha; *petB* in *Camellia sinensis* var. *sinensis* cv. Anhua and *Camellia ptilophylla*; *ndhD* in *Camellia ptilophylla*; *petD* in *Camellia grandibracteata*, r*pl16* in *Camellia sinensis* var. *sinensis* cv. Anhua, and *rpoC2* in *Camellia ptilophylla* were greater than 1.00, demonstrating that these genes were under diversifying selection pattern and more sensitive to the environment pressure (Fig. 6).

Phylogenetic Relationships of Camellia Species

The Bayesian Inference (BI) and Maximum Likelihood (ML) trees showed that all the *Camellia* species clustered into two clades (subgen. *Thea* and subgen. *Camellia)* by high bootstrap support. Tea trees from *Camellia taliensis*, *Camellia sinensis* var. *Pubilimba*, *Camellia sinensis* var. *Assamica*, and *Camellia sinensis* var. *sinensis* were separated on different branches, whereas *C. duntsa* was associated with Chinese type *Camellia sinensis* var. *sinensis* cv. Anhua, *Camellia sinensis* var. *sinensis* cv. Fudingdabaicha, and *Camellia ptilophylla*, indicating that its chloroplast genome is more closely related to that of the Chinese type tea tree (Fig. 6).

Discussion

In this study, we obtained the complete chloroplast genome of *C. duntsa* and analyzed it in comparison with the chloroplast genomes of other tea trees. The results showed that the full sequence length of the chloroplast genome of *C. duntsa* was very similar to the data of other members of the genus *Camellia* [15, 16].The complete chloroplast genome of tea tree shows a distinct tetrameric structure that the lengths of the whole chloroplast DNA sequences ranged from 72,756 bp (*Camellia sinensis* var. Dehungensis) to 81,408 bp (*Camellia sinensis* var. *sinensis* cv. Fudingdabaicha), the rRNAs lengths ranged from 9,044 bp (*Camellia sinensis* var. *sinensis* cv. Baiye 1) to 9,048 bp (*Camellia sinensis* cv. Kuntza, *Camellia sinensis* var. *sinensis* cv. Huangjinyacha, *Camellia sinensis* var. *sinensis* cv. Zijuanhua, *Camellia sinensis* var. *assamica* cv. Yunkang10). The GC content of angiosperm chloroplast genomes usually ranges from 34–40%, and the GC content of tea tree chloroplast genomes analyzed in this study was relatively stable at around 37.30% (Table 2). The GC content is crucial for the transmission of genetic information, as changes in base composition lead to variations in the composition of the chloroplast genome among different species [17, 18]. These findings provide strong evidence for the highly conserved and slow evolution of the chloroplast genome in this genus. This inference was further verified by an in-depth analysis of the contraction and expansion at the boundaries of the inverted repeat region (Fig. 2) and chloroplast genome comparison. The results of this study were analyzed more comprehensively in terms of RSCU analysis, IR boundaries, repeat sequences, SSR analysis, and sliding window analysis. These analyses provided a more comprehensive understanding of the chloroplast genome of tea tree.

Preferential codon usage is the result of long-term adaptation and evolution of species to their environment and is influenced by a variety of factors. RSCU values reflect the codon usage patterns of different genes, with higher RSCU values indicating a higher frequency of codon usage. It has been suggested that

plant synonymous codon usage bias is related to the number of introns, which may be due to DNA methylation [19]. The results of this study showed that 28 of the 31 codons with RSCU values greater than 1 ended in A or U (Supplemental Table 1), which is consistent with other plant chloroplast genome codon usage preferences [20, 21]. In the IR boundary analysis, high chloroplast genome sequence similarity was found in *C. duntsa*, *Camellia sinensis* var. *sinensis* cv. Anhua, and *Camellia sinensis* var. *sinensis* cv. Fudingdabaicha, suggesting that chloroplast genetics among these species are backgrounds of these species are closely related to each other. This conclusion was also verified in the constructed evolutionary tree.

The number and type of SSRs are species-dependent and can be used for species identification, phylogenetic analysis and molecular-assisted breeding [22, 23]. The tea tree chloroplast genome analyzed in this study revealed 241–248 SSR loci, including single nucleotide, dinucleotide, trinucleotide, tetranucleotide and hexanucleotide repeat sequences. These loci were mainly in the form of A/T and AT/TA, which may be related to the fact that A/T bases are more likely to be mutated than G/C bases (Wang et al. 1984). The twenty-one tea plant species exhibit notable differences in SSR markers within specific gene regions (Fig. 4). For example *Camellia sinensis* var. *sinensis* cv. Xilian 1 lacks AT and TA repeats, *Camellia sinensis* var. *sinensis* cv. Wuyi Narcissus lacks TGC repeats, *Camellia taliensis* lacks AAAT repeats, *Camellia sinensis* var. *sinensis* cv. Shuijingui lacks AGAT repeats, *Camellia sinensis* cv. *Kuntza* lacks TCTT repeats, and *Camellia sinensis* var. *Assamica*, *Camellia sinensis* var. *assamica* cv. Yunkang10, *Camellia grandibracteata*, and *Camellia sinensis* var. *sinensis* cv. Zijuanhua lack CCCT repeats. Additionally, *Camellia sinensis* var. *Assamica*, *Camellia sinensis* var. *assamica* cv. Yunkang10, and *Camellia sinensis* var. *sinensis* cv. Zijuanhua lack GAGG repeats, while *Camellia taliensis* lacks AAAAAG and CTTTTT repeats. Thus, these SSR loci could be further developed as molecular markers for the chloroplast genome of tea leaves.

Mutational hotspots in chloroplast genomes have been shown to be potential molecular markers for phylogenetic relationship analysis and identification of closely related species [24, 25]. In this study, nine regions with high Pi values were identified using DnaSP software v6.0, and these markers were recognized as potentially divergent regions in the complete chloroplast genomes of 24 species of *Camellia* sinensis, providing a basis for subsequent molecular identification. Among them, the *ndhF-rpl32* region has the highest Pi value, and *ycf1* also has high variability in the chloroplast genome, which can be used as molecular markers for subsequent studies. In addition, the divergent regions were mainly distributed in the LSC and SSC regions, which was in agreement with the results of others' studies.The analysis of Ka/Ks ratios indicated that the six mutation hotspots (*rpoC1, ycf1, petB, ndhD, rpl16, rpoC2*) were diversifying and were more sensitive to the environmental changes (Fig. 6).

Camellia, a genus in the *Theaceae* family with over 280 species, has a relatively high economic and ecological value[15, 26]. Among the 110 taxa that form the *Camellia* genus, the *Camellia sinensis* (L.) O. Kuntze species is the most prominent tea plant. *Camellia sinensis* var. *sinensis*, known as the Chinese type, and *Camellia sinensis* var. *assamica*, known as the Assam type, are the two primary groups to which cultivated cultivars of the tea plant belong. Previous research suggests that *C. duntsa* is a transitional tea tree resource with relatively considerable genetic variation in the tea tree population (Chen et al.1989). Comparative analysis of complete chloroplast genome sequences has proven to be a reliable tool to support plant phylogeny and taxonomy [27, 28]. In order to clarify the phylogenetic relationships between *C. duntsa* and other tea trees, a phylogenetic tree was constructed in this study using the chloroplast genome data of 23 published tea trees, and *Polyspora hainanensis*, *Coffea arabica*, and *Apterosperma oblata* were selected as outgroups. Based on the support values of the phylogenetic tree (Fig. 7), the genus *Camellia* was divided into 2 branches in this study, *C. duntsa* was placed in subgen. *Thea*. *C. duntsa* shows a relatively close evolutionary relationship with *Camellia sinensis* var. *sinensis* cv. Anhua, *Camellia sinensis* var. *sinensis* cv. Fudingdabaicha, and *Camellia ptilophylla*.

Conclusions

In this study, we sequenced, annotated and analyzed the complete chloroplast genome sequence of *C. duntsa*. Combined with the published chloroplast genome sequences of other *Camellia sinensis* genera, we performed comparative genome analysis and constructed a phylogenetic tree to illustrate the phylogenetic relationships between *C. duntsa* and other species within the genus. This study lays the foundation for species identification of the genus *Camellia*, provides more comprehensive and in-depth information for the study of the chloroplast genome of *C. duntsa*, and contributes to the development of new conservation and management strategies for the germplasm resources of the *C. duntsa* tea tree, as well as to the conservation of the species.

Methods and Materials

Fresh Leaves Sample and DNA Sequencing

The *C.duntsa* tea trees grown under natural environmental conditions in Chengbu Miao Autonomous County, Hunan Province ,China (26°05'52 "N, 110°28'78 "E) were selected, and the fresh leaves of one bud and two leaves were collected in spring and immediately fixed in liquid nitrogen. After returning to the laboratory, the samples were ground into powder and stored at -80°C for experimental analysis. Genomic DNA extraction using CTAB method [12]. After passing quality control (QC), the genomic DNA was mechanically fragmented by ultrasound. The DNA fragments obtained were purified, end-repaired, A-added at the 3' end, connected to sequencing junctions, and then fragment size-selected by agarose gel electrophoresis, and PCR-amplified to form sequencing libraries, which were first subjected to library QC, and then subjected to bipartite (PE) sequencing using the Illumina NovaSeq 6000 platform, with a sequencing read length of 150.

Genome Assembly and Annotation

FastQC software v0.20.0 filters the raw data. The filtering criteria were as follows: (1) truncate the sequencing linker and primer sequences in the Reads; (2) filter out the Reads whose average quality value was less than Q 5; (3) filter out N Reads whose number was more than 5. Using Bowtie2 very-sensitive-local pattern ratios, the original sequencing data were compared with a self-built chloroplast genome database, and the aligned sequence was used as the chloroplast genome sequence of the project sample. Assembly using Spades v3.10.1 [13], the chloroplast genome was annotated in two ways to improve annotation accuracy. First, prodigal v2.6.3 was used to annotate chloroplast PCGs, hmmer v3.1b2 software was used to predict rRNA, and aragorn v1.2.38

was used to predict tRNA. Secondly, the second annotation result was obtained by extracting the gene sequence of relative species published on NCBI, and then using blast v2.6 to compare the assembled sequence. The two annotation results were then manually checked for different genes, and the wrong and redundant annotations were removed, and the multi-exonic boundary was determined to obtain the final annotation.

Comparison and Marker Identification

The IRscope v2.01 software analyzed the boundaries of inverted repeat regions (IRs) and their contraction and extension for twenty-one chloroplast genomes. We analyzed the chloroplast genomes of *Camellia* species by performing pairwise alignments in m-VISTA in LAGAN mode using the *C. duntsa* reference sequence. DnaSP software v6.0 was utilized in order to determine the LSC, SSC, and IR regions, as well as the nucleotide diversity (Pi), for the *Camellia* species. A sliding window analysis with a 600-bp window and 200-bp increments was utilized to plot Pi values. Using DnaSP software v6.0, we calculated Non-synonymous (Ka)/synonymous (Ks) rates and found that the rates of evolution for twenty-nine PCGs varied significantly among the twenty species.

Relative Codon Usage Analysis

The analysis of relative synonymous codon usage (RSCU) in the chloroplast genome coding genes was conducted using the cloud service (cloud.genepioneer.com), with default parameter settings.

Repeats Identification

The forward, reverse, palindromic, and complementary repeat sequences were identified using REPuter (https://bibiserv.cebitec.unibielefeld.de/reputer/) with the following parameters: a hamming distance of three, a minimal repeat size of 30 bp, and an e-value of less than 1e-05. Additionally, simple sequence repeats (SSRs) of the assembled chloroplast genomes were identified using the website MISA (https://webblast.ipk-gatersleben.de/misa/) with the minimum numbers of mono-, di-, tri-, tetra-, penta-, and hexanucleotides set to 10, 5, 4,3, 3, and 3, respectively. In consideration of the possibility of dispersed repeats distributed across multiple chromosomes, we artificially concatenated the chromosomes in order to identify dispersed repeats and then divided the results.

Phylogenetic Analyse

To clarify the phylogenetic position of *C.duntsa* within the genus *Camellia*, we performed a phylogenetic analyses on twenty-five tea trees chloroplast genomes(*Polyspora hainanensis*, *Coffea arabica*, and *Apterosperma oblata* as outgroups). The chloroplast sequences were aligned using MAFFT, a utility in GENEIOUS R8, with the default parameters.Use software RAxML [14] v7.2.8 select GTRGAMMA model, bootstrap = 1000, build the maximum likelihood evolutionary tree. Bayesian analysis was performed using MrBayes v3.2.7, Markov Chain Monte Carlo (MCMC) iterative operation for 1 million generations, sampling every 100 generations.

Declarations

Acknowledgments

The authors thank Guo-Li Xiao, Da-Yong Liu, Wen-Hua Zhu, and Rong-Fu Yang for their help with sample collection.

Author Contributions

JL, L-ZX, and HL: conceived and designed the study. X-YQ, HT, and JT: database gathering. X-YQ, YQ, and T-TL analyzed the data.. JL, and X-YQ collected the sample. JL wrote the manuscript. L-ZX and HL: revised the paper. The final, published version of the work has been read and approved by all of the writers. All authors listed, have madesubstantial, direct and intellectual contribution to the work, and approved it for publication.

Funding

This work was supported by the Modern Tea Industry Technology System of Hunan Province (Xiang Cai Nong Index [2020] No. 65).

Availability of data and materials

The newly sequenced and annotated chloroplast genome has been submitted into NCBI (https://www.ncbi.nlm.nih.gov). Accession numbers: OL450397.

Ethics approval and consent to participate

Not applicable. No specific permits were required for the collection of specimens for this study. This research was carried out in compliance with the relevant laws and all methods were performed in accordance with the relevant guidelines and regulations.

Conflict of Interest

The authors state that they have no commercial or financial affiliations that could be perceived as a potential conflict of interest in the research that they conducted, and that is why it was conducted without any conflicts of interest.

References

1. Li J, Tang H, Luo H, Tang J, Zhong N, Xiao L. Complete mitochondrial genome assembly and comparison of Camellia sinensis var. Assamica cv. Duntsa. Front Plant Sci. 2023;14.

2. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97.
3. Daniell H, Lin C-S, Yu M, Chang W-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17:134.
4. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. 2011;28:583–600.
5. Palmer JD, Stein DB. Conservation of chloroplast genome structure among vascular plants. Curr Genet. 1986;10:823–33.
6. Maier RM, Neckermann K, Igloi GL, Kössel H. Complete Sequence of the Maize Chloroplast Genome: Gene Content, Hotspots of Divergence and Fine Tuning of Genetic Information by Transcript Editing. J Mol Biol. 1995;251:614–28.
7. Li H-T, Luo Y, Gan L, Ma P-F, Gao L-M, Yang J-B, et al. Plastid phylogenomic insights into relationships of all flowering plant families. BMC Biol. 2021;19:232.
8. Chen X, Zhou J, Cui Y, Wang Y, Duan B, Yao H. Identification of Ligularia Herbs Using the Complete Chloroplast Genome as a Super-Barcode. Front Pharmacol. 2018;9.
9. Jiao L, Lu Y, He T, Li J, Yin Y. A strategy for developing high-resolution DNA barcodes for species discrimination of wood specimens using the complete chloroplast genome of three Pterocarpus species. Planta. 2019;250:95–104.
10. Tian X, Ye J, Song Y. Plastome sequences help to improve the systematic position of trinerved Lindera species in the family Lauraceae. PeerJ. 2019;7:e7662.
11. Huang H, Shi C, Liu Y, Mao S-Y, Gao L-Z. Thirteen Camelliachloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol. 2014;14:151.
12. Fu Z, Song J, Jameson PE. A rapid and cost effective protocol for plant genomic DNA isolation using regenerated silica columns in combination with CTAB extraction. J Integr Agric. 2017;16:1682–8.
13. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11.
14. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
15. Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z. Comparative Chloroplast Genomes of Camellia Species. PLoS ONE. 2013;8:e73053.
16. Peng J, Zhao Y, Dong M, Liu S, Hu Z, Zhong X, et al. Exploring the evolutionary characteristics between cultivated tea and its wild relatives using complete chloroplast genomes. BMC Ecol Evo. 2021;21:71.
17. Gu C, Ma L, Wu Z, Chen K, Wang Y. Comparative analyses of chloroplast genomes from 22 Lythraceae species: inferences for phylogenetic relationships and genome evolution within Myrtales. BMC Plant Biol. 2019;19:281.
18. Zhu H, Liu J, Li H, Yue C, Gao M. Complete Chloroplast Genome Structural Characterization and Comparative Analysis of Viburnum japonicum (Adoxaceae). Forests. 2023;14:1819.
19. Qin Z, Cai Z, Xia G, Wang M. Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants. BMC Genomics. 2013;14:56.
20. Wang X, Zhou T, Bai G, Zhao Y. Complete chloroplast genome sequence of Fagopyrum dibotrys: genome features, comparative analysis and phylogenetic relationships. Sci Rep. 2018;8:12379.
21. Chen Y, Hu N, Wu H. Analyzing and Characterizing the Chloroplast Genome of *Salix wilsonii*. Biomed Res Int. 2019;2019:e5190425.
22. Yang A-H, Zhang J-J, Yao X-H, Huang H-W. Chloroplast microsatellite markers in Liriodendron tulipifera (Magnoliaceae) and cross-species amplification in L. chinense. Am J Bot. 2011;98:e123–6.
23. Chen X, Cui Y, Nie L, Hu H, Xu Z, Sun W, et al. Identification and Phylogenetic Analysis of the Complete Chloroplast Genomes of Three *Ephedra* Herbs Containing Ephedrine. Biomed Res Int. 2019;2019:e5921725.
24. Chi X, Wang J, Gao Q, Zhang F, Chen S. The Complete Chloroplast Genomes of Two Lancea Species with Comparative Analysis. Molecules. 2018;23:602.
25. Sun J, Wang Y, Liu Y, Xu C, Yuan Q, Guo L, et al. Evolutionary and phylogenetic aspects of the chloroplast genome of Chaenomeles species. Sci Rep. 2020;10:11466.
26. Vijayan K, Zhang W-J, Tsou C-H. Molecular taxonomy of Camellia (Theaceae) inferred from nrITS sequences. Am J Bot. 2009;96:1348–60.
27. Yang J, Vázquez L, Chen X, Li H, Zhang H, Liu Z et al. Development of Chloroplast and Nuclear DNA Markers for Chinese Oaks (Quercus Subgenus Quercus) and Assessment of Their Utility as DNA Barcodes. Front Plant Sci. 2017;8.
28. Li Y, Sylvester SP, Li M, Zhang C, Li X, Duan Y, et al. The Complete Plastid Genome of Magnolia zenii and Genetic Comparison to Magnoliaceae species. Molecules. 2019;24:261.

Figures

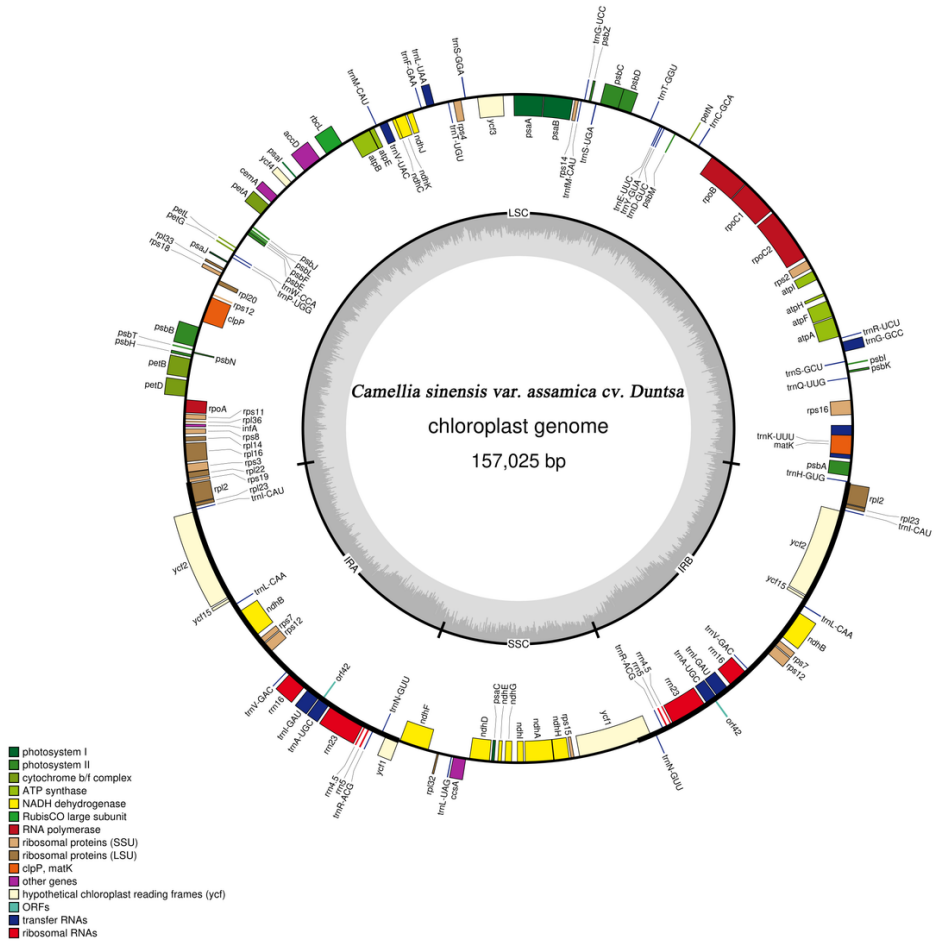


Figure 1

Chloroplast genome structural map of *C.duntsa*.

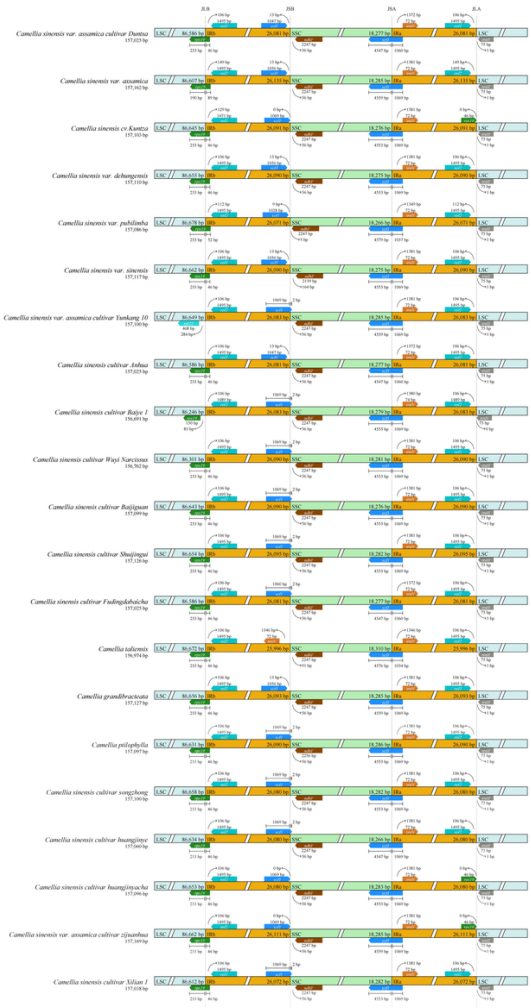


Figure 2

Chloroplast genome junctions in twenty-one *Camellia* species. The LSC, SSC, and IR scales are not to scale in the illustration.

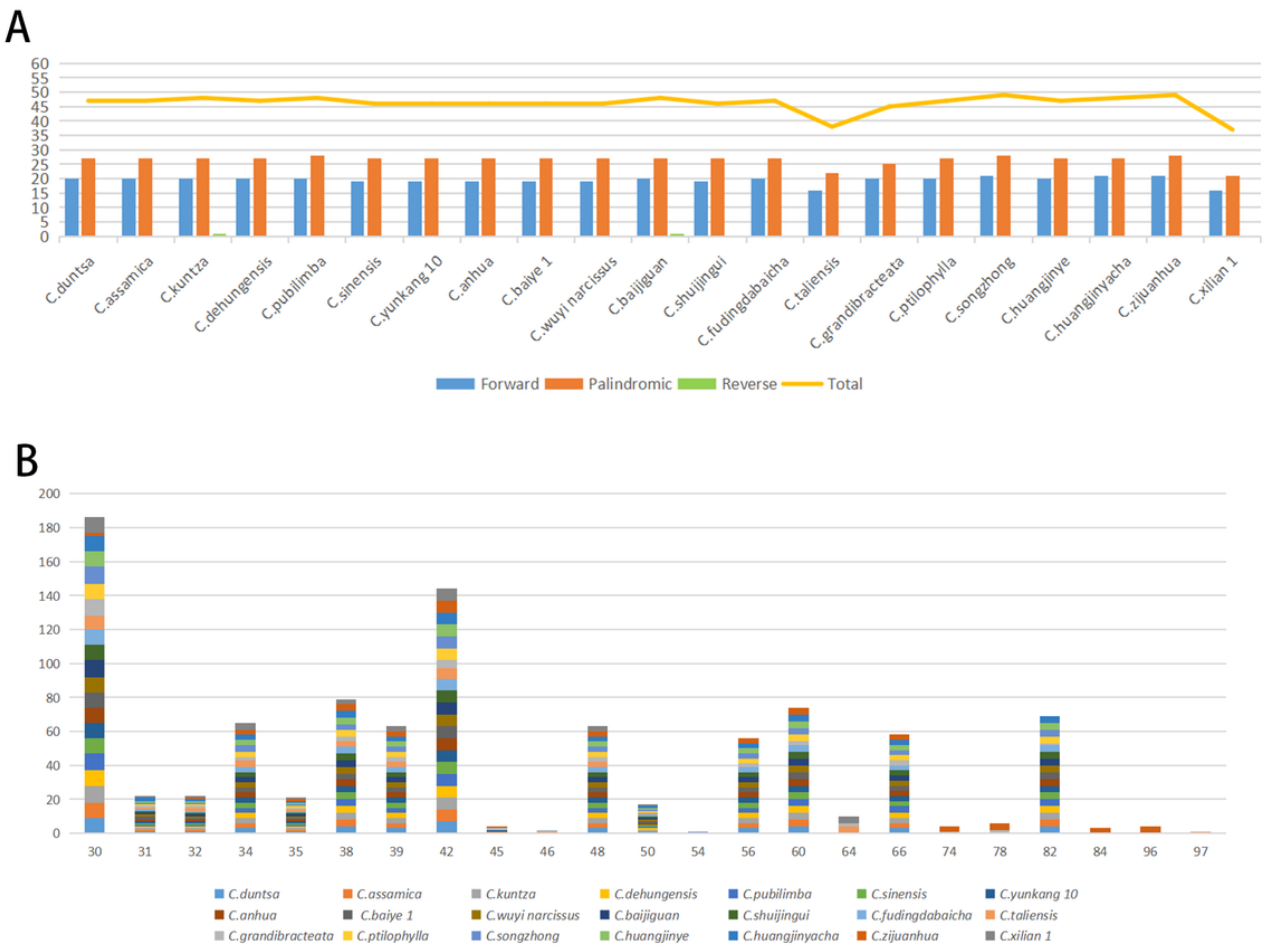


Figure 3

Number of long repetitive repeats on the complete chloroplast genome sequence of 21 tea tree species; (A) frequency of repeat type; (B) frequency of the repeats.

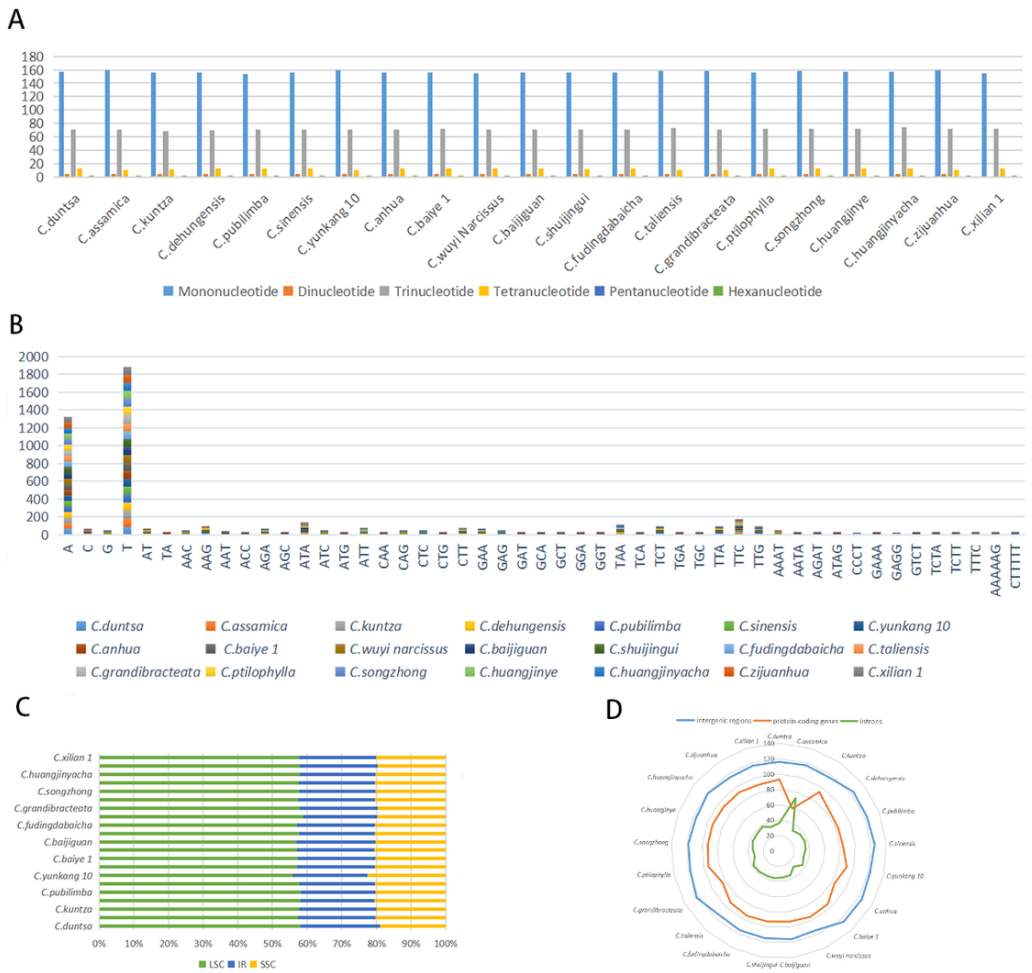


Figure 4

The comparison of SSR distribution in *Camellia*species chloroplast genomes; (A) number of different SSR types; (B) frequency of common motifs; (C) frequency of SSRs in the LSC, IR, SSC region; (D) frequency of SSRs in the intergenic regions, protein-coding genes and introns.

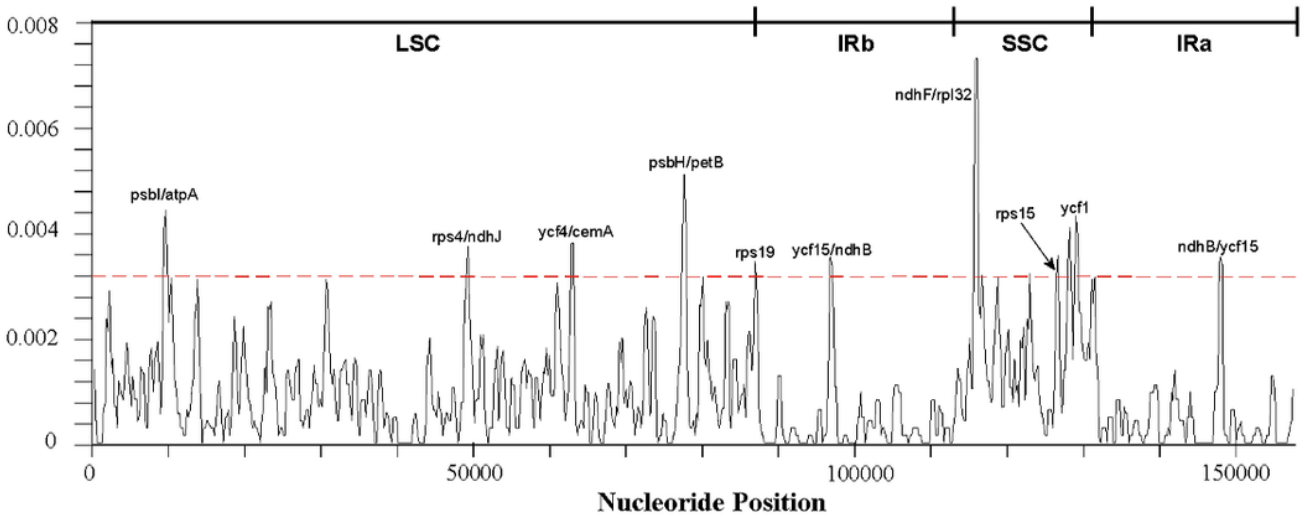


Figure 5

Comparison of nucleotide diversity (Pi) values among the *Camellia* species. X-axis, position of the midpoint of each window; Y -axis, nucleotide diversity (Pi) of each window.

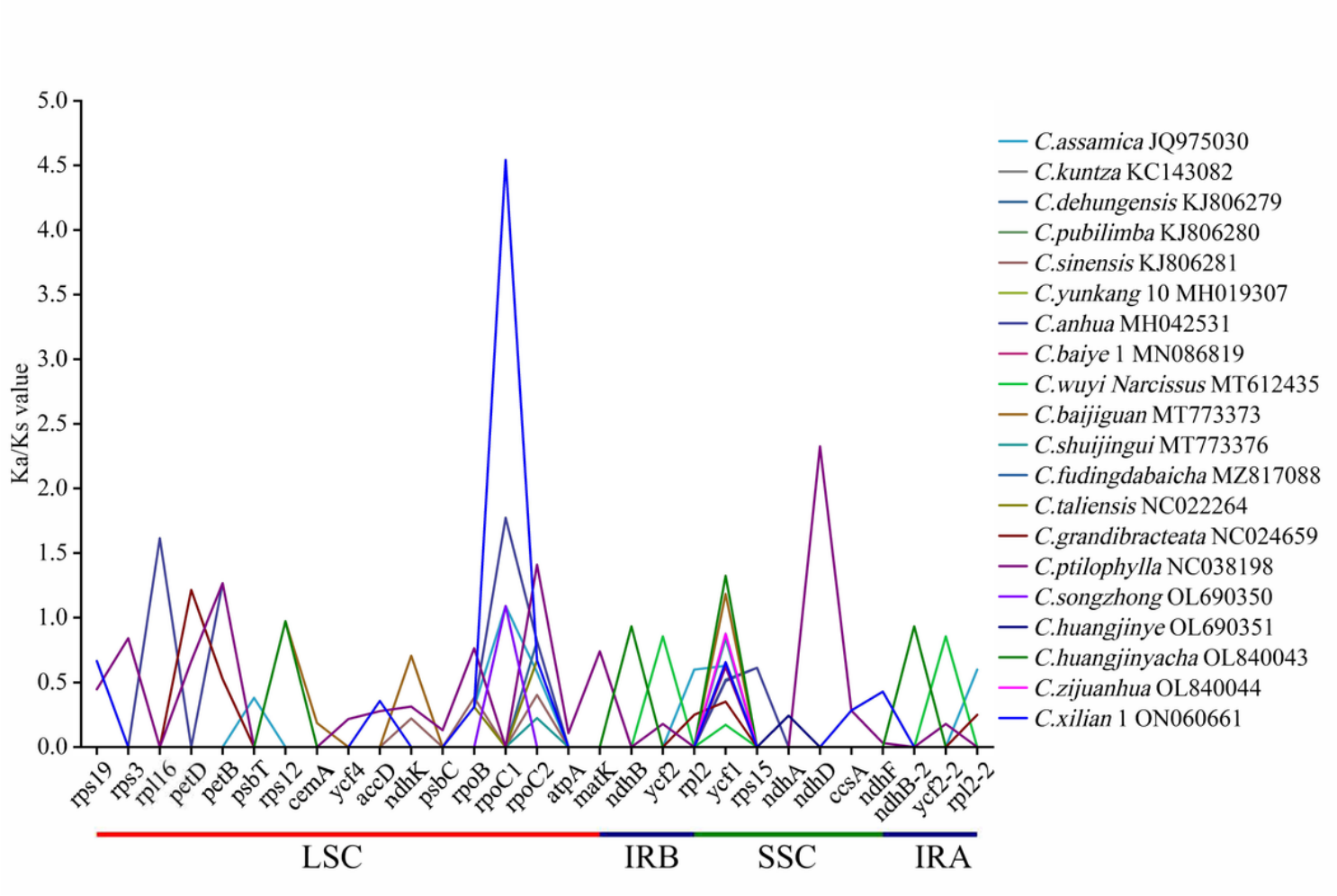


Figure 6

Ka/Ks ratios of 29 PCGs in tea tree species chloroplast genomes

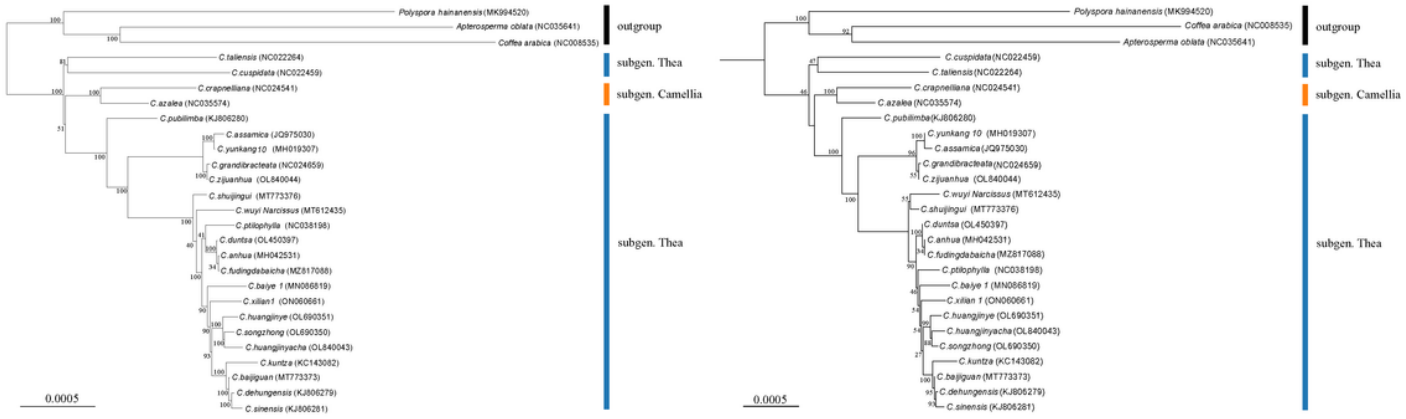


Figure 7

Phylogenetic relationships of *Camellia* species taxa inferred from Bayesian Inference (BI) and Maximum Likelihood (ML) analyses of protein-coding genes. Polyspora hainanensis, Coffea arabica, and Apterosperma oblata was used as the outgroup. (A)Bayesian phylogenies. (B) Maximum-likelihood phylogenies.

Supplementary Files

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

- SupplementaryMaterial.zip