Comparative chloroplast genomes of the tea plants and the implications for the different origins of the two Assam teas

LI li (zizheng2006@163.com)
Wuyi University
Yunfei Hu
Wuyi University
Min He
Wuyi University
Bo Zhang
Wuyi University
Yongcong Hong
Wuyi University
Wei Wu
Wuyi University
Pumo Cai
Wuyi University
Da Huo
Wuyi University

Research article

Keywords: Tea plant, Camellia, Chloroplast genome, Domestication origins, Taxonomy

Posted Date: June 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36917/v1

Version of Record: A version of this preprint was published on February 26th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07427-2.

Abstract

Background

Tea plants belong to the genus *Camellia*, whose species are taxonomically complex due to frequent hybridization and polyploidy nature. The genetic genealogy of *Camellia* has always been a focus of botanical and ecological research, including a debate about whether Assam tea has two different domestication origins (Chinese Assam type and Indian Assam type). The chloroplast genome resources were able to provide useful data for the analysis of the plastome evolutionary relationship and species classification. Here, we determined the first chloroplast genome of the natural triploid tea plant (*Camellia sinensis cv. Wuyi Narcissus*) and conducted the genome comparison with Chinese type tea (*Camellia sinensis var. sinensis*), Chinese Assam type tea (*Camellia sinensis var. assamica*) and Indian Assam type tea (*Camellia assamica*) to improve our understanding of the evolutionary mechanism and the taxonomic classification of *Camellia*.

Results

This study presented detailed sequences and structural variations of chloroplast genomes of four tea plants. The chloroplast genome of the natural triploid tea showed no obvious sequence difference from that of other two types of Chinese teas, while that of Chinese tea and Indian tea was significant sequence difference. The natural selection probably dominated in shaping the codon bias of the chloroplast genome in tea plant, and the codon usage distribution of genome in Indian tea was obviously different from that in Chinese tea. The phylogenetic status of Chinese and Indian Assam teas was in the different branches of the tea plant. Phylogenetic tree clustering was not consistent with the current some taxonomy of *Camellia*.

Conclusions

The sequence variation of the chloroplast genome of tea plant was mainly ascribed to the expansion and contraction of the border regions (IR/ SC), which were mainly due to the sequence repeat and indel mutation events of the genome. The codon usage pattern and phylogenetic analysis supported Chinese Assam type and Indian Assam type tea might have different domestication origins and suggested the current some taxonomy of *Camellia* might need revision.

Introduction

Tea is the most popular non-alcoholic beverage with huge economic values in the world [1]. Botanically, cultivated tea is a member of the *camellia* family of angiosperms, which mainly comes from two different regions: China and India, so it can be mainly classified as Chinese type tea (*Camellia sinensis var. sinensis*) and Assam type tea (*Camellia sinensis var. assamica*) [2-4]. On the origin of cultivated tea domestication, it has long been suggested that Chinese and Assam type tea might have distinct origins, but the idea that Assam type tea consists of two distinct lineages (Chinese Assam type and Indian Assam type) that were domesticated separately is more controversial [5].

Chloroplast (cp) genomes are highly conserved in sequence and structure due to their non-recombinant, haploid, and uniparentally inherited nature [6]. Previous studies have found that the cp genome resources were able to provide useful data for the analysis of tea tree evolutionary relationship and had a good classification effect [7]. With the development of high-throughput sequencing technology, more than 30 complete cp genomes of *Camellia* species have been sequenced [8]. These massive data, together with the conservation of cp sequences, make it become a more increasingly used and effective tool for plant phylogenomic analysis than nuclear and mitochondrial genomes [9].

Because of frequent hybridization and polyploidization, the traditional taxonomic classification based upon the morphological characteristics is always incomplete for *Camellia* species identification [10]. Therefore, the genetic genealogy of *Camellia* has also always been a focus of botanical and ecological research [11,12]. However, the complete cp genome of polyploid tea plant has not been reported so far. In this study, the first completed cp genome of the natural triploid tea tree (*C. sinensis cv. Wuyi Narcissus*) from china was assembled and characterized. A good knowledge of its cp genomic information will contribute to germplasm resources protection and further phylogenetical studies. And then we systematically examined the sequence variation and codon usage pattern of cp genomes among *C. sinensis cv. Wuyi Narcissus* (*CWN*, Chinese type natural triploid tea), *C. sinensis var. sinensis* (*CSS,* Chinese type tea), *C. sinensis var. assamica* (*CSA,* Chinese Assam type tea) and *C. assamica* (*CIA,* Indian Assam type tea). Furthermore, a phylogenetic analysis was conducted using 37 complete cp genomes of *Camellia* species to improve our understanding of the evolutionary mechanism and the taxonomic classification of *Camellia*.

Materials And Methods

Plant material and DNA extraction

Young and healthy leaves of an individual plant of *CWN* were collected for DNA extraction from Wuyi University, Fujian Province, China. Fresh leaves were harvested and immediately frozen in liquid nitrogen after collection, followed by the preservation at -80 °C in the laboratory prior to DNA extraction. High-quality genomic DNA was extracted from leaves using CTAB extraction method. RNase A and proteinase K were separately used to remove RNA and protein contamination. The quality and quantity of the isolated DNA were checked by electrophoresis on a 0.8% agarose gel and a NanoDrop spectrophotometer (Thermo Scientific, Carlsbad, CA, USA), respectively.

Cp genome sequencing, assembly and annotation

In this study, cp genome was sequenced by PacBio with Illumina paired-end data support and quality checked by the NGS service at Sangon Biotech (Shanghai) Co., Ltd., China. For PacBio: the genome was sequenced using PacBio technology at PacBio Sequel platform. For Illumina: total DNA was used to generate libraries with the Illumina Novaseq 6000 platform. PacBio and Mate-pair reads were mapped against three published cp genomes, including two types of China tea tree (*CSS*, KJ806281.1; *CSA*, MH019307.1) and one India tea tree (*CIA*, MH460639.1) [7,8,13], using CLC Genomics Workbench 11.0.1 (CLC Bio, Arhus, Denmark) to filter out the chloroplast reads. The extracted chloroplast reads of *C. Wuyi Narcissus* were de-novo assembled by guidance-based assembly approach using CLC genomics workbench. The assembled contigs were ordered with reference cp genomes using ABACAS v1.3.1 (http://abacas.sourceforge.net/) to generate draft chloroplast assembly of *CWN*. The orientation and inverted repeat (IR) regions were checked with cpGAVAS (http://www.herbalgenomics.org/cpgavas) followed by manual curation and gap filling with extracted chloroplast reads to obtained the final cp genome of *C. Wuyi Narcissus*. To annotate the cp genome, we used initial annotation by cpGAVAS and verified the sequence coordinates of each of the annotated genes using BLAST search against annotated chloroplast genes available at NCBI. Annotation errors were manually corrected. The final annotated cp genome sequence of *C. Wuyi Narcissus* was subjected to OGDRAW software (http://ogdraw.mpimp-golm.mpg.de) to generate the circular cp genome map and deposited to NCBI GenBank. For the accuracy of the cp genome-wide comparison, the inverted repeat (IR) regions and the annotated data of three published cp genomes (*CSS*, *CSA* and *CIA*) was verified with the above methods.

Comparative analysis of cp genomes

Four cp genome sequences of *Camellia* species (Tab. 1) were aligned using MAFFT Version 7.017 [14] and adjusted manually where necessary. The cp genome sequence divergences between four *Camellia* species were compared and plotted using mVISTA program [15]. Genetic divergence parameter (p-distance) was calculated by MEGA 6.0 [16]. A sliding window analysis was conducted to compare π among the chloroplast genomes, using DnaSP v5.0 [17]. The window length was 600 bp with a 200 bp step size. The percentage of variable characters for coding and noncoding regions in the genome was calculated as described previously [18].

Repeat sequences were searched by REPuter [19] with a minimal size of 30 bp and >90% identity (Hamming distance equal to 3) between the two repeats. Gap size between the repeats was restricted to a maximal length of 3 kb. Tandem repeats were identified by Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) [20] with default settings. Simple sequence repeats (SSRs) were predicted using MISA (http://pgrc.ipk-gatersleben.de/misa/) with the parameters: monomer (one nucleotide, n ≥ 8), dimer (two nucleotides, n ≥ 4), trimer (three nucleotides, n ≥ 4), tetramer (four nucleotides, n ≥ 3), pentamer (five nucleotides, n ≥ 3), hexamer (six nucleotides, n ≥ 3).

Codon Usage Bias Analyses

In order to avoid sampling bias, each CDS in cp genome were checked for being full-length and for the presence of proper start and stop codons. CDS shorter than 300 bp in length were excluded in codon usage calculations [21]. GC3s, ENc, CAI and RSCU for CDS were calculated using CodonW v1.4.4 [22].

ENc value is a measure of general non-uniformity of usage within synonymous groups of codons, ranging from 20 (extreme bias where only one codon is used in each amino acid) to 61 (random codon usage) [23]. ENc plot analysis (ENc vs GC3s) was used to examine whether the codon usages were affected only by mutation or other factors. If codon usage is constrained only by mutation pressure, ENc value lie on or slightly below the expected curve, and if codon usage is subject to natural selection, ENc value will lie considerably below the expected curve [24].

Neutrality plot (GC12 vs. GC3) was used to investigate the effects of mutation pressure and natural selection on codon use patterns. GC12 and GC3 were calculated by Perl script. GC3 was calculated excluding the three termination codons (TAA, TAG and TGA) and the three codons for Ile (ATT, ATC and ATA). Meanwhile, two single codons for Met (ATG) and Trp (TGG) were also excluded in all three patterns [25]. The slope of the plot regression is zero which indicates that there was no effect on directional mutation pressure (complete selection constraint). Slope 1 indicates that the codon usage bias is completely affected by the directional mutation pressure, and represented complete neutrality [26].

RSCU value for a particular codon refers to the ratio of its actual usage frequency to expected frequency when it is used without bias. The preferred codons with RSCU > 1.0 occur when they are used with higher frequencies than random, and the rare codons with RSCU < 1.0 means the opposite [27]. The distribution of codon usage for the four species was shown in the form of a heatmap using HemI 1.0 [28], according to the RSCU value.

CAI value is widely used to evaluate the gene expression level and ranges from 0 to 1. The larger the CAI value, the stronger the codon usage bias, otherwise, the weaker the codon usage bias [29]. The chloroplast genes in the upper and lower 5% of CAI values were respectively defined as the high- and low-expression gene datasets. A statistical chi-squared test was performed with SPSS 18.0 to compare the RSCU values between two datasets. If a codon whose frequency in the high-expression genes was significantly higher ($p < 0.05$) than in the low-expression genes, it will be classified as an optimal codon [30].

Phylogenomic Analyses

The cp genome sequences of 37 *Camellia* species and one outgroup (*Apterosperma oblata*) were aligned with the program MAFFT version and adjusted manually when necessary. Maximum likelihood (ML) analyses were implemented in RAxML version 7.2.6 [31]. RAxML searches relied on the general time reversible (GTR) model of nucleotide substitution with the gamma model of rate heterogeneity. Non-parametric bootstrapping test was implemented in the “fast bootstrap” algorithm of RAxML with 1000 replicates. Bayesian analyses were performed using the program MrBayes version 3.1.2 [32]. The best-fitting models were determined by the Akaike Information Criterion [33] as implemented in the program Modeltest 3.7 [34]. In all analyses, *A. oblata* was set as an outgroup.

Results And Discussion

Cp genome sequencing and assembly

PacBio (10838 long reads, >5kb) and Mate-pair reads (7.04G of raw data were produced with 2×150 bp pair-end read lengths) were used to assemble the cp genome of *CWN* into a circular contig of 156,762 bp length. Circular genome maps were drawn with OGDRAW software (Fig. 1). Raw reads have been deposited in the NCBI Sequence Read Archive (SRA, SRR12002624). Assembled cp genome sequences and accompanying gene annotations of *C. sinensis cv. Wuyi Narcissus* have been deposited in the NCBI GenBank (Accession numbers: MT612435).

Comparative analysis of four cp genomes

Gene content

All four complete *Camellia* cp genomes displayed the typical quadripartite structure of most angiosperms, including the large single copy (LSC), the small single copy (SSC) and a pair of inverted repeats (IRa and IRb). Among these cp genomes, genome size ranged from 156,762 bp (*CWN*) to 157,353 bp (*CIA*). The length varied from 86,301 bp (*CWN*) to 87,214 bp (*CIA*) in the LSC region, from 18,079 bp (*CWN*) to 18,285 bp (*CSA*) in the SSC region, and from 26,030 bp (*CIA*) to 26,090 bp (*CSS* or *CWN*) in IR region (Tab. 1).

Four plastomes are highly conserved in gene content, gene order, and intron number. Each cp genome contained a total of 137 genes, including 92 protein-coding genes, 37 transfer RNA (tRNA) genes and 8 ribosomal RNA (rRNA) (Supplementary Tab. S1). Of them, 60 protein-coding and 22 tRNA genes were located within LSC, 16 protein-coding genes, 14 tRNA coding genes and eight rRNA coding genes were located within IRs and 11 protein-coding and one tRNA gene were located within SSC. The rps12 gene was a divided gene with the 5' end exon located in the LSC region while two copies of 3' end exon and intron were located in the IRs. The ycf1 was located in the boundary regions between IRa/SSC, leading to incomplete duplication of the gene within IRs. The rps19 genes in *CSS*, *CSA*, and *CWN* were crossed the LSC/IR region while in *CIA* was located within LSC. There were 18 genes containing introns, including 6 tRNA genes and 12 protein-coding genes. Except for two introns in the ycf3 and clpP genes, all other genes contained only one intron. MatK gene was located within the intron of trnK-UUU with the largest intron (2,489 bp). Overlaps of adjacent genes were found in the complete genome, rps3-rpl22, atpB-atpE, and psbD-psbC had a 16 bp, 4 bp, and 53 bp overlapping region, respectively. Unusual initiator codons were observed in rps19 with GTG and orf42 with ATC in four cp genomes. The initiation codon of ndhD in *CIA* was ATG, while that of other three cp genomes was GTG.

Sequence variation and IR expansion/contraction

To elucidate the level of the genome divergence, the sequences were plotted to check their identity using the program mVISTA by aligning the four *Camellia* cp genomes with *CWN* (Fig. 2) as a reference. The whole aligned sequences showed high similarities with only a few regions below 90%, suggesting that tea plant plastomes were rather conserved (Figure 4).

Sequence divergence analyses of four cp genomes revealed Pi values in the range from 0 to 0.01917 with an average of 0.00093, indicating moderate genetic divergence existed within the four cp genomes. However, four regions (including rp12/trnH-UGU, psaA/ycf3, atpB/rbcL and psbT/psbH) had relatively higher divergence values (Pi > 0.006) (Fig. 3). These four regions were all located within LSC, indicating that the LSC region had more gene mutations than the IR region and the SSC region.

Mutations may cause changes in the length of the coding gene sequence, leading to changes in the coding and non-coding regions. Therefore, the number and distribution patterns of variable characters in coding and non-coding regions among four cp genomes were further analyzed. The result showed that the proportion of variability in non-coding regions was with a mean value of 1.82%, while in the coding regions was 1.15%. Five coding genes had over 4% variability proportion, such as rps19, ndhF, ndhD, ndhI and ycf1. Five non-coding regions had over 10% variability proportions, such as rpl12/trnH-GUG, trnE-UUC/trnT-GGU, ndhD/psaC, ndhI/ndhA and rps15/ycf1(Fig. 4). These divergence hotspot regions might provide information for marker development in phylogenetic analyses of tea plants, but further verification is needed. Among them, the coding gene rps19, ndhF, ycf1 and non-coding region rpl2/trnH-GUG, rps15/ycf1 were located in the junctions of IR/SC region, which supported that the length of angiosperm cp genomes was variable primarily due to the expansion and contraction of IR/SC boundary regions [35].

To further elucidate the potential contraction and expansion of IR regions, the gene variation at the IR/SSC and IR/LSC boundary regions of the four plastomes was compared (Fig. 5). The genes rps19, ycf1-5'end/ndhF, ycf1 and rp12/trnH-GUG were located in the junctions of LSC/IR and SSC/IR regions. The rps19 gene in *CSS*, *CSA*, and *CWN* was 279 bp, and crossed the LSC/IRa region by 46 bp while the rps19 gene in *CIA* was just 150 bp, and all located in the LSC region, 1bp away from the IRa region. The ycf1-5'end gene in *CSS*, *CSA*, and *CWN* was 1071 bp, and crossed the IRa/SSC region by 2 bp while in *CIA* was 1065 bp, and crossed the IR/SSC region by 33 bp. The ndhF gene in all four cp genomes was located in the SSC region. The ndhF gene in *CSA*, *CIA*, and *CWN* was 2247 bp while in *CSS* was 2139. The ndhF gene in *CSS* was 165 bp away from the IRa region, in *CSA* or *CWN* was 57 bp away from the IRa region while in *CIA* was 88 bp away from the IRa region. The ycf1 gene in *CSS* or *CWN* was 5622 bp, in *CSA* was 5628 bp while in *CIA* was only 1038 bp. The ycf1 genes in all four cp genomes crossed the IR/SSC region. The ycf1 gene in *CSS* or *CWN* was with 4553 bp located in the SSC region and 1069 bp in IRb region, in *CSA* was with 4559 bp located in the SSC region and 1069 bp in IRb region while in *CIA* was with only 6 bp located in the SSC region and 1032 bp in IRb region. The rpl2 gene in *CSS*, *CSA* or *CWN* was 107 bp away from the LSC region while in *CIA* was 82 bp away from the LSC region. The trnH-GUG gene in *CSS*, *CSA* or *CWN* was 2 bp away from the IRb region while in *CIA* was 637 bp away from the IRb region.

Among four species, three Chinese tea varieties (*CSS*, *CSA*, and *CWN*) were similar in both gene sequence and IR/LSC boundary pattern, except for length variations in ndhF and ycf1. The triploid *CWN* was closer to *CSS*, indicating that triploidy did not cause significant changes in cp genome. However, there were relatively obvious differences in IR/LSC boundary of cp genomic between Chinese tea and Indian tea, suggesting that environmental selection may be responsible for the differences.

Repeat and indel sequence analyses

SSRs are small repeating units of cpDNA, and have been widely used to characterize genetic variation among plant genotypes [36-38]. A total of 671 SSRs were identified in four cp genomes, of which 57% were in IGS, 34% were in CDS, and 9% were in Intron (Fig. 6C). 74.0% of these SSRs were monomers, 19.3% of dimers, 0.5% of trimers, 5.3% of tetramers, 0.9% of hexamers and no pentamers found (Fig. 6A). Comparing the four genomes, except for 167 SSRs of *CIA*, the other three were all 168. A total of 128 SSRs were fully shared among four cp genomes (Fig. Additional file 1: Table). There were 47 loci with different SSR types, most of which existed in the LSC region. Among them, *CSS* had 7 unique types, *CSA* had 18 unique types, *CIA* had 9 unique types, and *CWN* had 14 unique types. (Fig. 6B, Supplementary Tab. S2).

Long repeat sequences of plastomes had been reported to play roles in genomic rearrangement and sequence variation [39,40]. In total, 270 repeats were detected in four plastomes, including three categories of long repeats: tandem, forward and palindromic. The number of the three repeated types is consistent in *CSS* and *CWN*, as follows: 23, 20, 23. However, it is 19, 20, 23 in *CSA* and 21, 23, 32 in *CIA*. The sizes of repeats ranged from 11 to 82 bp (Fig. 7A, 7C). The four cp genomes have a total 57 identical repeat sequences. In addition, *CSS* had 1 unique repeat, *CIA* had 18 unique repeats, *CWN* had 2 unique repeats, while *CSA* had no unique repeats (Fig. 7B). These unique repeats were found mainly in the intergenic psaA/ycf3, atpB/rbcL, trnW-CCA/ trnP-UGG, rps19/rpl2, psbT/psbN, rpl2/trnH-GUG and gene rpl2, ycf1, ycf2. Only one repeat was in the intron regions (ndhA) (Supplementary Tab. S3).

Indels played an important role in sequence evolution [41]. In indels analysis, the indel events in simple sequence repeats were filtered out. By comparing four cp genomes, a total 67 indels were found. Indels ranged in size from 1 to 637 bp (Fig. 8A), and most of the Indels events occurred in IGS regions (72%), with 24% in CDS regions and only 4% in Intron regions (Fig. 8B). As expected, single-nucleotide Indels (1 bp) were the most common, but 16 long Indels (>10bp) were found. The longest one was an insertion of 637 bp in *CIA* (intergenic rp12/trnH-GUG), followed by a 335 bp deletion in *CWN* (intergenic trnE-UUC/trnT-GGU) and a 107 bp deletion in *CIA* (gene rps19). The largest number of long inserts were found in intergenic psaA/ycf3, where *CSA* had a unique 10 bp deletion, *CIA* had a unique 12 bp insertion, and *CSA* and *CWN* had a common 17 bp deletion. Secondly, two deletions occurred in the gene coding region of nhdA in *CIA*, which were respectively 53bp and 66bp deletions (Fig. 8C, Supplementary Tab. S4).

The above results showed that the divergent regions of cp genomes were almost all associated with the long repeat sequences (intergenic rp12/trnH-UGU, atpB/rbcL, psbT/psbN and gene ycf1) and the indel sequences (intergenic rp12/trnH-UGU, psaA/ycf3, ndhI/ndhA, trnE-UUC/trnT-GGU, psbN/psbH and gene rps19, ndhF, ndhD, ndhI, ycf1, ycf2). Furthermore, these regions also contained genes from all IR/SC boundary (rps19, ycf1-5'end/ndhF, ycf1 and rp12/trnH-GUG). Therefore, the sequence repeats and indel events might play an important role in the expansion and contraction of the border regions. These repeats and indels might be further served as genetic markers for phylogenetic and population genetic studies [42,43].

Codon usage pattern analyses

Codon use bias can be used to reflect the origin, evolution and mutation mode of species or genes [44]. The ENc plots showed only a few points lie near the curve, however, most of the genes with lower ENc values than expected values lay below the curve (Fig. 9). Therefore, the codon usage bias of the chloroplast genome was slightly affected by the mutation pressure, but natural selection and other factors play an important role.

To further investigate the extent of influence between mutation pressure and natural selection on the codon usage patterns, Neutrality plot (GC12 vs. GC3) was performed. The correlation between GC1 and GC2 was strong (*CSS*: r = 0.445; *CSA*: r = 0.453; *CIA*: r = 0.445; *CWN*: r = 0.464, p <0.01). However, no significant correlation was found for GC1 with GC3 (*CSS*: r =0.141; *CSA*: r = 0.139; *CIA*: r = 0.078; *CWN*: r = 0.141) or GC2 with GC3 (*CSS*: r = 0.146; *CSA*: r = 0.143; *CIA*: r = 0.078; *CWN*: r = 0.152), which suggested mutation pressure had a minor effect on the codon usage bias. Moreover, the slope of Neutrality plot showed that mutation pressure accounts for only 0.52% - 8.42% on the codon usage patterns in four cp genomes, while natural selection accounts for 91.58% - 99.48% (Fig. 10). These results further suggested that natural selection played an important role in the codon usage patterns.

The patterns of codon usage are strongly correlated with GC content [45]. The overall GC content was almost identical with each other among the four cp genomes, about 37.3%, indicating a higher AU content. Furthermore, the distributions of codon usage in the heatmap for 4 *Camellia* species showed that about half of the codons with low RSCU values were infrequently used and almost all codons with RSCU >1 ended with A/U (Fig. 11). These data indicated that the four cp genomes tended to use A/U bases and A/U-ending codons. In addition, we also found that the codon usage patterns of the three Chinese teas were more similar, and the cp genome of *CWN* has a closer relationship with that of *CSS* in RSCU cluster analysis. In contrast, the codon usage patterns of Indian tea (*CIA*) were quite different from that of Chinese teas. The RSCU values of the 36 codons (36/64, 56.25%) were identical in the three Chinese teas, but different from those in Indian tea (Fig. 11, Tab. 2), indicative of a different selection on codon usage in the cp genomes between Chinese tea and Indian tea. This also suggested that the two Assam tea plants (Chinese Assam type and Indian Assam type) might undergo different domestications.

Phylogenetic analysis

To further understand the evolutionary mechanism and taxonomic classification of *Camellia*, a most comprehensive phylogenetic analyse was performed based on complete cp genome from 37 *Camellia* species with *Apterosperma. oblata* as an outgroup so far, including Chinese type tea (*CSS*), Chinese Assam type tea (*CSA*) and Indian Assam type tea (*CIA*) and a nature triploid species (*CWN*) (Fig. 12). Phylogenetic trees were generated by Bayesian inference (BI) and Maximum likelihood (ML) based on the complete cp genome sequence have similar topologies, and almost all relationships have high bootstrap values. Both trees showed that natural triploid *CWN* was formed into a monophyletic clade with a bootstrap value of 100%, suggesting a polyploid event in the evolution of tea plants. Further, our results clearly indicated that the ancestors of tea plants differentiated into two branches at a certain node, with one branch continuing to differentiate into Chinese Assam type tea and the other branch continuing to differentiate into Indian Assam type tea and Chinese type tea, respectively. This result suggested that these three species might have different domestication origins, which supported an earlier idea that there are three independently domesticated tea species. The one is Chinese type tea, and the other two are distinct Assam tea: a Chinese tea from the southwestern province

of Yunnan (Chinese Assam type tea) and an Indian tea from the Assam region (Indian Assam type tea) [46]. The existence of two distinct domestication origins of Assam tea had long been a major topic of debate [5]. Our study further suggested that there were two distinct domesticated origins of Assam tea.

Because of the frequent hybridization and other factors, the classification of the genus *Camellia* based on morphology had been controversial. Some previous studies had shown that cp genomes can provide useful information for solving complex taxonomic problem of *Camellia* species [7,47]. In terms of morphological classification, Chang et al. classified the genus *Camellia* into 4 subgenera with 22 sections [48], while Ming et al. revised the classification of Chang and classified the genus *Camellia* into 2 subgenera with a total of 14 sections. In Ming et al. classification revision, the section *Chrysantha* established in Chang's classification was merged into the section *Archecamellia*, and the section *Heterogenea* was separately established [49, 50]. Corresponding to Ming's classification, our results showed that not all subgenus *tea* or subgenus *camellia* were individually clustered, suggesting that this classification needed to be reconsidered to revise: (i). at the subgenus level, both BI tree and ML tree (BS=100%) showed that *C. danzaiensis* should belong to subgenus *camellia*, rather than subgenus *tea*, which was consistent with a previous study by Huang et al. 2014. In addition, (ii). *C. nitidissima* and *C. petelotii* of sect. *Archecamellia* of subgenus *tea* (Syn. sect. *Chrysantha* Chang) were clustered with *C. szechuanensis* of sect. *Heterogenea* of subgenus *Camellia* in BI tree (BS=88%), and not all species of sect. *Heterogenea* clusters came together. This result supported a previous study which pointed out that sect. *Chrysantha* Chang should not be merged into sect. *Archecamellia*, and sect. *Heterogenea* should not be recognized in taxonomic treatments of *Camellia* species by analyzing the secretory structure [51]. Moreover, (iii). The positions of *C. yunnanensis*, *C. luteoflora and C. tachangensis* were with low bootstrap values in ML tree, which were also in doubt and need further verification. More cp genomes of *Camellia* would be needed for further studies to solve these problems as phylogenomic analysis tends.

Conclusion

The first complete cp genome of natural triploid tea tree and its phylogenetic status were determined. The specific genome features, including repeat sequences, SSRs, and indels were compared among the natural triploid tea, Chinese type tea, Chinese Assam type tea and Indian Assam type tea. The identified divergence regions could be used as potential molecular markers for further analysis. This study is also the first to provide information on the domestication origin of tea plant based on the complete cp genome. Our results supported that Chinese Assam type tea and Indian Assam type tea might have different domestication origins. Moreover, phylogenetic tree analysis also suggested that the current taxonomy of *Camellia* species might need further revision. Our results might facilitate germplasm resources protection for these economically important tea plants, and offer useful genetic information for the purposes of phylogenetics, taxonomy and species identification in the genus *Camellia*.

Abbreviations

NGS: Next-Generation Sequencing;

ABACAS: Algorithm Based Automatic Contiguation of Assembled sequences;

cpGAVAS: an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences;

OGDRAW: Organellar Genome DRAW soft;

CLC Genomics Workbench: a toolkit for NGS data analysis;

Indel: Insertion and deletion;

CDS: Coding DNA sequence; CNS: Conserved noncoding sequence;

ENc: Effective number of codons;

GC3s: GC content at the third synonymously variable coding position;

GC1: the GC content at the first position; GC2: the GC content at the second position;

GC12: the average value of GC contents at the first and second positions of codon;

GC3: the GC content at the third position; CAI: the codon adaptation index;

RSCU: Relative synonymous codon usage; SC: Single-copy; LSC: Large single copy;

SSC: Small single copy; IR: Inverted repeat; SSR: Simple sequence repeat;

Pi: Nucleotide diversity; BI: Bayesian inference; ML: Maximum likelihood.

Declarations

Acknowledgments

We are very grateful to Bo Zhang and Yongcong Hong (Germplasm Bank of Tea Species in Wuyi University) for their help with experiments and data analyses. We also thank Department of Mathematics and Computer Science in Wuyi University for providing the computer resources.

Funding

This study is financially supported by Scientific Research Launch Fund of Wuyi University (YJ201902), Discipline Team Construction Fund of Wuyi University (Wu Yuan Zong [2017] No. 66), Construction Fund of Wuyi Tea Industrial Technology Research Institute (2018N2004) and the Open Fund of The Key Laboratory of Cognitive Computing and Intelligent Information Processing of Fujian Education Institutions (No: KLCCIIP2018104).

Availability of data and materials

The chloroplast genomes generated during the current study were deposited in NCBI with accession number MT612435 (*C. sinensis cv. Wuyi Narcissus*), and Raw reads have been deposited in the NCBI Sequence Read Archive (SRA, SRR12002624).

Authors' contributions

LL designed the experiments. LL, YFH and MH performed the experiments. LL, YCH, WW, PMC and DH analyzed data. LL wrote the paper. All authors reviewed and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[a] College of Tea and Food Science, Wuyi University, 358# Baihua Road, Wuyishan 354300, China

[b] Department of Mathematics and Computer Science, Wuyi University, 358# Baihua Road, Wuyishan 354300, China.

References

1. Mondal TK, Bhattacharya A, Laxmikumaran M, Ahuja PS. Recent advances in tea (*Camellia sinensis*) biotechnology. Plant Cell Tiss Org. 2004; 76: 195–254.
2. Wight Tea Classification Revised. Curr Sci. 1962; 31: 298–299.
3. Benerjee B. Botanical classification of tea. In: Wilson K, Clifford N ed. Tea: cultivation and consumption. Chapman and Hall, London. 1992.
4. Ming TL, Bartholomew B. Theaceae. In: Flora of China Volume 12. Wu ZY, Raven PH, Hong DY. editors. Science Press and Missouri Botanical Garden Press, Beijing, St. Louis. 2007; pp. 366–478.
5. Drew Genetic studies of today's tea trees are providing clues to how the plant was first domesticated. Nature. 2019. DOI: 10.1038/d41586-019-00395-4.
6. Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011; 76: 273–297.
7. Huang H, Shi C, Liu Y, et al. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships[J]. BMC Evolutionary Biology. 2014; 14(1): 151.
8. Zhang F, Li W, Gao C, et al. Deciphering tea tree chloroplast and mitochondrial genomes of *Camellia sinensis var. assamica*. Sci Data. 2019; 6: 209. https://doi.org/10.1038/s41597-019-0201-8.
9. Zeng S, Zhou T, Han K, Yang Y, Zhao J, Liu ZL. The Complete Chloroplast Genome Sequences of Six *Rehmannia* Genes. 2017; 8: 103.
10. Lu H, Jiang W, Ghiassi M, Lee S, Nitin M. Classification of *Camellia* (*Theaceae*) species using leaf architecture variations and pattern recognition techniques, PLoS One. 2012; 7: e29704. https://doi.org/10.1371/journal.pone.0029704.
11. Meegahakumbura MK. Genetic Assessment of Asian Tea Germplasm and the Domestication History of the Tea Plant (*Camellia sinensis*). PhD dissertation, University of Chinese Academy of Sciences, Beijing. 2016.
12. Yang H, Wei CL, Liu HW, Wu JL, Li ZG, Zhang L, Jian JB, Li YY, Tai YL, Zhang J, Zhang ZZ, Jiang CJ, Xia T, Wan XC. Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome wide SNPs from RAD sequencing. PLoS One. 2016; 11: e0151424.
13. Rawal HC, Kumar PM, Bera B, Singh NK, & Mondal TK. Decoding and analysis of organelle genomes of Indian tea (*camellia assamica*) for phylogenetic confirmation. Genomics. 2020; 112(1): 659-668. https://doi.org/10.1016/j.ygeno.2019.04.018.
14. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 2013; 30: 772–780.

15. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: Computational tools for comparative genomics. Nucleic Acids Res. 2004; 32 (Suppl. 2): W273–W279.
16. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. 2013; 30: 2725–2729.
17. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11): 1451–2.
18. Zhang YJ, Ma PF, Li DZ. High-Throughput Sequencing of Six Bamboo Chloroplast Genomes: Phylogenetic Implications for Temperate Woody Bamboos (*Poaceae*: *Bambusoideae*). PLoS One. 2011; 6: e20596.
19. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29: 4633–4642.
20. Benson G. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. 1999; 27: 573–580.
21. Rosenberg MS, Subramanian S, Kumar S. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol. 2003; 20: 988–993.
22. Peden JF. Analysis of codon usage. PhD thesis. UK: University of Nottingham. 1999.
23. Wu Y, Li Z, Zhao D, Tao J. Comparative analysis of flower-meristem-identity gene APETALA2 (AP2) codon in different plant species. Journal of Integrative Agriculture. 2018; 17: 867-877. DOI: 10.1016/S2095-3119(17)61732-5.
24. Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87: 23–29
25. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proceedings of the National Academy of Sciences of the United States of America. 1988; 85: 2653-2657. DOI 10.1073/pnas.85.8.2653.
26. Wen Y, Zou Z, Li H, Xiang Z, He N. Analysis of codon usage patterns in *Morus notabilis* based on genome and transcriptome data. Genome. 2017; 60: 473-484. DOI: 10.1139/gen-2016-0129.
27. Gupta SK, Bhattacharyya TK, Ghosh TC. Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. J Biomol Struct Dyn. 2004; 21: 527–536.
28. Deng WK, Wang YB, Liu ZX, Cheng H, Xue Y. HemI: a toolkit for illustrating heatmaps. PloS One. 2014; 9(11): e111988. DOI: 10.1371/journal.pone.0111988.
29. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution. 1986; 24: 28-38. DOI: 10.1007/BF02099948.
30. Liu Q. Analysis of codon usage pattern in the radioresistant bacterium Deinococcus radiodurans. Biosystems. 2006; 85: 99–106.
31. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22: 2688–2690.
32. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19: 1572–1574.
33. Posada D, Buckley TR. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches over Likelihood Ratio Tests. Syst. Biol. 2004; 53: 793–808.
34. Posada D, Crandall KA. Modeltest: Testing the model of DNA substitution. Bioinformatics. 1998; 14: 817–818.
35. Kim KJ, Lee HL. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng Nees*) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 2004; 11: 247–261.
36. Gaudeul M, Giraud T, Kiss L, Shykoff JA. Nuclear and chloroplast microsatellites show multiple introductions in the worldwide invasion history of common ragweed, Ambrosia artemisiifolia. PloS One. 2011; 6(3): e17658.
37. Piya S, Nepal MP. Characterization of nuclear and chloroplast microsatellite markers for *Falcaria vulgaris* (*Apiaceae*). Am J Plant Sci. 2013; 4: 590–5.
38. Redwan R, Saidin A, Kumar S. Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass *Commelinidae*. BMC Plant Biol. 2015; 15: 196.
39. Cavalier-Smith Chloroplast evolution: Secondary symbiogenesis and multiple losses. Curr. Biol. 2002; 12: R62–R64.
40. Timme RE, Kuehl JV, Boore JL, Jansen RK. A comparative analysis of the Lactuca and Helianthus (*Asteraceae*) plastid genomes: Identification of divergent regions and categorization of shared repeats. Am. J. Bot. 2007; 94: 302–312.
41. Britten RJ, Rowen L, Williams J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci USA. 2003; 100: 4661–4665.
42. Gao L, Yi X, Yang YX, Su YJ, Wang T. Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. BMC Evol Biol. 2009; 9: 130–144.
43. Bapteste E, Philippe H. The potential value of indels as phylogenetic markers: position of Trichomonads as a case study. Mol Biol Evol. 2002; 19: 972–977.
44. Pyo YJ, Kwon KC, Kim A, Cho MH. Seedling Lethal1, a pentatricopeptide repeat protein lacking an E/EC or DYW domain in Arabidopsis, is involved in plastid gene expression and early chloroplast development. Plant Physiology. 2013; 163: 1844-1858. DOI: 10.1104/pp.113.227199.
45. Shackelton LA, Parrish CR, Holmes EC. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. Journal of Molecular Evolution. 2006; 62: 551-563.
46. Meegahakumbura MK, Wambulwa MC, Thapa KK, et al. Indications for Three Independent Domestication Events for the Tea Plant (*Camellia sinensis (L.) O. Kuntze*) and New Insights into the Origin of Tea Germplasm in China and India Revealed by Nuclear Microsatellites. Plos one. 2016; 11(5): e0155369. DOI: 10.1371/journal.pone.0155369.

47. Yang JB, Yang SX, Li HT, Yang J, Li DZ. Comparative chloroplast genomes of *Camellia* PLoS One. 2013; 8: e73053.
48. Chang HD, Ren SX. Flora of China. Science Press. Tomus. 1998; 49(3): 1–251.
49. Ming Monograph of the genus *Camellia*, Yunnan Science and Technology Press. Kunming. 2000.
50. Ming TL, Bruce B. Flora of China. Beijing, China: Science Press. 2010.
51. Jiang B, Peng QF, Shen ZG, et al. Taxonomic treatments of Camellia (Theaceae) species with secretory structures based on integrated leaf characters [J]. Plant Systematics and Evolution. 2010; 290(1-4): p.1-20.

Tables

Table 1. Summary of four chloroplast genome features.

Genome Features	*CWN* (MT612435)	*CSS* (KJ806281.1)	*CSA* (MH019307.1)	*CIA* (MH460639.1)
Location of sample	Fujian, China	Yunnan, China	Yunnan, China	Assam, India
Longitude	118.004001	102.714601	102.714601	94.228661
Latitude	27.72846	25.04915	25.04915	26.73057
Genome size (bp)	156762	157117	157100	157353
LSC length (bp)	86301	86662	86649	87214
SSC length (bp)	18281	18275	18285	18079
IR length (bp)	26090	26090	26083	26030
Number of genes	137	137	137	137
Number of Protein-coding genes	92	92	92	92
Number of tRNA genes	37	37	37	37
Number of rRNA genes	8	8	8	8
GC content of LSC (%)	35.32	35.31	35.31	35.38
GC content of SSC (%)	30.55	30.56	30.51	30.59
GC content of IR (%)	42.94	42.95	42.95	42.96
overall GC content (%)	37.3	37.3	37.29	37.34

CWN, C. sinensis cv. wuyi narcissus; *CSS, C. sinensis var. sinensis*;

CSA, C. sinensis var. assamica; *CIA, C. assamica*

Table 3. The relative synonymous codon usage (RSCU) values of four chloroplast genomes. The optimal codon, marked with * ($p < 0.05$) and @ ($p < 0.01$), was defined as a codon whose usage frequency in the high-expression genes was significantly higher than in the low-expression genes by the chi-squared test. The RSCU values are identical in all three Chinese teas (Blue background), but different in Indian teas (Yellow background).

AA	Condon	Species				AA	Condon	Species				AA	Condon	Species		
		CSS	*CSA*	*CIA*	*CWN*			*CSS*	*CSA*	*CIA*	*CWN*			*CSS*	*CSA*	*CIA*
Phe	UUU	1.32	1.32	1.3	1.32	Pro	CCU	1.65@	1.66@	1.69@	1.65@	Lys	AAA	1.53	1.53	1.51
	UUC	0.68	0.68	0.7	0.68		CCC	0.71	0.71	0.7	0.71		AAG	0.47	0.47	0.49
Leu	UUA	1.97	1.97	1.96	1.98		CCA	1.17	1.16	1.14	1.16	Asp	GAU	1.63*	1.63	1.62
	UUG	1.24	1.23	1.25	1.24		CCG	0.47	0.47	0.47	0.47		GAC	0.37	0.37	0.38
	CUU	1.25	1.25	1.24	1.25	Thr	ACU	1.66@	1.66@	1.68@	1.66@	Glu	GAA	1.53@	1.54	1.51
	CUC	0.39	0.39	0.4	0.39		ACC	0.74	0.74	0.74	0.74		GAG	0.47	0.46	0.49
	CUA	0.78	0.78	0.8	0.77		ACA	1.22	1.22	1.2	1.22	Cys	UGU	1.52	1.52	1.51
	CUG	0.37	0.37	0.35	0.37		ACG	0.38	0.38	0.38	0.38		UGC	0.48	0.48	0.49
Ile	AUU	1.46**	1.46**	1.48	1.46	Ala	GCU	1.84	1.84	1.85	1.84	Trp	UGG	1	1	1
	AUC	0.58	0.58	0.59	0.58		GCC	0.62	0.62	0.63	0.63	Arg	CGU	1.39@	1.39@	1.44@
	AUA	0.96	0.96	0.93	0.96		GCA	1.16	1.15@	1.14	1.16		CGC	0.33	0.34	0.35
Met	AUG	1	1	1	1		GCG	0.38	0.38	0.39	0.38		CGA	1.42	1.42	1.44
Val	GUU	1.49	1.48	1.48	1.49	Tyr	UAU	1.62	1.62	1.63	1.62		CGG	0.39	0.39	0.4
	GUC	0.43	0.43	0.44	0.43		UAC	0.38	0.38	0.37	0.38	Ser	AGU	1.31	1.31	1.33
	GUA	1.53	1.54**	1.53**	1.53@	TER	UAA	1.56	1.5	1.5	1.5		AGC	0.32	0.32	0.33
	GUG	0.55	0.55	0.56	0.55		UAG	0.63	0.63	0.63	0.63	Arg	AGA	1.86	1.85	1.75
Ser	UCU	1.78**	1.78**	1.8@	1.79@		UGA	0.81	0.87	0.87	0.87		AGG	0.6	0.61	0.62
	UCC	0.91	0.91	0.92	0.91@	His	CAU	1.57	1.57	1.56	1.57	Gly	GGU	1.29@	1.29@	1.31
	UCA	1.16	1.16	1.11	1.16		CAC	0.43	0.43	0.44	0.43		GGC	0.45	0.45	0.45
	UCG	0.51	0.51	0.52	0.51	Gln	CAA	1.53	1.53	1.51	1.53		GGA	1.59	1.59	1.57
Asn	AAU	1.57	1.57	1.55	1.57		CAG	0.47	0.47	0.49	0.47		GGG	0.67	0.67	0.67
	AAC	0.43	0.43	0.45	0.43											

Figures

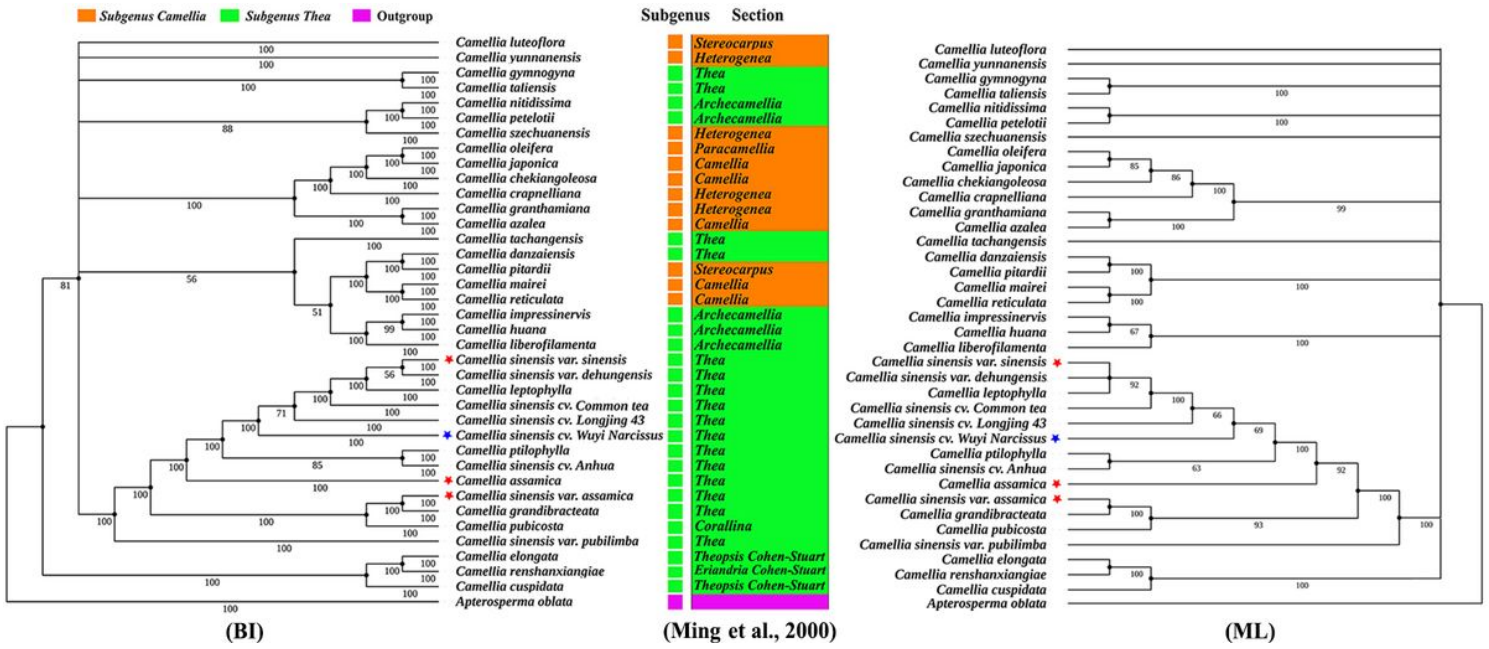

Figure 1

Phylogenetic relationships of 37 Camellia species based on complete chloroplast genome sequences with A. oblata as outgroup. (A) Bayesian tree (B) Maximum likelihood tree The bootstrap support values (>50%) were shown above the branches. C. sinensis cv. wuyi Nacrissus was highlighted with blue star mark. C. sinensis var. sinensis, C. sinensis var. assamica and C. assamica were highlighted with red star mark. Species classification was based on Ming et al., 2000.

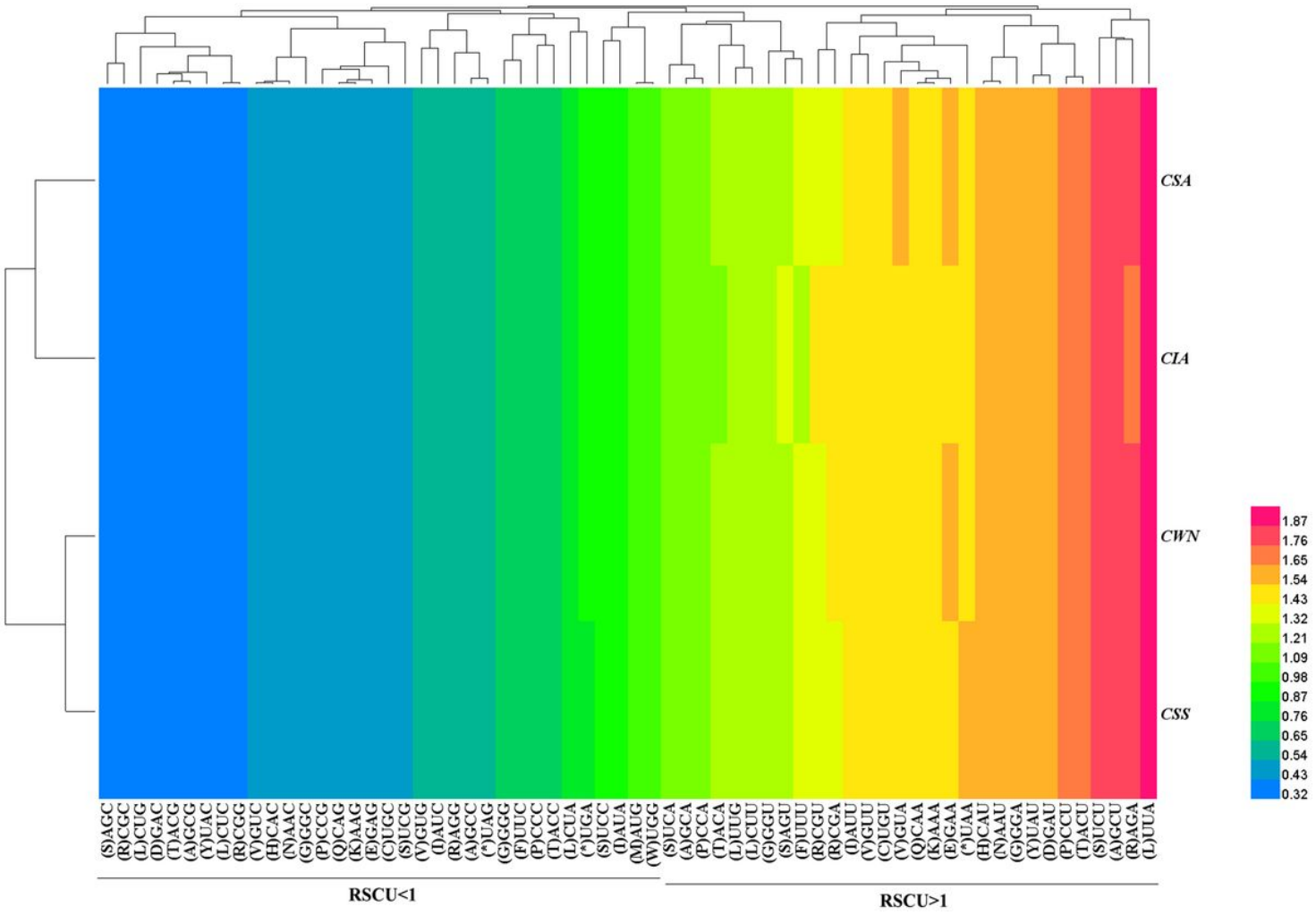

Figure 2

Neutrality plot of chloroplast genomes of four tea species. (A) C. sinensis var. sinensis (B) C. sinensis var. assamica (C) C. assamica (D) C. sinensis cv. wuyi Nacrissus

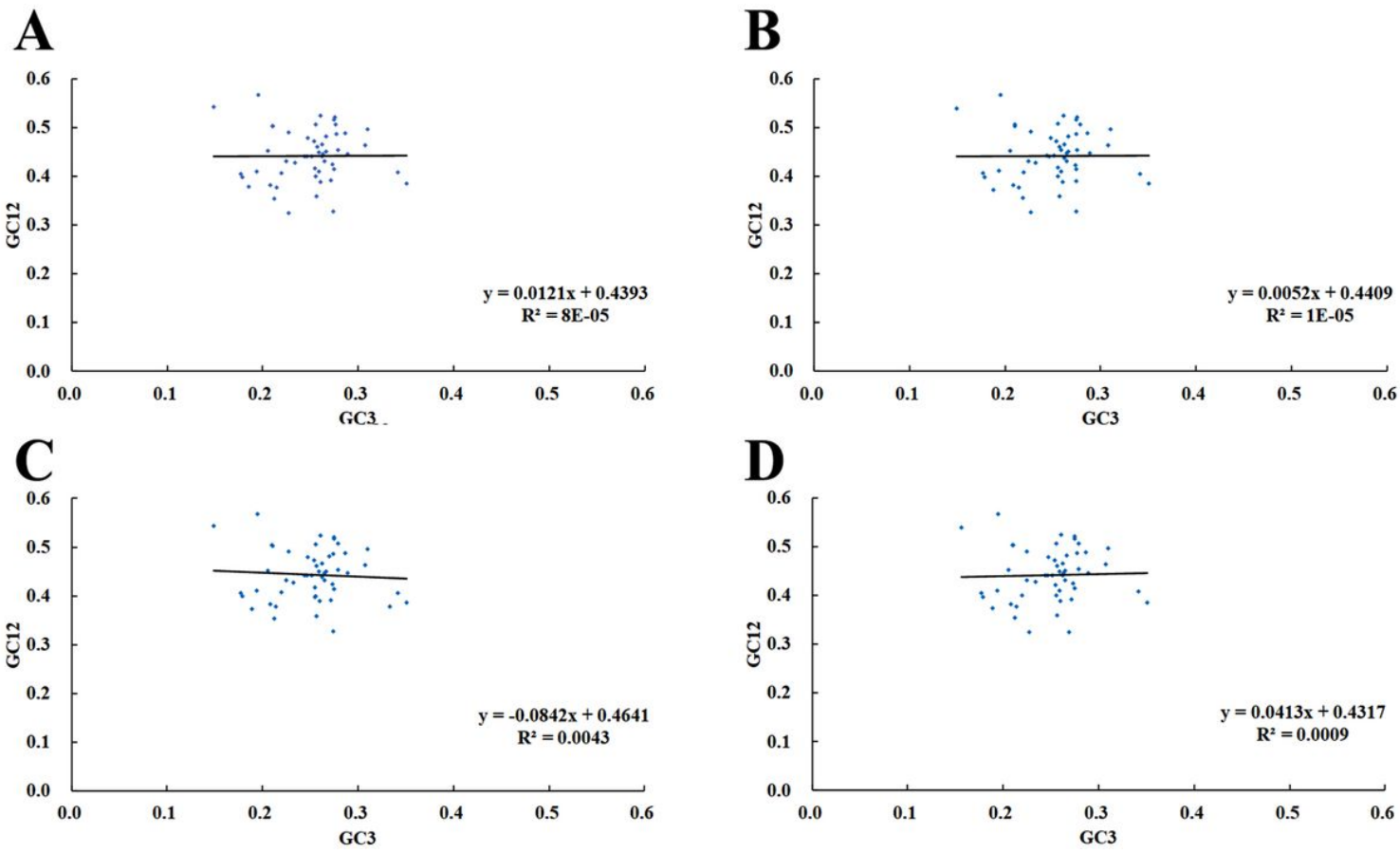

Figure 3

ENc-plot of chloroplast genomes of four tea species. (A) C. sinensis var. sinensis (B) C. sinensis var. assamica (C) C. assamica (D) C. sinensis cv. wuyi Nacrissus

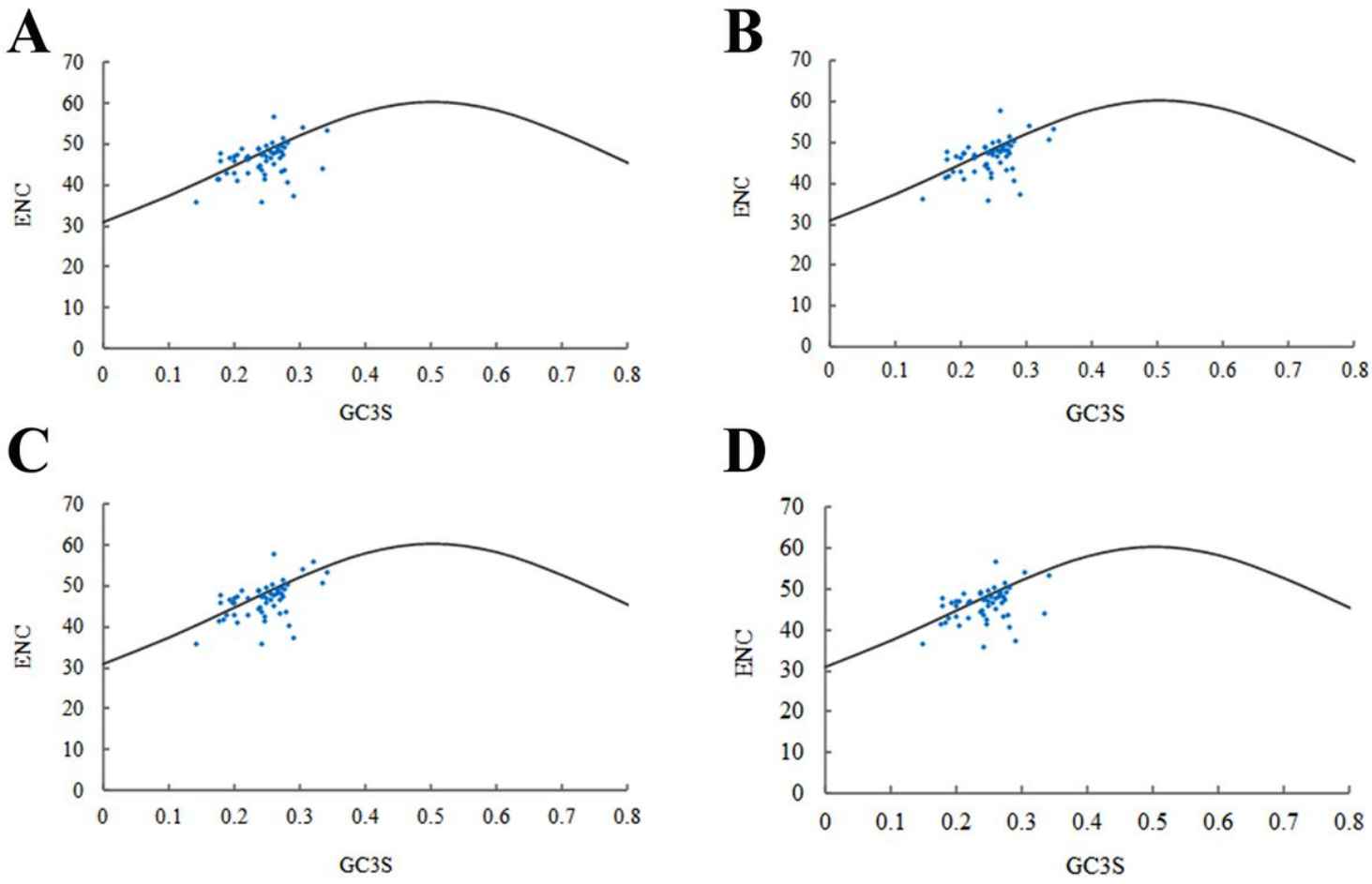

Figure 4

The distributions of codon usage in the form of heat maps for 4 tea species. Color indication: red represents the larger RSCU values and blue represents the smaller RSCU values. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

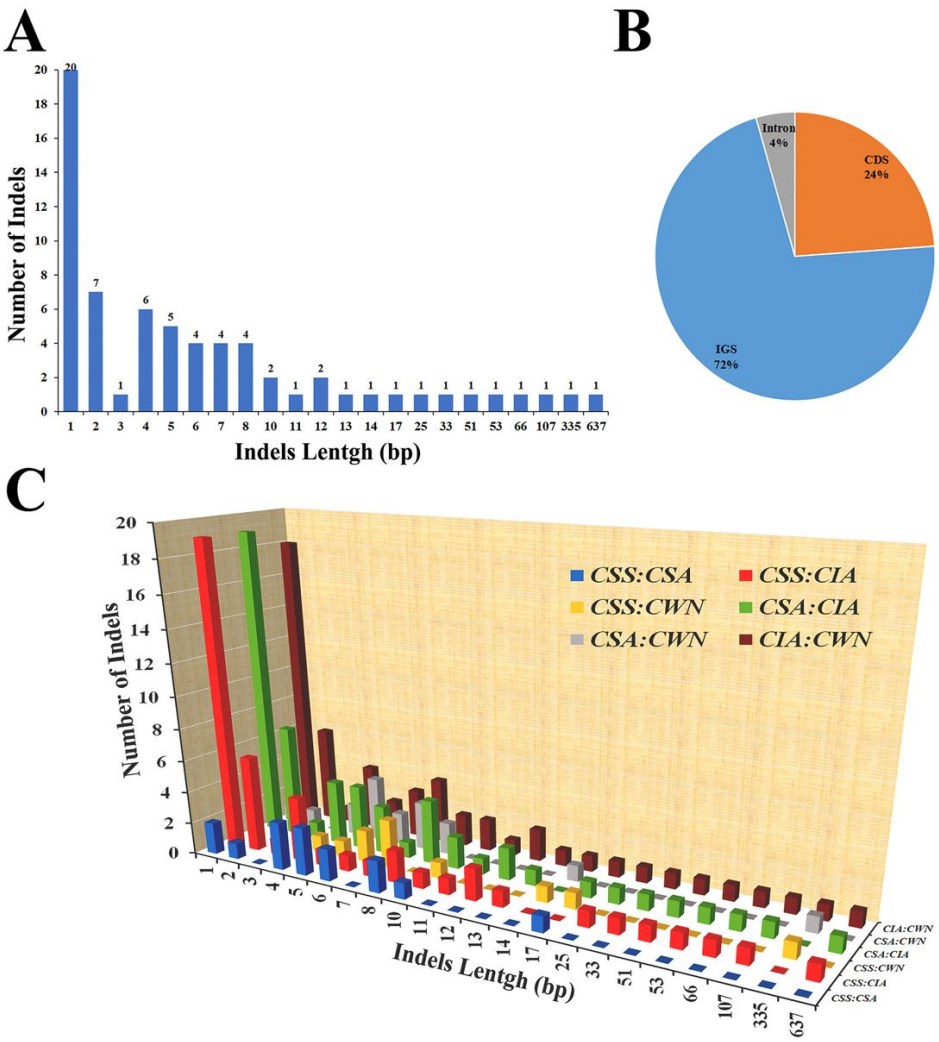

Figure 5

Analyses of the Indel sequences in the four chloroplast genomes. (A) Number of the Indel types by length. (B) Location of the all Indel from four species. (C) The pairwise comparisons among the four chloroplast genomes. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

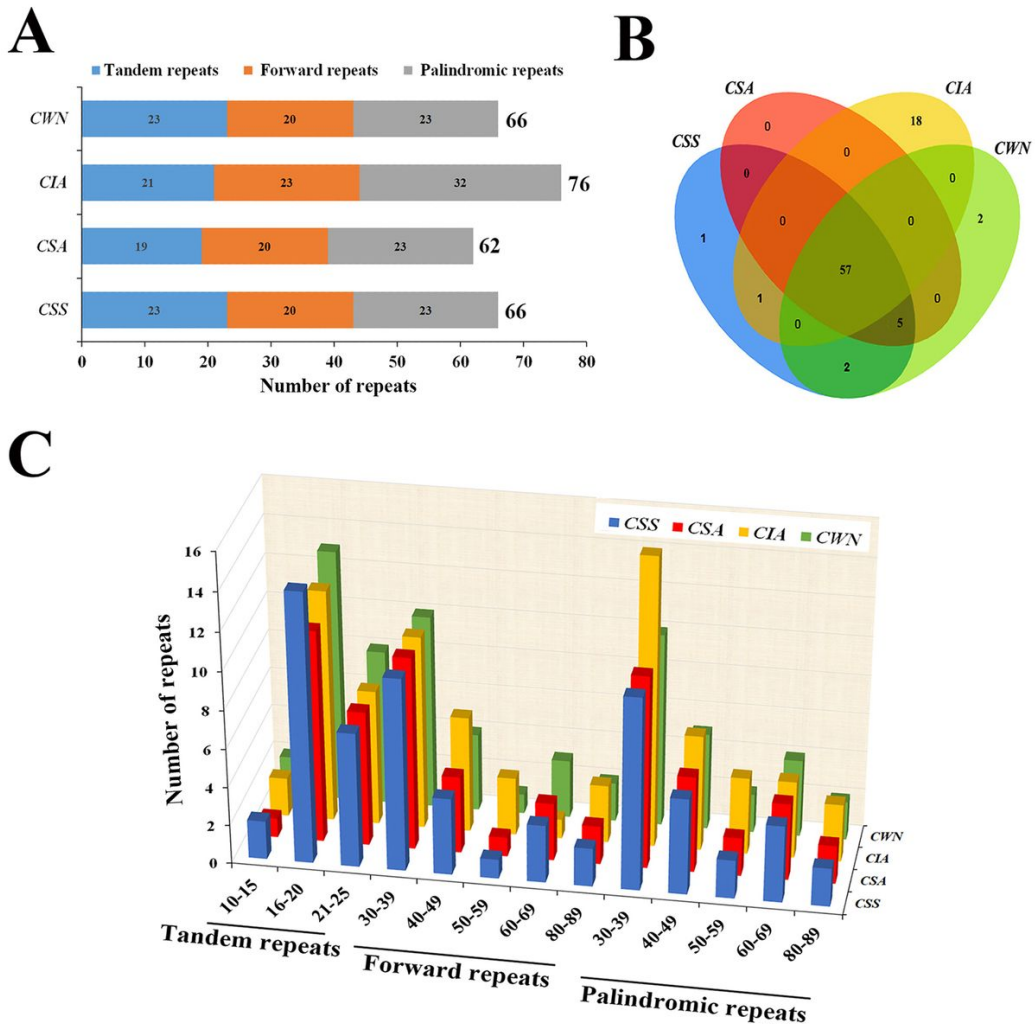

Figure 6

Analyses of repeated sequences in the four chloroplast genomes. Number of the three repeat types. (B) Number of repeated sequences in the four chloroplast genomes by Venn diagram. (C) Number of the repeats by different length. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

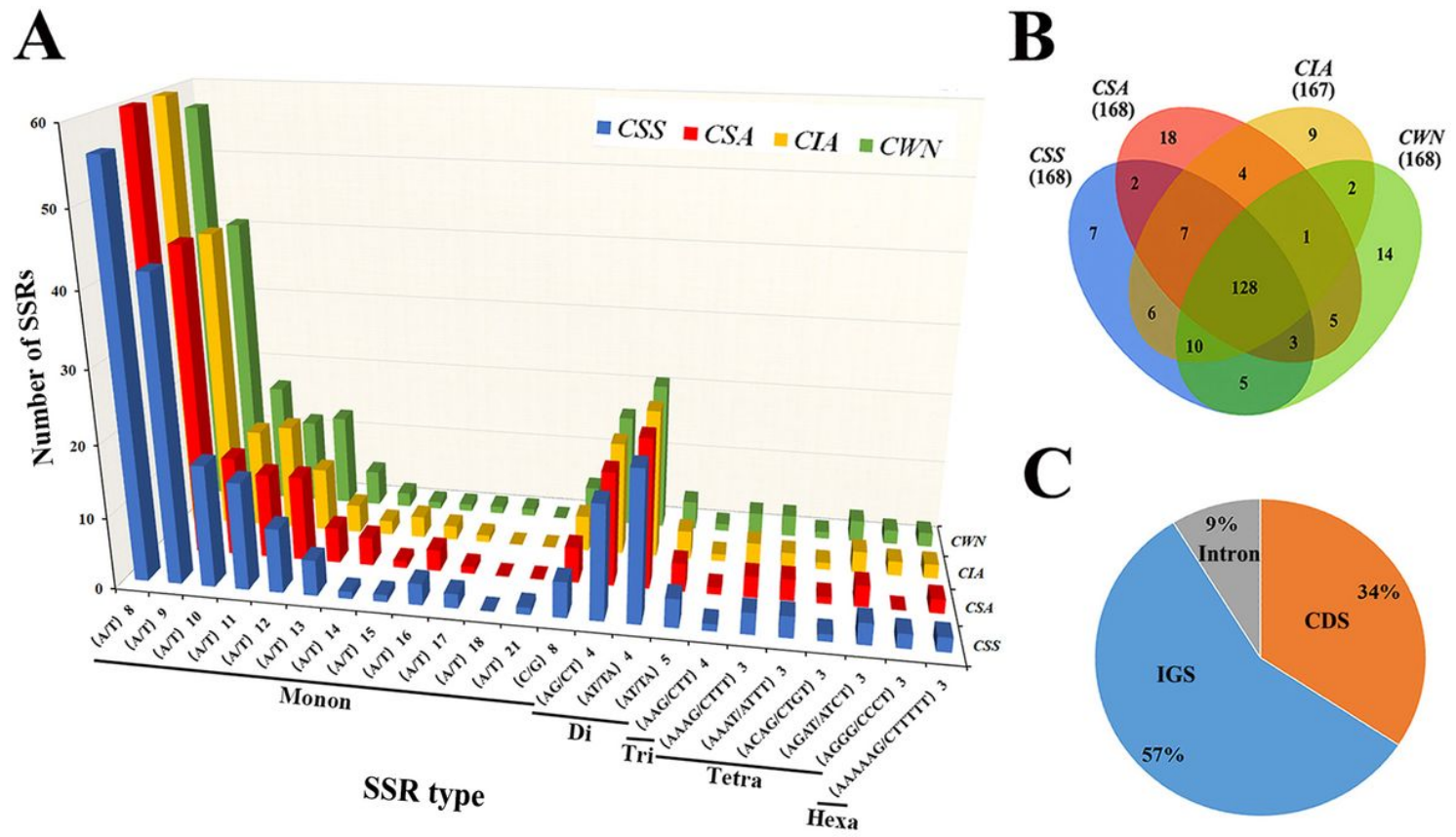

Figure 7

Analyses of simple sequence repeat (SSR) in the four chloroplast genomes. (A) Number different SSRs types detected by MISA. (B) Number of simple sequence repeats (SSRs) in the four chloroplast genomes by Venn diagram. (C) Location of the all SSRs from four species. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

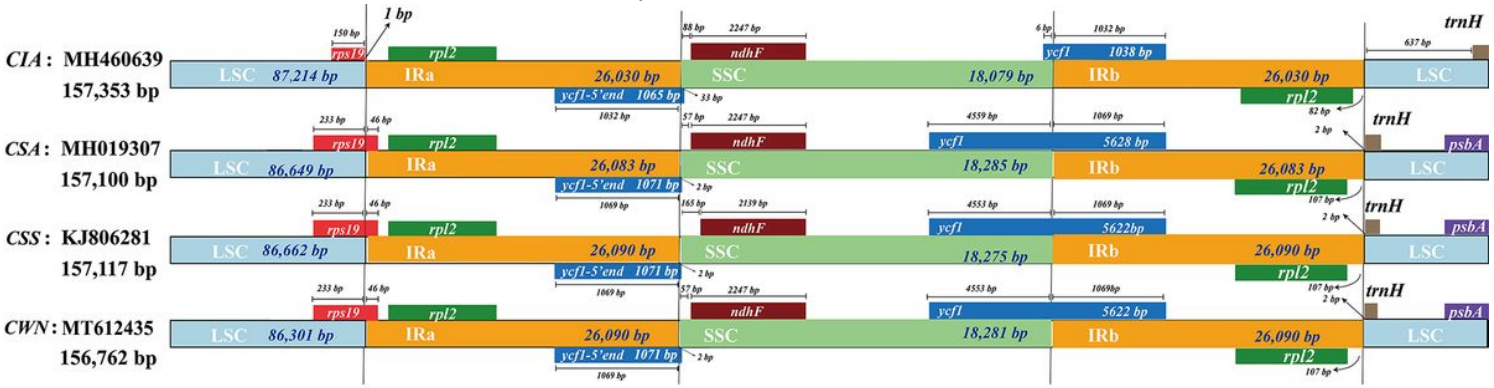

Figure 8

The comparison of the LSC, IR and SSC border regions among the four chloroplast genomes. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

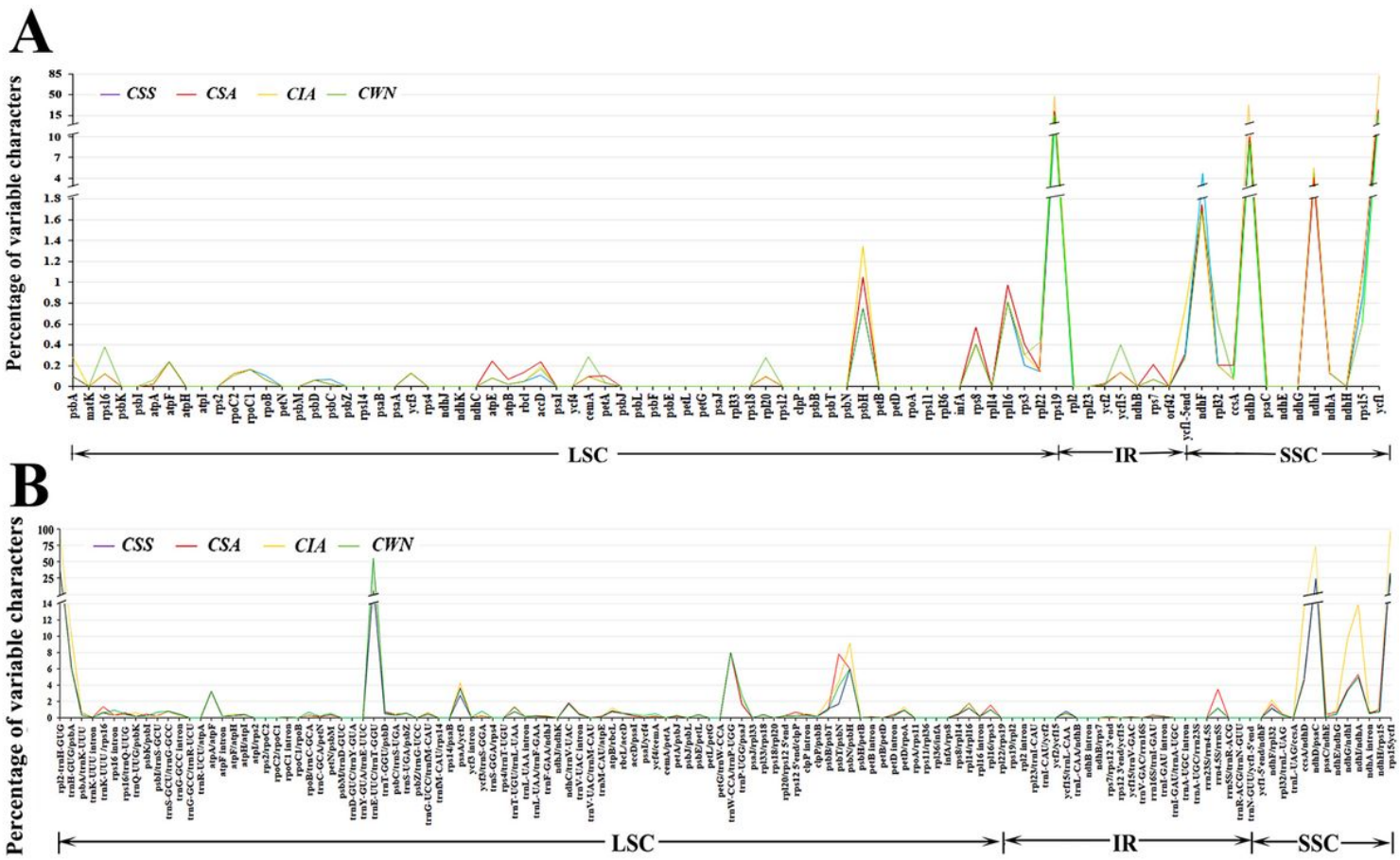

Figure 9

Percentages of variable characters in homologous regions across the four chloroplast genomes. (A) Coding regions. (B) Non-coding regions. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

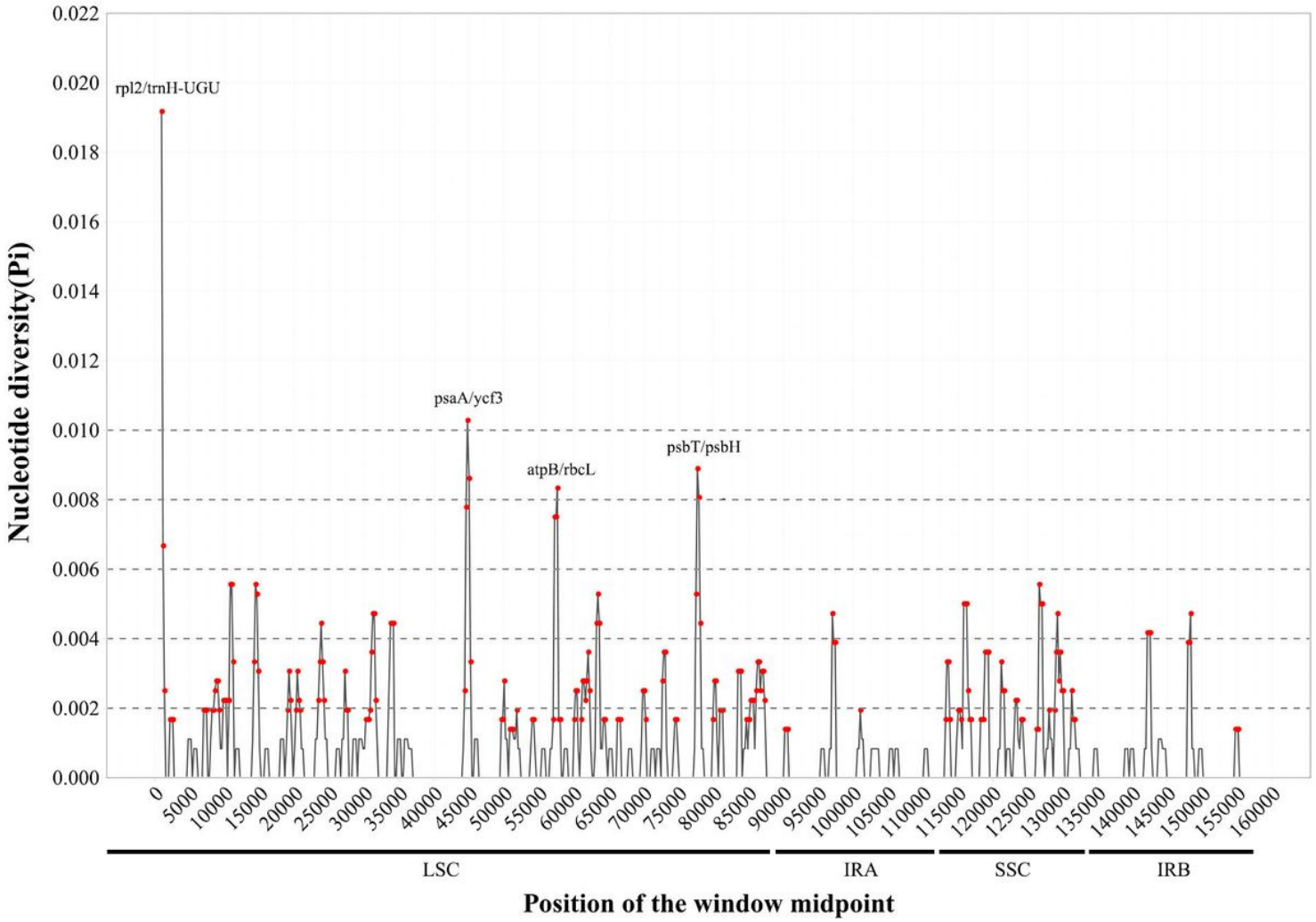

Figure 10

Sliding window analysis of the complete chloroplast genomes of four tea species (window length: 600 bp, step size: 200 bp). X-axis: position of the window midpoint, Y-axis: nucleotide diversity within each window.

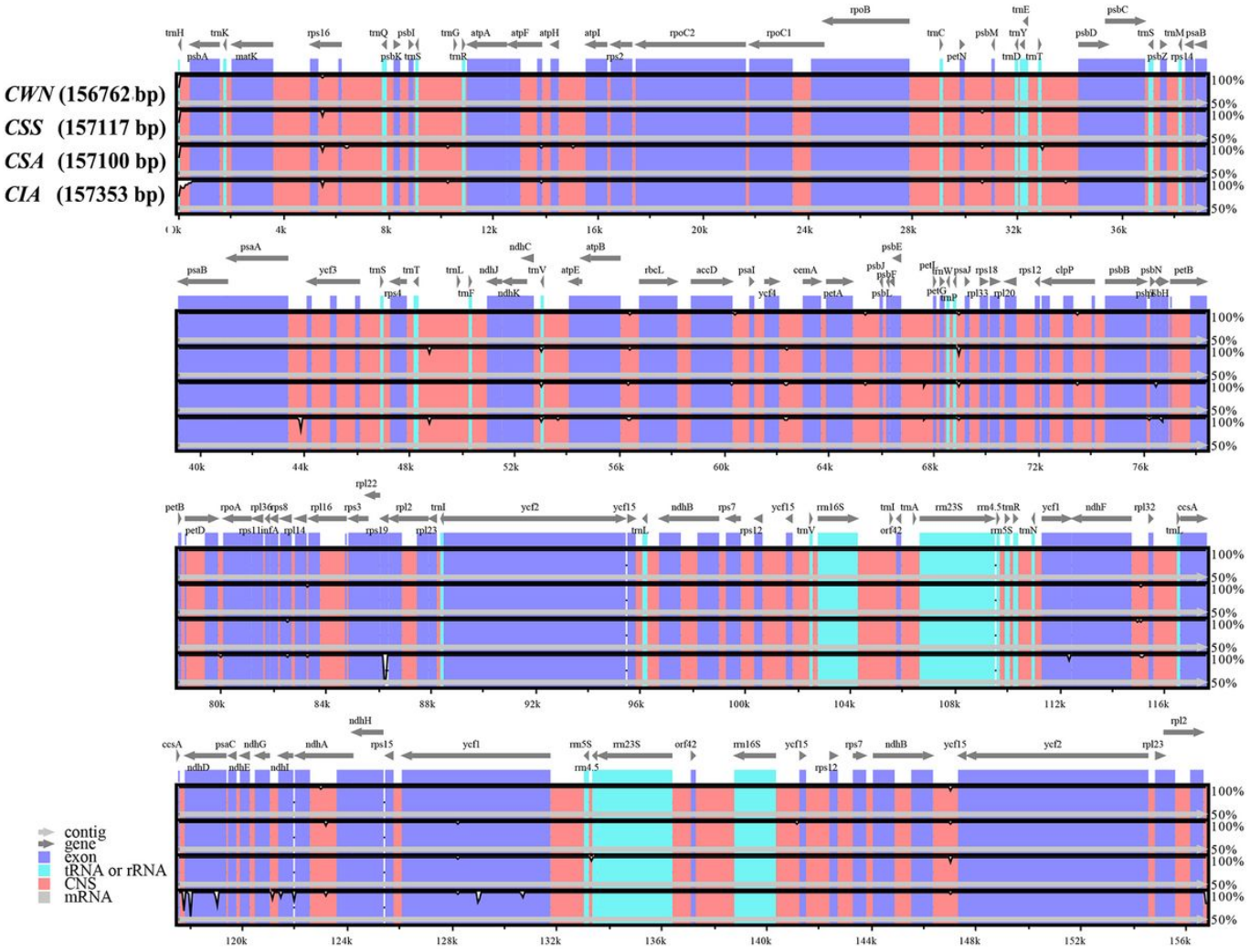

Figure 11

Visualization of alignment of the four tea species chloroplast genome sequences. VISTA-based identity plots showed sequence identity of four chloroplast genomes with C. sinensis cv. Wuyi Narcissus as a reference. Genome regions are color coded as protein coding, rRNA coding, tRNA coding or conserved noncoding sequences (CNS). The vertical scale indicates the percentage identity, ranging from 50% to 100%. CSS, C. sinensis var. sinensis; CSA, C. sinensis var. assamica; CIA, C. assamica; CWN, C. sinensis cv. wuyi Nacrissus

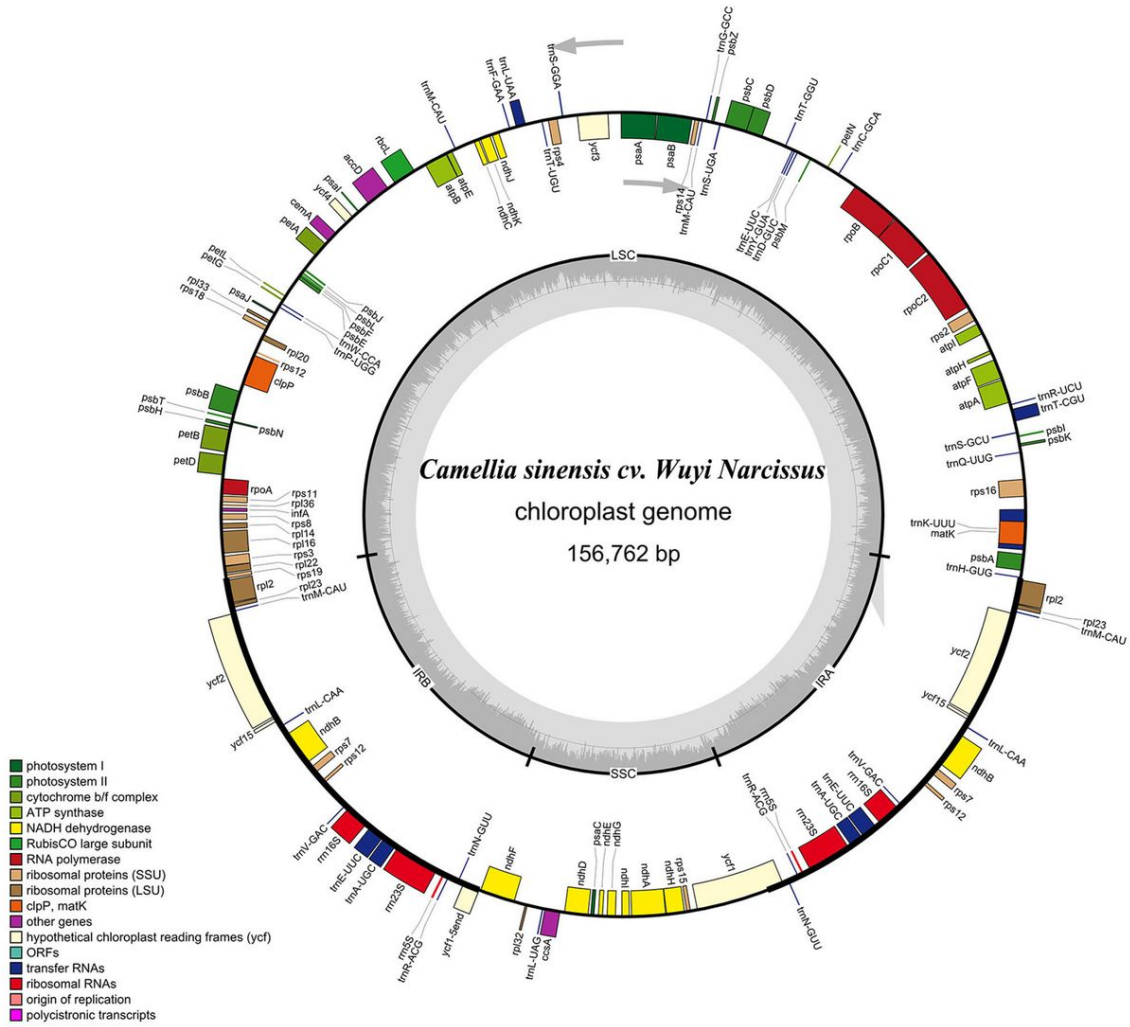

Figure 12

Chloroplast genome map of Camellia sinensis cv. Wuyi Narcissus. Genes shown outside the outer circle are transcribed clockwise and those inside are transcribed counterclockwise. Genes belonging to different functional groups are color coded. Dashed area in the inner circle indicates the GC content of the chloroplast genome. ORF: open reading frame.

Supplementary Files

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

- SupplementaryTable.1.doc
- SupplementaryTable.2.docx
- SupplementaryTable.3.docx
- SupplementaryTable.4.docx