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

Jin Li
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 genomes, 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 (tpoC1, 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 ptlichyllia. The results can provide valuable information for better understanding Camellia species 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 Chengdu 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. fiscerii and L. jaenaensis 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 Camphorae species using chloroplast whole-genome and nuclear gene data from 191 individuals of 25 genera and 133 species of Camphoraeae. 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 *ycf3*) have two introns (Table 1).

### Table 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Gene group</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis</td>
<td>Subunits of photosystem I</td>
<td>psaA, psaB, psaC, psaL, psaJ</td>
</tr>
<tr>
<td></td>
<td>Subunits of cytochrome b/f complex</td>
<td>petA, petB*, petD*, petG, petL, petN</td>
</tr>
<tr>
<td></td>
<td>Subunits of ATP synthase</td>
<td>atpA, atpB, atpE, atpF*, atpH, atpl</td>
</tr>
<tr>
<td></td>
<td>Large subunit of ribulose</td>
<td>rbcL</td>
</tr>
<tr>
<td>Self-replication</td>
<td>Proteins of large ribosomal subunit</td>
<td>rpl14, rpl16*, rpl2*(2), rpl20, rpl22, rpl23(2), rpl32, rpl33, rpl36</td>
</tr>
<tr>
<td></td>
<td>Proteins of small ribosomal subunit</td>
<td>rps11, rps12**(2), rps14, rps15, rps16*, rps18, rps19, rps2, rps3, rps4, rps7(2), rps8</td>
</tr>
<tr>
<td></td>
<td>Subunits of RNA polymerase</td>
<td>rpoA, rpoB, rpoC1*, rpoC2</td>
</tr>
<tr>
<td></td>
<td>Ribosomal RNAs</td>
<td>rm16(2), rm23(2), rm4.5(2), rm5(2)</td>
</tr>
<tr>
<td>Other genes</td>
<td>Maturase</td>
<td>matK</td>
</tr>
<tr>
<td></td>
<td>Protease</td>
<td>clpP**</td>
</tr>
<tr>
<td></td>
<td>Envelope membrane protein</td>
<td>cemA</td>
</tr>
<tr>
<td></td>
<td>Acetyl-CoA carboxylase</td>
<td>accD</td>
</tr>
<tr>
<td></td>
<td>c-type cytochrome synthesis gene</td>
<td>ccsA</td>
</tr>
<tr>
<td></td>
<td>Translation initiation factor</td>
<td>infA</td>
</tr>
<tr>
<td>Genes of unknown function</td>
<td>Conserved hypothetical chloroplast ORF</td>
<td>#ycf1, #ycf15(2), orf42(2), ycf1, ycf2(2), ycf3**, ycf4</td>
</tr>
</tbody>
</table>

**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>
<thead>
<tr>
<th>species</th>
<th>Accession number</th>
<th>Length (bp)</th>
<th>GC (%)</th>
<th>Length (bp)</th>
<th>GC (%)</th>
<th>Length (bp)</th>
<th>GC (%)</th>
<th>Length (bp)</th>
<th>GC (%)</th>
<th>Cis_spliced intron (bp)</th>
<th>tRN/ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Camellia sinensis</em> var. Assamica cv. Dunsia</td>
<td>OL450397</td>
<td>157025</td>
<td>37.30</td>
<td>86586</td>
<td>35.33</td>
<td>18277</td>
<td>30.55</td>
<td>26081</td>
<td>42.95</td>
<td>79413</td>
<td>37.58</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. Assamica cv. Kunta</td>
<td>JQ975030</td>
<td>157162</td>
<td>31.07</td>
<td>86607</td>
<td>35.33</td>
<td>18285</td>
<td>30.51</td>
<td>26135</td>
<td>42.91</td>
<td>79176</td>
<td>30.87</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. Dehunensis</td>
<td>KC143082</td>
<td>157103</td>
<td>37.31</td>
<td>86645</td>
<td>35.34</td>
<td>18276</td>
<td>30.54</td>
<td>26091</td>
<td>42.95</td>
<td>79698</td>
<td>37.53</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. Pubilimba</td>
<td>KJ806279</td>
<td>157086</td>
<td>37.32</td>
<td>86678</td>
<td>35.32</td>
<td>18266</td>
<td>30.6</td>
<td>26071</td>
<td>42.98</td>
<td>80622</td>
<td>37.58</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis</td>
<td>KJ806281</td>
<td>157117</td>
<td>37.30</td>
<td>86662</td>
<td>35.31</td>
<td>18275</td>
<td>30.56</td>
<td>26090</td>
<td>42.95</td>
<td>80542</td>
<td>37.58</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. assamica cv. Yunkang10</td>
<td>MH019307</td>
<td>157100</td>
<td>37.29</td>
<td>86649</td>
<td>35.31</td>
<td>18285</td>
<td>30.51</td>
<td>26083</td>
<td>42.95</td>
<td>79092</td>
<td>37.47</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Anhua</td>
<td>MH042531</td>
<td>157025</td>
<td>37.30</td>
<td>86586</td>
<td>35.33</td>
<td>18277</td>
<td>30.55</td>
<td>26081</td>
<td>42.95</td>
<td>80620</td>
<td>37.57</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Baiye 1</td>
<td>MN086819</td>
<td>155691</td>
<td>37.32</td>
<td>86246</td>
<td>35.35</td>
<td>18279</td>
<td>30.55</td>
<td>26083</td>
<td>42.95</td>
<td>80208</td>
<td>37.52</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Wuyi Narcissus</td>
<td>MT612435</td>
<td>156762</td>
<td>37.30</td>
<td>86301</td>
<td>35.32</td>
<td>18281</td>
<td>30.55</td>
<td>26090</td>
<td>42.94</td>
<td>81318</td>
<td>37.61</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Baijiguan</td>
<td>MT773373</td>
<td>157099</td>
<td>37.30</td>
<td>86643</td>
<td>35.32</td>
<td>18276</td>
<td>30.56</td>
<td>26090</td>
<td>42.95</td>
<td>81234</td>
<td>37.62</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Shuangui</td>
<td>MT773376</td>
<td>157126</td>
<td>37.30</td>
<td>86654</td>
<td>35.32</td>
<td>18282</td>
<td>30.56</td>
<td>26095</td>
<td>42.94</td>
<td>81357</td>
<td>37.63</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Fudingdabaicha</td>
<td>MZ817088</td>
<td>157025</td>
<td>37.30</td>
<td>86586</td>
<td>35.33</td>
<td>18277</td>
<td>30.55</td>
<td>26081</td>
<td>42.95</td>
<td>81408</td>
<td>37.62</td>
</tr>
<tr>
<td><em>Camellia taiensis</em></td>
<td>NC022264</td>
<td>156974</td>
<td>37.32</td>
<td>86672</td>
<td>35.32</td>
<td>18436</td>
<td>30.59</td>
<td>25933</td>
<td>43.05</td>
<td>79577</td>
<td>37.57</td>
</tr>
<tr>
<td><em>Camellia grandibracteata</em></td>
<td>NC024659</td>
<td>157127</td>
<td>37.29</td>
<td>86656</td>
<td>35.32</td>
<td>18285</td>
<td>30.52</td>
<td>26093</td>
<td>42.94</td>
<td>80656</td>
<td>37.56</td>
</tr>
<tr>
<td><em>Camellia ptilophylla</em></td>
<td>NC038198</td>
<td>157097</td>
<td>37.30</td>
<td>86631</td>
<td>35.32</td>
<td>18286</td>
<td>30.54</td>
<td>26090</td>
<td>42.96</td>
<td>81177</td>
<td>37.58</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Songzhong</td>
<td>OL690350</td>
<td>157100</td>
<td>37.29</td>
<td>86658</td>
<td>35.31</td>
<td>18282</td>
<td>30.54</td>
<td>26080</td>
<td>42.96</td>
<td>80928</td>
<td>37.58</td>
</tr>
<tr>
<td><em>Camellia sinensis</em> var. sinensis cv. Huangjinye</td>
<td>OL690351</td>
<td>157060</td>
<td>37.30</td>
<td>86634</td>
<td>35.32</td>
<td>18266</td>
<td>30.54</td>
<td>26080</td>
<td>42.96</td>
<td>81093</td>
<td>37.57</td>
</tr>
</tbody>
</table>


### 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 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* 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 *C. duntsa*, *C. duntsa* var. *Assamica*, *C. duntsa* var. *Kuntza* and *C. duntsa* var. *Publiclimba*, 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 *C. 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 *ycf1* span the LSC-IRb region and the SSC-IRa region, respectively. In Fig. 2, it can be seen that *Camellia sinensis* var. *Assamica*, *Duntsa*, *C. sinensis* var. *sinensis*, *Anhua* and *C. 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, 203 were Complement type and 203 were Reverse type (Fig. 3). Among them, the Reverse type repeats were found in *Camellia sinensis* cv. *Kuntza* and *C. sinensis* var. *sinensis* cv. *Bajiguan*, 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 0.000729, 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: *psbl-atpA*, *rps4-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*).
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 sliding window analysis. These analyses provided a more comprehensive understanding of the chloroplast genome of tea tree.

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.

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.

Preference 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 talienensis lacks AAAT repeats, Camellia sinensis var. sinensis cv. Shujingui 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 talienensis lacks AAAAG 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 Apterosperrma 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 was used 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 subject 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, hmmr 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.uni-bielefeld.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 GENIEUS 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 made substantial, 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**


**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**


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 *Apterasperma 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