Exploring evolution characteristic between cultivated tea and its wild relatives using complete chloroplast genomes

Jiao Peng

Central South University of Forestry and Technology <https://orcid.org/0000-0002-0495-8510>

Yunlin Zhao

Central South University of Forestry and Technology

Meng Dong

Hunan City University

Shiquan Liu

Hunan City University

Zhiyuan Hu

Hunan City University

Xiaofen Zhong

Hunan City University

Zhenggang Xu (✉ rssq198677@163.com)

<https://orcid.org/0000-0002-7401-5163>

Research article

Keywords: chloroplast genome, cultivated tea, evolution, ycf1, Camellia

Posted Date: July 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-15441/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at BMC Ecology and Evolution on April 30th, 2021. See the published version at <https://doi.org/10.1186/s12862-021-01800-1>.

Abstract

Background: The cultivated tea is one of the most important economic and ecological trees distributing worldwide. The cultivated tea suffered from long-term targeted selection of traits and overexploitation of habitats by human beings, which may change the genetic structure. Chloroplast is an organelle with a conserved cyclic structure, and can help us better understand the evolutionary relationship of *Camellia* plants.

Results: The study conducted comparative analysis and evolution analysis between cultivated tea and wild tea, and detected the evolution characteristic in cultivated tea. Chloroplast genome sizes of cultivated tea were slightly different, ranged from 157,025 bp to 157,085 bp. These cultivated species were more conservative than wild species, in terms of the genome length, genes number, genes arrangement and GC contents. However, the IRs length of cultivated species was about 20 bp shorter than that of *C. sinensis* var. *sinensis*. We also found that the nucleotide diversity of 14 sequences in cultivated tea was higher than that of wild tea. These results provided the evidence on the variation of chloroplast genomes of cultivated tea. Detail analysis on the chloroplast genome variation and evolution of cultivated tea showed that 67 SNPs and 46 indels and 16 protein coding genes had nucleotide substitutions. The most common variation gene was *ycf1*. It has the largest number of nucleotide substitutions. At the same time, in *ycf1*, five amino acid sites were exhibiting site-specific selection, and a 9 bp sequence insertion was found in the *C. sinensis* cultivar. Anhua. The phylogenetic tree constructed by *ycf1* sequence shows that two cultivated tea were not completely clustered, and the evolutionary relationship between *C. sinensis* var. *sinensis* and *C. sinensis* cultivar. Longjing is closer than that of *C. sinensis* cultivar. Anhua.

Conclusions: The cultivated species were more conservative than wild species in terms of architecture and linear sequence order. The variation of chloroplast genome of cultivated tea was mainly manifested in the nucleotide polymorphism of some sequences. The *ycf1* gene played an important role in the adaptive evolution of cultivated tea. These results provided evidence regarding the influence of human activities on tea.

Background

From ancient times, numerous plant species have been taken from their habitats and introduced into cultivation—that is, into various human-made systems [1]. The cultivated process has played an important role in human history and cultivated environments often presented strong ecological contrasts with the wild environments [2]. Wild species are exposed to natural selection that operates to survival under abiotic and biotic stress, while cultivated species are subjected to artificial selection that emphasizes a steady supplying, improving quality and yields. Criteria for fitness are expected to change dramatically under both regimes. Therefore, alterations in vegetation phenology, growth and reproductive traits occur essentially because plants are subjected to different levels of stress and distinctive selection pressures [3]. Pot experiment showed there is significant differences of flowering and pod set in wild and cultivated types of soybean[4]. Besides, the compounds, microstructures are also surveyed for many horticultural plants[5]. However, current study does not pay enough attention to the genetic variation between cultivated and wild species. The inadequate genetic information prevents us to understand the spreading process of cultivated plants. We need to compare the genetic differences between cultivated species and wild species in order to use the species more effectively.

Camellia, containing about 280 species, is a genus with high economic, ecological and phylogenetic values in the family Theaceae [6, 7]. The *Camellia* is native to Asia and has been cultivated for more than 1300 years [8]. Because of its variety use, the cultivated species are now found all over the world [9, 10]. *Camellia* species can provide many valuable products, including making tea with young leaves and extracting edible oil with seeds. Moreover, most *Camellia* species are of great ornamental values [11]. The genus *Camellia* is composed of more than 110 taxa [12], of which only one species, *Camellia sinensis* var. *sinensis* (CSVS), is used as the source of the beverage tea [13]. Due to long-term cultivation and manual selection, the *Camellia sinensis* formed many local varieties plant, such as *Camellia sinensis* cultivar Anhua (CSCA) and *Camellia sinensis* cultivar Longjing (CSCL). The leaves of CSCA, well known for its specific area, are the main sources of dark tea [14].

The abundant cultivars germplasm resources are the fundamental condition for the dark tea products quality. The quality of dark tea could be more related to cultivar of CSVS and geographical tea are formed [15].

Chloroplast (cp) genome is often used to analyze evolution process and phylogenetics status because of its high conservation and relative compact gene alignment. Moreover, cp genome sequences are useful in the identification of closely related, breeding-compatible plant species [16]. Although the cp genome is very useful, there are still a very limited number of cp genomes available for the *Camellia* species so far. Until now, there are only sixteen *Camellia* complete cp genomes which has been sequenced, including 2 cultivated species and 14 wild species [7, 14, 17-21]. It has been proved that human interference has effects on the genetic structure, leaf nutrients and pollen morphology of *Camellia* [22-24]. Due to human overexploitation of habitats and long-term targeted selection of traits, the genetic diversity of *Camellia* germplasm resources have been significantly reduced [25]. Nevertheless, it remains unclear what the evolution mechanism of the cp genome of artificially selected cultivated *Camellia* is.

Current researches always ignore material differences between cultivated and wild species. Basing on our sequencing complete chloroplast genome of CSCA (MH042531) data, we want to explore evolution characteristic between cultivated tea and its wild relatives[14]. To assess the variations of the chloroplast genome in wild and cultivated species of *Camellia*, and detect the evolution characteristic in cultivated tea, we selected earlier published *Camellia* chloroplast genomes and did comparative and evolutionary analysis. It can help us to better understand the structure of *Camellia* chloroplast genomes and the phylogenetic relationships among species, and provide more information about the influence of human activities on tea. We believe that this research can make other researchers pay more attention to the source of the tea tree.

Results

Chloroplast genome features of cultivated tea

Length of the whole genome of cultivated tea were slightly different for CSCA and CSCL, ranged from 157,025 bp to 157,085 bp. Protein-coding regions (CDS), intergenic spacer (IGS), intron regions showed the same results that are 30 bp, 28 bp, 2 bp difference in length, respectively. Both the two kinds of cultivated tea consist of 81 unique CDS genes, 30 tRNA, 4 rRNA and 3 pseudogenes (*ycf1*, *ycf2* and *ycf15*). Among them, *atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps16*, *trnG-GCC*, *trnI-GAU*, *trnL-UAA*, *trnV-UAC* contained a single intron, *clpP* and *ycf3* contained two introns (Fig. 1).

Comparison of chloroplast genomes between cultivated tea and wild tea

Chloroplast genomic similarity

We used CSVS as a reference sequence and compared it with cpDNA genomes of two cultivated species and 13 wild relatives (Table 1). Compared with the genome length of CSVS, the average length variation of cultivated species and wild species are 62 bp and 186 bp respectively. Three of the wild species, named *C. pitardii*, *C. cuspidata*, *C. yunnanensis*, have a 500 bp length variation compared with CSVS. Similarly, the number of genes and GC content of cultivated species were more stable than that of wild species. Comparing the gene and intron deletion between cultivated and wild species, the results showed that the introns of *rps12* gene were deleted in all cultivated species and some wild species. The *orf42* and *ycf1* genes of some wild species were deleted, but these two genes are common in cultivated species. The differences of GC content of intron and IGS in cultivated species were about 0.01% to 0.03%, but we found that the differences of intron and IGS in wild species were 0% to 1.5% and 0% to 0.46%, respectively. mVISTA and BRIG were used to compare the genomic sequence identity. The results showed that the cp genomes of these 16 species were highly similar at the genomic level, indicating that their genomes are relatively conservative. The regions with relatively low identity were *atpH_atpI*, *trnE-UCC_trnT-GGU*, *psaA_ycf3*, *ycf15_trnL-CAA*, *ycf1_ndhF* and *ndhG_ndhI* (Figs. 2 & 3). In conclusion, at the genomic level, cultivated species were more conservative than wild species.

The expansion and contraction of IR regions

The locations of IRs regions were extracted via a self-BLASTN search, and the characteristics of IR/LSC and IR/SSC boundary regions were analyzed. The IRs boundary regions of the 16 complete *Camellia* cp genomes were compared, showing slight differences in junction positions (Fig. 4). In order to detect possible IR border polymorphism, first of all, we compared four IR-SC boundaries among cultivated tea and CSVS cp genome, but no difference was found at the LSC-IRb, IRa-LSC border. At the IRb-SSC, SSC-IRa border, only minor differences were discovered. Then, we compared the cp genome boundaries of wild tea and CSVS. The *rps19* gene at the LSC-IRb boundary expanded 52 bp from the LSC region to the IRb side in *C. sinensis* var. *pubilimba*, while it stopped at the 46 bp from the LSC region in the rest of the species. On the other side of the IRa-LSC boundary, the lengths of the spacers between the IRa-LSC junction and the *rp12* gene (in IRa) were 112 bp for *C. sinensis* var. *pubilimba* while those of the rest species were all 106 bp. Consistently, in all of the comparative cp genomes, the *ycf1* gene spanned the SSC-IRb region and the length of *ycf1* were from 963bp to 1069bp in IRa. Remarkably, most species have created the *ycf1* pseudogene at the IRa-LSC junction while it was not observed in *C. sinensis* var. *assamica*, *C. taliensis*, *C. impressinervis*, *C. pitardii*, *C. crapnelliana*, *C. cuspidate*, *C. yunnanensis*. Like most plants, the *ndhF* gene involved in photosynthesis was located in the SSC region. However, *ndhF* gene was located at the IRb-SSC boundary of *C. reticulata*, and there was a 35 bp overlap between *ndhF* and *ψycf1*.

Nucleotide diversity

Comparison based on the Pi values of the sixteen species' cp genomes were presented, including the intergeneric regions (IGS), protein-coding genes and introns (Table S1, Fig. 5). The results showed that the Pi values of the genes, introns and IGS in wild species were higher than cultivated species. The average values for the genes, introns and IGS in wild species were about 16.6, 3.5 and 9.1 times to that of cultivated species.

The genes and IGS in wild species with higher values, such as *psbB* and *psbD*, while Pi values were 0 in cultivated species. These results suggest that these genes and IGS were more conserved in cultivated species than in wild species.

The Pi values of genes and IGS, such as *psbB*, *psbD*, *trnI-CAU_ycf2*, *trnI-GAU_rrn16*, and *trnI-CAU_rpl23*, were higher of wild species, but they were 0 in cultivated species. Furthermore, except for *ndhD*, *ndhF*, *ndhH* and *psbC*, the Pi values of the photosynthetic genes of cultivated tea were 0. The Pi values of these genes were smaller than that of wild species. These results indicate that these genes and IGS were more conserved of cultivated species than of wild species. Although the Pi values of cultivated species were less, we still found that the Pi values of *rps16*, *rps4*, *trnL-UAA_intron*, *rps4_trnT-UGU*, *ndhC_trnV-UAC*, *cemA_petA*, *rpl33_rps18*, *psbN_psbH*, *rpl36_infA*, *rpl14_rpl16*, *rps7_rps12*, *ndhG_ndhI*, *trnV-GAC_rps12*, and *rps12_rps7* in cultivated species were higher than that in wild species, and these difference sequences were mainly located in LSC region (Fig. 5).

Phylogenetic analysis of cultivated tea and wild tea

We constructed three phylogenetic trees of cultivated and wild tea, namely, the complete cp genomic tree (complete cp-Tree), all shared coding protein genes among all species tree (SCDS-Tree) and the *ycf1* gene tree (*ycf1*-Tree) (Fig. 6-8). All phylogenetic trees strongly supported that Thea subgenus could be divided into two clades: clade I, including CSVS, CSCL, CSCA, CSVA, CGR, CPU and CSVP, and clade II, including CIM and CTA. The clade I was strongly supported, because the posterior probabilities or bootstrap values obtained by the neighbor-joining (NJ), the maximum parsimony (MP), the Bayesian inference (BI) and the maximum likelihood (ML) were very high for each lineage. These results suggested that the evolution direction of seven species in the clade I was the same, but not the same as two species in the clade II. All phylogenetic trees proved that CSVS was the closest relative species to CSCA and CSCL, and the three species were in the same branch. In particular, in *ycf1*-Tree, the posterior probabilities or bootstrap values of CSVS, CSCL and CSCA were lower than that of complete cp-Tree and SCDS-Tree. And the value of CSCL was less than 50%. These results suggested that the *ycf1* gene has evolved in cultivated tea.

However, we found conflict in three trees (Fig. 6-8). The topological structures consisted of Protocamellia subgenus (CPE and CYU), Camellia subgenus (CPI, CRE, CAZ and CCR), Metacamellia subgenus (CCU) and Thea subgenus (CIM and CTA) were

poorly supported by the complete cp-Tree, SCDS-Tree and *ycf1*-Tree. Because most bootstrap values or posterior probabilities were less than 50% for each lineage. These results may be caused by unbalanced sampling.

The complete cp-Tree showed some structural variations in *Camellia* cp genomes (Fig. 6). The clade, which made up of CSVS, CSCL, CSCA, CSVA, CGR, CPU, CSVP and CPE, was reported the loss of *rps12* intron, the pseudo *ycf1* gene, and the pseudo *ycf2* gene (except for CSVA). The other species, except for CRE and CAZ, was reported the loss of *ycf1* gene and *orf42* gene.

Chloroplast genome variation and evolution in cultivated tea

In order to explain the changes of cp genome structure of cultivated tea group, we took CSVS, CSCA and CSCL as objects to research SNPs and indels in the cp genome of cultivated tea. After comparing the whole cp genome of three species, 67 SNPs and 46 indels were found. The LSC, IRb, SSC and IRa regions contain 43, 3, 13, 8 SNPs and 37, 2, 5, 2 indels, respectively (Table S2). Most of SNPs and indels were located in the non-coding protein region (IGS and intron). There were 39 SNPs and 41 Indels in this region, while 28 SNPs and 5 Indels are found in the coding protein region. The two *ycf1* genes, which located in the junction of SSC and IRa, contained the most SNPs and indels, with the number of 6 and 2 respectively. For photosynthetic genes, *psbC*, *ndhD*, *ndhF* and *ndhH* presented SNPs variations, while the *psbI* gene presented indel variation. For the 14 sequences with higher Pi value in cultivated species than in wild species, *trnV-GAC_rps12* and *ndhG_ndhI* contained the most abundant SNPs, with 5 and 2 respectively (Fig. 5).

In order to have a clear view of the evolution of cultivated species, we used their 80 shared coding protein genes to calculated Ka, Ks and Ka/Ks ratio. The results showed that only 16 coding protein genes had synonymous or non-synonymous mutations (Fig. 9, Table S3). Among them, there are non-synonymous mutations in *matK*, *rps16*, *rpoC2*, *rpoB*, *accD*, *clpP*, *rps8*, *ycf1*, *ndhD*, *ndhH* and *rps15*. The genes with the highest rate of non-synonymous mutations are *rps16*, *rps8* and *rps15*. There are synonymous mutations in *rpoB*, *psbC*, *rps4*, *ycf4*, *rpoA* and *ndhF*. The highest mutation rates are *rps4*, *ycf4* and *rpoA*. Of the 80 genes, 79 have a Ka / Ks value of 0, and only *rpoB*, has a Ka/Ks value of 0.3004 < 0.5, suggesting very strong purifying selective pressure.

The site specific selection events of 16 genes with synonymous or non-synonymous mutations were analyzed by Bayesian Empirical Bayes (BEB), and found some amino acid sites of *ycf1* and *rps15* exhibited site-specific selection (Table S4). In *ycf1*, there were six sites under position selection, and in *rps15*, there was one site under position selection. For example, in *rps15* gene, the codon ACC (threonine) of CSVS mutated to AAC (asparagine) of two cultivated species.

Discussion

Understanding the genetic variation between cultivated and wild species is crucial for introducing the interest traits from wild species into cultivars [26]. Organelle genome sequencing is proven to be an effective way to resolve phylogenetic relationships among closely related species [27, 28]. Here, we constructed and compared the complete cpDNA genome sequences of two cultivars and fourteen wild species of *Camellia*. At the genomic level, cultivated species were more conservative than wild species, in terms of architecture and linear sequence order (the length, genes number, genes arrangement, and GC contents) (Table 2, Figs. 2 & 3). As for other land plant species, such as peanuts, cherries and radishes, the cp genome size and structure, as well as gene content and order are highly conserved among the cultivated and wild species acquired cp genomes [29-31].

We found that the IR regions of two kinds of cultivated tea expanded or contracted. The IRs length of the two cultivated species was about 20 bp shorter than that of CSVS, accounting for 32% of the difference in the complete genome length (Fig. 4). In fact, the contraction and expansion of IRs is considered to be one of the important reasons for cp genome length variation [32]. Further analysis showed that in CSCA, the sum of the length of two *ycf1* genes located in the boundary region

of IRb/SSC and SSC/IRa, was 18 bp less than that of CSVS and CSCL (Fig. 4). This is mainly because a 9 bp sequence (TCCTTCTTC/GAAGAAGAAGGA) was inserted into the *ycf1* gene, based on the results of SNPs and Indels analysis (Table S2). It is suggested that *ycf1* is one of the important reasons for the expansion or contraction of the IRs of cultivated species. The same results were also found in Zheng's study [33]. He analyzed the cp genome length variation in 272 species and found that *atpA*, *accD* and *ycf1* accounted for 13% of the difference in length. Therefore, *ycf1*, which is associated with plant survival, may play a key role in the cp genome size variations of cultivated tea.

In addition to the variations in genome size, there are also some nucleotide mutations in the cultivated species. In this study, the nucleotide diversity of cultivated tea was lower than that of wild tea (Fig. 5), the unbalanced sampling between the wild tea (14) and cultivated tea (2) may lead to the nucleotide diversity difference of cpDNA fragments. The nucleotide diversity comparison of 358 cultivated rice and 54 wild rice also presented similar results [34]. Nevertheless, we also found that the nucleotide diversity of 14 sequences in cultivated tea was higher than that of wild tea (*rps16*, *rps4*, *trnL-UAA_intron*, *rps4_trnT-UGU*, *ndhC_trnV-UAC*, *cema_petA*, *rpl33_rps18*, *psbN_psbH*, *rpl36_infA*, *rpl14_rpl16*, *rps7_rps12*, *ndhG_ndhI*, *trnV-GAC_rps12*, and *rps12_rps7*) (Fig. 5). These sequences proved the variations in cp genomes of cultivated tea, and they were potential molecular markers for distinguishing *Camellia* species and for the phylogenetic analysis of *Camellia*.

Previous studies have proved that human interference has effects on the genetic structure, leaf nutrients and pollen morphology of *Camellia*. Yan et al. analyzed the genetic relationship of five semi wild tea which were lack of human management for a long time was studied by using genome-wide SNP. He find that human interference will affect the genetic structure of tea. After the human interference stopped, the tea from five different geographical regions could be divided into three different groups because of the absence of free pollination [22]. Xiong et al. make comparative analyses of the nutrient content in leaves in cultivated and wild *C. nitidissima*. He find that cultivated *C. nitidissima* had significantly higher contents of essential amino acids (26.05%) and total amino acids (33.27%) than that of wild *C. nitidissima* [23]. Shu et al. proved that there are obvious differences in pollen morphology and exine morphology between cultivated and wild species of *Camellia* [24]. Therefore, to explore specific evolution characteristics between cultivated tea and its wild relatives, we subsequently did evolution research on cultivated tea.

First, to have a clear view of the cp genomic adaptative evolution of cultivated tea, we did evolution analysis on protein-coding sequence. The Ka/Ks ratio is very useful for measuring selective pressure at the protein level[35]. In this study, Ka/Ks value of 79 genes are 0, and the only *rpoB* is 0.3004. Besides, some amino acid sites of *ycf1* and *rps15* exhibiting site-specific selection (Table S3 & S4). *rpoB* were crucial for genetic information transmission, which affect transcription of DNA into RNA and translation of RNA to protein. They were also under selective pressure in beverage crops[36]. *rps15* gene have function in chloroplast ribosome subunits[35]. *ycf1*, encoding a component of the chloroplast's inner envelope membrane protein translocon, is one of the largest plastid genes[36], and is also essential for almost all plant lineages[37]. These positively selected genes may have played key roles in the adaptation of cultivated tea to various environments.

Generally, the deletion or insertion of amino acids and bases in the coding protein gene will affect the structure and function of this gene [38-40]. In our study, 16 protein coding genes have nucleotide substitutions, among which *ycf1* gene has the largest number of nucleotide substitution. At the same time, in *ycf1*, five amino acid sites were exhibiting site-specific selection, and a 9 bp sequence insertion was found in the CSCA (Table S2-S4, Fig. 9). Therefore, we hypothesized that the *ycf1* gene played an important role in the adaptive evolution of cultivated tea.

ycf1 is an open reading framework of unknown function, but some studies infer that *ycf1* is very important for plant survival [33, 41]. In tobacco, a chimeric add a gene conferring resistance to aminoglycoside antibiotics is transferred into *ycf1* in its cp genome. Then, it is cultured in the plant regeneration medium containing the presence of the antibiotic spectinomycin. After that, the maintenance of a fairly constant ratio of wild-type versus transformed genome copies is found. However, the wild-type genome is still present in all samples whereas the transplastomic fragments are missing from several samples after culturing in antibiotic-free medium. This experiment proved that *ycf1* encode products that are essential for cell survival. *ycf1* is also an important molecular marker of plants [42, 43]. Because it has higher variability than other known cp molecular

markers (such as widely used *rbcL* and *matk* genes), in total number of parsimony informative characters and in percent variability.

Phylogenetic analysis of cultivated and wild tea showed that two cultivar tea were closely related to the CSVS (Figs. 6 & 7), which supports the previous finding that most of the tea originated directly from CSVS [13]. But in *ycf1*-Tree, the posterior probabilities or bootstrap values of CSVS, CSCL and CSCA were lower than that of complete cp-Tree and SCDS-Tree, which suggested that the *ycf1* gene has evolved in cultivated tea (Figs. 6-8). Similar results have been found in *Corylus* [44]. The *ycf1* gene of *Corylus chinensis* and *Corylus avellana* has similar evolutionary history, which is different from that of *Corylus heterophylla*. This evolution of cultivated plants may be related to the utilization efficiency of photosynthesis. Because of photosystem biogenesis regulator 1 (PBR1), the RNA binding protein encoded by the nuclear genome, can improve the translation efficiency of *ycf1* in *Arabidopsis thaliana* cp genome. Then, the symbiosis and stability maintenance of the three photosynthetic complexes were regulated [45]. However, at present, the effect of the mutation in the single amino acid site and insertion or deletion of the short sequence on the function of *ycf1* is still not clear, and cultivated tea may be ideal materials for this kind of research.

Conclusion

In this work, the complete cp genomes of two cultivated and 14 wild species of *Camellia* were selected. Genomic variation and evolution press were compared in these species. Our results shown that the cultivated species were more conservative than wild species in terms of architecture and linear sequence order. The variation of chloroplast genome of cultivated tea was mainly manifested in the nucleotide polymorphism of some sequences. This nucleotide polymorphism also led to the mutation of amino acid sites in some genes, among which *ycf1* was the gene with the most mutation sites. In addition to amino acid mutation, there was a 9 bp base insertion in the *ycf1* gene. *ycf1* is believed to be a critical gene for plant survival, and might influence photosynthesis and is related to plant adaptation. In cultivated tea, two cultivated tea are not completely clustered, and the evolutionary relationship between CSVS and CSCL was closer than that of CSCA. However, at present, the effect of the mutation in the single amino acid site and insertion or deletion of the short sequence on the function of *ycf1* was still not clear, and cultivated tea may be ideal materials for this kind of research.

Material And Methods

Genomic materials collection of cultivated tea

The complete cp genome of CSCA has been presented and annotated in our previous study [14] with GenBank accession number MH042531. Meanwhile, we searched in National Center for Biotechnology Information (NCBI) dataset to find published cultivated tea's complete cp genome, only CSCL with accession number KF562708 have been published [46]. Gene map of cultivated tea was generated using OGDRAW v1.2 [47].

Comparative analysis between cultivated tea and wild tea

The Basic Local Alignment Search Tool (BLAST) was used to find similarity cp genomes of CSCA in NCBI. After screening the cp genome of *Camellia*, there were 16 *Camellia* cp genomes with sampling information, including 2 cultivated species (only CSCA and CSCL) and 14 wild species (Table 2). Previous studies have shown that both CSCA and CSCL originated directly from CSVS[13, 46]. Therefore, we used CSVS as the reference sequence to study genomic variations and evolution direction between cultivated tea and wild tea.

The comparison of cp genomes was based on the method of Li [48] using mVista in Shuffle-LAGAN mode and Blast Ring Image Generator (BRIG), respectively. CSVS was set as the reference sequence.

We annotated the IRs boundary of *Camellia* cp genomes by Plastid Genome Annotator (PGA) [49]. The location of their IRs was extracted, and then the expansion and contraction of IR regions were visualized by using the Visio professional 2016.

The comparison of nucleotide diversity (π) was finished according to the method of Njuguna [50]. First, we manually extracted 211 loci shared among 16 *Camellia* species, including 80 coding protein genes, 117 intergenic regions (IGS), and 14 intron regions. After multiple alignments, a sliding window analysis was conducted to compare the nuclear diversity among the cp genomes using DnaSP v6.10.04 [51]. The window length was 600 bp with a 200 bp step size.

Phylogenetic analysis of *Camellia*

Three datasets were used to construct the following phylogenetic trees for 14 wild species and two cultivated species in the *Camellia*: (I) the complete cp genomes, (II) the all shared coding protein genes among all species (SCDS), (III) the *ycf1* gene sequences. All datasets were aligned using MAFFT v7.380 [52] under the FFT-NS-2 default setting. The alignments were used for phylogenetic analysis. According to the method described by Xie et al.[53] and Zhang et al.[54], we used four methods to construct phylogenetic trees, that were the NJ method, the MP method, the BI method and the ML method. *Coffea canephora* and *Coffea arabica* were selected as the outgroup.

The NJ analysis was reconstructed via MEGA7.0 [55] under the default settings with 1000 bootstrap values. The MP analysis was performed in PAUP 4.0a167 [56] with heuristic searches with 1000 bootstrap replicates. The BI analysis was performed with MrBayes 3.2.7 [57] under the best substitution models and parameters. The analysis parameters were set as four chains that were run simultaneously for 10,000,000 generations or until the average standard deviation of the split frequencies fell below 0.01. The best substitution models and parameters were computed by jmodeltest 2.1.7 [58]. The ML analysis was carried out in IQ-TREE [59] using the default settings, with 1000 bootstrap values for tree evaluation. The best substitution models were computed by IQ-TREE.

Evolution research on cultivated tea

The number and position of single nucleotide polymorphisms (SNPs) and insertion/deletions (Indels) in the genomes of CSVS, CSCA and CSCL were extracted in DnaSP v6.10.04 according to the Wu's method [60].

We detected the synonymous (K_s) and nonsynonymous (K_a) nucleotide substitution rates as well as the K_a/K_s ratio in homologous protein-coding regions of CSVS, CSCA and CSCL. After aligning each gene using the ClustalW (Codons) program in MEGA7, the K_s , K_a and K_a/K_s values of each gene were determined according to Dong's method [61] with the program from the PAML package [62]. For identification of site-specific selection, four models, M1 (neutral), M2 (selection), M7 (beta) and M8 (beta & ω), were used in codeml from the PAML package. The Bayes empirical Bayes (BEB) was used to calculate the posterior probabilities for site classes. Only sites with posterior probabilities > 0.9 were selected.

Declarations

Acknowledgements

We would like to thank Dr. Huang Hui from Kunming Institute of Botany for providing us with the information of species origin.

Authors' contributions

Peng J and Xu ZG conceived the study. All authors collected field samples. Dong M, Liu SQ, Hu ZY, Zhong XF analyzed the final data. Zhao YL and Xu ZG acquired funds for this study. Peng J wrote the original manuscript, and all authors commented on an early draft of the manuscript.

Funding

This study was funded by the Major Science and Technology Program of Hunan Province (2017NK1014), Key Technology R&D Program of Hunan Province (2016TP2007, 2016TP1022), Natural Science Foundation of Hunan Province (2019JJ50027) and Innovation Experiment Program for Undergraduate of Hunan City University.

Availability of data and materials

Raw sequences data of CSCA were submitted to National Center for Biotechnology Information (NCBI) database with accession number MH042531. The other genomic data mentioned in articles also can be accessed from NCBI.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Zohary D: **Domestication of Crop Plants**. 2001:217-227.
2. Denison RF, Kiers ET, West SA: **Darwinian agriculture: when can humans find solutions beyond the reach of natural selection?** *The Quarterly review of biology* 2003, **78**(2):145-168.
3. Matesanz S, Gianoli E, Valladares F: **Global change and the evolution of phenotypic plasticity in plants**. In: *Year in Evolutionary Biology*. Edited by Schlichting CD, Mousseau TA, vol. 1206; 2010: 35-55.
4. Kuniyuki Saitoh KN, Toshiro Kuroda: **Characteristics of Flowering and Pod Set in Wild and Cultivated Types of Soybean**. *PLANT PRODUCTION SCIENCE* 2004, **7**(2):172-177.
5. Aharoni A, Giri AP, Verstappen FWA, Bertea CM, Sevenier R, Sun Z, Jongsma MA, Schwab W, Bouwmeester HJ: **Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species**. *The Plant cell* 2004, **16**(11):3110-3131.
6. Vijayan K, Zhang W-J, Tsou C-H: **MOLECULAR TAXONOMY OF CAMELLIA (THEACEAE) INFERRED FROM NRITS SEQUENCES**. *American Journal of Botany* 2009, **96**(7):1348-1360.
7. Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z: **Comparative Chloroplast Genomes of Camellia Species**. *Plos One* 2013, **8**(8).
8. Gao Jiying CRP: **Collected Species of The Genus Camellia-An Illustrated Outline**. Zhejiang: Zhejiang Science and Technology Press; 2005.
9. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE: **Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(10):4623-4628.
10. Wachira F, Tanaka J, Takeda Y: **Genetic variation and differentiation in tea (Camellia sinensis) germplasm revealed by RAPD and AFLP variation**. *Journal of Horticultural Science & Biotechnology* 2001, **76**(5):557-563.
11. Jianfei Z: **Flora of China**, vol. 12. Beijing: Flora of China Editorial Committee.; 2007.
12. Ming TL: **A Systematic Synopsis of the Genus Camellia**. *Acta Botanica Yunnanica* 1999.
13. Liu Y, Yang S-x, Ji P-z, Gao L-z: **Phylogeography of Camellia taliensis (Theaceae) inferred from chloroplast and nuclear DNA: insights into evolutionary history and conservation**. *Bmc Evolutionary Biology* 2012, **12**.

14. Dong M, Liu S, Xu Z, Hu Z, Ku W, Wu L: **The complete chloroplast genome of an economic plant, *Camellia sinensis* cultivar Anhua, China.** *Mitochondrial DNA Part B-Resources* 2018, **3**(2):558-559.
15. Chen S, Deng Y, Gong Z, Zhu H, University HA: **Analysis of the Evolution of Technological Innovation Model of Anhua Dark Tea Industry.** *Journal of Agriculture* 2015.
16. Daniell H, Lin C-S, Yu M, Chang W-J: **Chloroplast genomes: diversity, evolution, and applications in genetic engineering.** *Genome Biology* 2016, **17**.
17. Ye X, Zhao Z, Zhu Q, Wang Y, Lin Z, Ye C, Fan L, Xu H: **Entire chloroplast genome sequence of tea(*Camellia sinensis* cv.Longjing 43):a molecular phylogenetic analysis.** *Journal of Zhejiang University* 2014.
18. Huang H, Shi C, Liu Y, Mao S-Y, Gao L-Z: **Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships.** *Bmc Evolutionary Biology* 2014, **14**.
19. Xu X, Zheng W, Wen J: **The complete chloroplast genome of the long blooming and critically endangered *Camellia azalea*.** *Conservation Genetics Resources* 2018, **10**(1):5-7.
20. Yang J-B, Li D-Z, Li H-T: **Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs.** *Molecular Ecology Resources* 2014, **14**(5):1024-1031.
21. Zeng C-X, Hollingsworth PM, Yang J, He Z-S, Zhang Z-R, Li D-Z, Yang J-B: **Genome skimming herbarium specimens for DNA barcoding and phylogenomics.** *Plant Methods* 2018, **14**.
22. Yan G, Da-he Q, Chun Y, Yan L, Zheng-wu C, Juan C: **Genetic Diversity of Old Tea Plant Resources in Jiuan City of Guizhou Province,Using Genome-Wide SNP.** *Journal of Plant Genetic Resources* 2019.
23. Xiong Z, Qi X, Wei X, Chen Z, Tang H, Chai S: **NUTRIENT COMPOSITION IN LEAVES OF CULTIVATED AND WILD *CAMELLIA NITIDISSIMA*.** *Pakistan Journal of Botany* 2012, **44**(2):635-638.
24. Shu J, Liang C, Wang H, Wang P, Mei X, Song W: **Pollen Morphology,Ultrastructure and Evolution of Tea Plant and other Genus *Camellia* Plants.** *Journal of Tea Science* 1998.
25. Zhao D-w, Yang J-b, Yang S-x, Kato K, Luo J-p: **Genetic diversity and domestication origin of tea plant *Camellia taliensis* (Theaceae) as revealed by microsatellite markers.** *Bmc Plant Biology* 2014, **14**.
26. Amar MH, Magdy M, Wang L, Zhou H, Zheng B, Jiang X, Atta AH, Han Y: **Peach chloroplast genome variation architecture and phylogenomic signatures of cpDNA introgression in *Prunus*.** *Canadian Journal of Plant Science* 2019, **99**(6):885-896.
27. Ivanova Z, Sablok G, Daskalova E, Zahmanova G, Apostolova E, Yahubyan G, Baev V: **Chloroplast Genome Analysis of Resurrection Tertiary Relict *Haberlea rhodopensis* Highlights Genes Important for Desiccation Stress Response.** *Frontiers in Plant Science* 2017, **8**.
28. Ma P-F, Zhang Y-X, Zeng C-X, Guo Z-H, Li D-Z: **Chloroplast Phylogenomic Analyses Resolve Deep-Level Relationships of an Intractable Bamboo Tribe *Arundinarieae* (Poaceae).** *Systematic Biology* 2014, **63**(6):933-950.
29. Cho M-S, Yoon HS, Kim S-C: **Complete chloroplast genome of cultivated flowering cherry, *Prunus xyedoensis* 'Somei-yoshino' in comparison with wild *Prunus yedoensis* Matsum. (Rosaceae).** *Molecular Breeding* 2018, **38**(9).
30. Wang J, Li Y, Li C, Yan C, Zhao X, Yuan C, Sun Q, Shi C, Shan S: **Twelve complete chloroplast genomes of wild peanuts: great genetic resources and a better understanding of *Arachis* phylogeny.** *Bmc Plant Biology* 2019, **19**(1).
31. Yamagishi H, Terachi T, Ozaki A, Ishibashi A: **Inter- and intraspecific sequence variations of the chloroplast genome in wild and cultivated *Raphanus*.** *Plant Breeding* 2009, **128**(2):172-177.
32. Palmer JD, Nugent JM, Herbon LA: **Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families.** *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(3):769-773.
33. Zheng X, Wang J, Feng L, Liu S, Pang H, Qi L, Li J, Sun Y, Qiao W, Zhang L *et al*: **Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants.** *Scientific Reports* 2017, **7**(1).

34. Cheng L, Nam J, Chu S-H, Rungrapa P, Min M-h, Cao Y, Yoo J-m, Kang J-S, Kim K-W, Park Y-J: **Signatures of differential selection in chloroplast genome between japonica and indica.** *Rice* 2019, **12**(1).
35. Gao C, Deng Y, Wang J: **The Complete Chloroplast Genomes of Echinacanthus Species (Acanthaceae): Phylogenetic Relationships, Adaptive Evolution, and Screening of Molecular Markers.** *Frontiers in Plant Science* 2019, **9**.
36. Liu M-L, Fan W-B, Wang N, Dong P-B, Zhang T-T, Yue M, Li Z-H: **Evolutionary Analysis of Plastid Genomes of Seven Lonicera L. Species: Implications for Sequence Divergence and Phylogenetic Relationships.** *International Journal of Molecular Sciences* 2018, **19**(12).
37. Dong W-L, Wang R-N, Zhang N-Y, Fan W-B, Fang M-F, Li Z-H: **Molecular Evolution of Chloroplast Genomes of Orchid Species: Insights into Phylogenetic Relationship and Adaptive Evolution.** *International Journal of Molecular Sciences* 2018, **19**(3).
38. Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK: **Extensive reorganization of the plastid genome of Trifolium subterraneum (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions.** *Journal of molecular evolution* 2008, **67**(6):696-704.
39. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT *et al*: **Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP.** *Scientific Reports* 2015, **5**.
40. Wu Y, Liu F, Yang D-G, Li W, Zhou X-J, Pei X-Y, Liu Y-G, He K-L, Zhang W-S, Ren Z-Y *et al*: **Comparative Chloroplast Genomics of Gossypium Species: Insights Into Repeat Sequence Variations and Phylogeny.** *Frontiers in Plant Science* 2018, **9**.
41. Drescher A, Ruf S, Calsa T, Jr., Carrer H, Bock R: **The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes.** *The Plant journal : for cell and molecular biology* 2000, **22**(2):97-104.
42. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S: **ycf1, the most promising plastid DNA barcode of land plants.** *Scientific Reports* 2015, **5**.
43. Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M: **Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK.** *Plant Systematics and Evolution* 2009, **277**(1-2):75-84.
44. Yongle W, Zhifang W, Fang L, Jiwei Z, HUANGWugang, Yanping L, Lili C, Qingchang C, Guanglong H: **Bioinformatics Analysis of ycf1 Gene in Corylus.** *Journal of Shanxi Agricultural Sciences* 2018, **46**(08):1244-1247, 1333.
45. Yang X-F, Wang Y-T, Chen S-T, Li J-K, Shen H-T, Guo F-Q: **PBR1 selectively controls biogenesis of photosynthetic complexes by modulating translation of the large chloroplast gene Ycf1 in Arabidopsis.** *Cell Discovery* 2016, **2**.
46. Xiaoqian Y, Zhonghui Z, Quanwu Z, Yingying W, Zhangxiang L, Chuyu Y, Longjiang F, Hairong X: **Entire chloroplast genome sequence of tea (Camellia sinensis cv.Longjing 43): a molecular phylogenetic analysis** *Journal of Zhejiang University (Agric & Life Sci)* 2014, **40**(04):404-412.
47. Lohse M, Drechsel O, Kahlau S, Bock R: **OrganellarGenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets.** *Nucleic Acids Research* 2013, **41**(W1):W575-W581.
48. Li C, Zhao Y, Xu Z, Yang G, Jiao P, Xiaoyun P: **Initial characterization of the chloroplast genome of Vicia sepium, an important wild resource plant, and related inferences about its evolution.** *Frontiers in Plant Science* 2020.
49. Qu XJ, Moore MJ, Li DZ, Yi TS: **PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes.** *Plant Methods* 2019, **15**:12.
50. Njuguna AW, Li Z-Z, Saina JK, Munywoki JM, Gichira AW, Gituru RW, Wang Q-F, Chen J-M: **Comparative analyses of the complete chloroplast genomes of nymphoides and menyanthes species (menyanthaceae).** *Aquatic Botany* 2019, **156**:73-81.
51. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A: **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets.** *Molecular Biology and Evolution* 2017, **34**(12):3299-3302.

52. Katoh K, Standley DM: **MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.** *Molecular Biology and Evolution* 2013, **30**(4):772-780.
53. Xie D-F, Yu Y, Deng Y-Q, Li J, Liu H-Y, Zhou S-D, He X-J: **Comparative Analysis of the Chloroplast Genomes of the Chinese Endemic Genus Urophysa and Their Contribution to Chloroplast Phylogeny and Adaptive Evolution.** *International journal of molecular sciences* 2018, **19**(7):1847.
54. Zhang Y-b, Yuan Y, Pang Y-x, Yu F-I, Yuan C, Wang D, Hu X: **Phylogenetic Reconstruction and Divergence Time Estimation of Blumea DC. (Asteraceae: Inuleae) in China Based on nrDNA ITS and cpDNA trnL-F Sequences.** *Plants* 2019, **8**(7):210.
55. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Molecular Biology and Evolution* 2016, **33**(7):1870-1874.
56. Swofford D: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10**, vol. Version 4.0; 2002.
57. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754-755.
58. Darriba D, Taboada GL, Doallo R, Posada D: **jModelTest 2: more models, new heuristics and parallel computing.** *Nat Methods* 2012, **9**(8):772-772.
59. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.** *Molecular Biology and Evolution* 2014, **32**(1):268-274.
60. Wu Z, Gu C, Tembrock LR, Zhang D, Ge S: **Characterization of the whole chloroplast genome of Chikusichloa mutica and its comparison with other rice tribe (Oryzeae) species.** *Plos One* 2017, **12**(5):e0177553.
61. Dong M, Zhou X, Ku W, Xui Z: **Detecting useful genetic markers and reconstructing the phylogeny of an important medicinal resource plant, Artemisia selengensis, based on chloroplast genomics.** *Plos One* 2019, **14**(2).
62. Yang Z, Nielsen R: **Synonymous and nonsynonymous rate variation in nuclear genes of mammals.** *Journal of molecular evolution* 1998, **46**(4):409-418.
63. Huang H, Shi C, Liu Y, Mao SY, Gao LZ: **Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships.** *Bmc Evolutionary Biology* 2014, **14**.
64. Zeng C-X, Hollingsworth PM, Yang J, He Z-S, Zhang Z-R, Li D-Z, Yang J-B: **Genome skimming herbarium specimens for DNA barcoding and phylogenomics.** *Plant methods* 2018, **14**:43-43.
65. Yang JB, Yang SX, Li HT, Yang J, Li DZ: **Comparative Chloroplast Genomes of Camellia Species.** *Plos One* 2013, **8**(8).

Tables

Table 1 Chloroplast genomic features of sixteen *Camellia* species, with the *C.sinensis* var. *sinensis* as the outgroup

Abbreviations: *Camellia sinensis* var. *sinensis*, CSVS; *Camellia sinensis* cultivar Anhua, CSCA; *Camellia sinensis* cultivar Longjing, CSCL; *Camellia sinensis* var. *assamica*, CSA; *Camellia sinensis* var. *pubilimba*, CSP; *Camellia grandibracteata*, CGR; *Camellia taliensis*, CTA; *Camellia impressinervis*, CIM; *Camellia pubicosta*, CPU; *Camellia azalea*, CAZ; *Camellia pitardii*, CPI; *Camellia reticulata*, CRE; *Camellia crapnelliana*, CCR; *Camellia cuspidate*, CCU; *Camellia petelotii*, CPE; *Camellia yunnanensis*, CYU.

Table 2 Summary of research species information

Species	CSVS	CSCA	CSCL	CSA	CSP	CGR	CTA	CIM	CPU	CAZ	CPI	CRE	CCR	CCU	CPE	CYU
Length																
(bp)	157117	157025	157085	157028	157086	157127	156974	156892	157076	157039	156585	156971	156997	156618	157121	156592
Genes	115	115	115	113	115	115	115	115	115	115	115	115	114	115	115	115
CDS																
genes*	81	81	81	79	81	81	81	81	81	81	81	81	81	81	81	81
tRNA																
genes	30	30	30	30	30	30	30	30	30	30	30	30	29	30	30	30
Introns	18	18	18	22	18	18	21	21	18	18	21	18	20	21	18	21
CDS																
region	80542	80620	80650	79093	80622	80656	79577	79655	80665	80629	79619	76224	79649	79643	80650	79655
Introns																
region	15192	15196	15198	17902	15210	15205	16947	16897	15198	15195	16937	15182	16239	16917	15196	16935
IGS																
region	49535	49361	49389	48200	49405	49418	48591	48481	49365	49367	48171	53717	49321	48199	49427	48143
tRNA																
region	2802	2802	2802	2789	2802	2802	2813	2813	2802	2802	2812	2802	2742	2813	2802	2813
rRNA																
region	9046	9046	9046	9044	9047	9046	9046	9046	9046	9046	9046	9046	9046	9046	9046	9046
Genome																
GC	37.3	37.3	37.29	37.3	37.32	37.29	37.32	37.33	37.3	37.3	37.34	37.31	37.3	37.31	37.29	37.33
CDS																
GC	37.58	37.57	37.56	37.47	37.58	37.56	37.57	37.54	37.57	37.56	37.56	37.54	37.54	37.56	37.56	37.54
Introns																
GC	36.41	36.38	36.38	37.91	36.42	36.39	37.25	37.28	36.42	36.41	37.22	36.4	37.54	37.25	36.41	37.25
IGS GC	32.93	32.94	32.94	32.48	32.97	32.94	32.68	32.72	32.93	32.95	32.71	33.39	32.64	32.63	32.92	32.68
tRNA																
GC	52.86	52.86	52.86	52.99	52.86	52.86	52.86	52.9	52.89	52.86	52.92	52.86	52.88	52.9	52.86	52.9
rRNA																
GC	55.39	55.41	55.41	55.4	55.41	55.41	55.38	55.41	55.42	55.39	55.36	55.34	55.41	55.38	55.39	55.41
Gene				<i>orf42</i> ,									<i>orf42</i> ,			
losses				<i>ycf1</i> ,			<i>orf42</i> ,	<i>orf42</i> ,			<i>orf42</i> ,		<i>ycf1</i> ,	<i>orf42</i> ,		<i>orf42</i> ,
				<i>ycf15</i>			<i>ycf1</i>	<i>ycf1</i>			<i>ycf1</i>		<i>trnG</i>	<i>ycf1</i>		<i>ycf1</i>
Intron																
losses	<i>rps12</i>	<i>rps12</i>	<i>rps12</i>		<i>rps12</i>	<i>rps12</i>			<i>rps12</i>	<i>rps12</i>		<i>rps12</i>			<i>rps12</i>	

Species	GenBank ID	Subgenus	Types	Sample location	Location	References
<i>Camellia sinensis</i> var. <i>sinensis</i>	KJ806281	Thea	wild	Yunnan Academy of Agricultural Science	Yunnan, China	[63]
<i>Camellia sinensis</i> cultivar <i>Anhua</i>	MH042531	Thea	cultivar	Hunan City University	Hunan, China	[14]
<i>Camellia sinensis</i> cultivar <i>Longjing</i>	KF562708	Thea	cultivar	Huajiachi campus of Zhejiang University	Zhejiang, China	[17]
<i>Camellia sinensis</i> var. <i>assamica</i>	MH394410	Thea	wild	Kunming Institute of Botany, Kunming	Yunnan, China	[64]
<i>Camellia sinensis</i> var. <i>pubilimba</i>	KJ806280	Thea	wild	Yunnan Academy of Agricultural Science	Yunnan, China	[63]
<i>Camellia grandibracteata</i>	NC024659	Thea	wild	Yunnan Academy of Agricultural Science	Yunnan, China	[63]
<i>Camellia taliensis</i>	NC022264	Thea	wild	Kunming Institute of Botany	Yunnan, China	[65]
<i>Camellia impressinervis</i>	NC022461	Thea	wild	Kunming Institute of Botany	Yunnan, China	[65]
<i>Camellia pubicosta</i>	NC024662	Thea	wild	International Camellia Species Garden	Zhejiang, China	[63]
<i>Camellia azalea</i>	NC035574	Camellia	wild	Yangchun County	Guangdong, China	[19]
<i>Camellia pitardii</i>	NC022462	Camellia	wild	Kunming Institute of Botany	Yunnan, China	[65]
<i>Camellia reticulata</i>	NC024663	Camellia	wild	Kunming Institute of Botany	Yunnan, China	[63]
<i>Camellia crapnelliana</i>	NC024541	Camellia	wild	Kunming Botanical Garden	Yunnan, China	[20]
<i>Camellia cuspidata</i>	NC022459	Metacamellia	wild	Kunming Institute of Botany	Yunnan, China	[65]
<i>Camellia petelotii</i>	NC024661	Protocamellia	wild	International Camellia Species Garden	Zhejiang, China	[63]
<i>Camellia yunnanensis</i>	NC022463	Protocamellia	wild	Kunming Institute of Botany	Yunnan, China	[65]

Figure 1

Figure 1

Gene map of the complete chloroplast genome of cultivated tea. Genes lying inside of the circle are transcribed clockwise, and those outside are transcribed counterclockwise. Different color of blocks represents different functional groups. The darker gray color of the inner circle corresponds to the GC contents, and the lighter gray color corresponds to the AT contents.

Figure 2

The sequence identity of sixteen *Camellia* species. The inner circle is the reference genome. Next circles represent the sequence identity between *C. sinensis* var. *sinensis* and fifteen other species. The outermost circle corresponds to the protein-coding genes and intergenic spacer regions. Genes with clockwise arrows represent reverse strands, while genes with the counterclockwise arrow represent forward strands.

Figure 3

Alignment visualization of the sixteen *Camellia* chloroplast genome sequences using *C.sinensis* var. *sinensis* as reference. Vertical scale indicates the percentage of identity, ranging from 50% to 100%. Arrows indicate the annotated genes and their transcriptional direction. The different colored boxes correspond to exons, tRNA or rRNA, and noncoding sequences (CNSs).

Figure 4

Comparison of IRs boundary regions among the 16 *Camellia* chloroplast genomes, using *C. sinensis* var. *sinensis* as the reference. Boxes above or below the line are forward strands and reverse strands, respectively.

Figure 5

Comparative analysis of nucleotide variability (Pi) values between the cultivated tea and wild tea cp genome sequences. X-axis: position of the midpoint of a window, Y-axis: nucleotide diversity of each window.

Figure 6

The phylogenetic tree of *Camellia* species based on the complete cp genomes (complete cp-Tree). *Coffea canephora* and *Coffea arabica* were selected as the outgroup. Tree constructed by the neighbor-joining (NJ), the maximum parsimony (MP), the Bayesian inference (BI) and the maximum likelihood (ML) with bootstrap values or posterior probabilities above the branches, respectively. The bootstrap values less than 50% were represented by "-".

Figure 8

The phylogenetic tree of *Camellia* species based on the *ycf1* gene (*ycf1*-Tree). *Coffea canephora* and *Coffea arabica* were selected as the outgroup. Tree constructed by the neighbor-joining (NJ), the maximum parsimony (MP), the Bayesian inference (BI) and the maximum likelihood (ML) with bootstrap values or posterior probabilities above the branches, respectively. The bootstrap values less than 50% were represented by "-".

Figure 9

Nonsynonymous nucleotide substitution (Ka) and synonymous nucleotide substitution (Ks) of homologous protein-coding genes from *C. sinensis* var. *sinensis*, *C. sinensis* cultivar Longjing and *C. sinensis* cultivar Anhua.

Supplementary Files

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

- [TableS1.Pi.xlsx](#)
- [renamed42f8a.xlsx](#)
- [TableS3.KaKs.xlsx](#)

- [TableS4.Sitemodel.xlsx](#)