www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Deciphering tea tree chloroplast and mitochondrial genomes of *Camellia sinensis* var. *assamica*

Received: 6 March 2019

Accepted: 10 July 2019

Published online: 17 October 2019

Fen Zhang[1], Wei Li[1], Cheng-wen Gao[2], Dan Zhang[1] & Li-zhi Gao[1,3]

Tea is the most popular non-alcoholic caffeine-containing and the oldest beverage in the world. In this study, we *de novo* assembled the chloroplast (cp) and mitochondrial (mt) genomes of *C. sinensis* var. *assamica* cv. *Yunkang10* into a circular contig of 157,100 bp and two complete circular scaffolds (701719 bp and 177329 bp), respectively. We correspondingly annotated a total of 141 cp genes and 71 mt genes. Comparative analysis suggests repeat-rich nature of the mt genome compared to the cp genome, for example, with the characterization of 37,878 bp and 149 bp of long repeat sequences and 665 and 214 SSRs, respectively. We also detected 478 RNA-editing sites in 42 protein-coding mt genes, which are ~4.4-fold more than 54 RNA-editing sites detected in 21 protein-coding cp genes. The high-quality cp and mt genomes of *C. sinensis* var. *assamica* presented in this study will become an important resource for a range of genetic, functional, evolutionary and comparative genomic studies in tea tree and other *Camellia* species of the Theaceae family.

Background & Summary

Tea is the most popular non-alcoholic caffeine-containing and the oldest beverage in the world since 3000 B. C.[1,2]. The production of tea made from the young leaves of *Camellia sinensis* var. *sinensis* and *C. sinensis* var. *assamica*, together with ornamentally well-known camellias (e.g., *C. japonica*, *C. reticulata* and *C. sasanqua*) and worldwide renowned wooden oil crop *C. oleifera*[3] has made the genus *Camellia* possess huge economic values in Theaceae. Besides its industrial, cultural and medicinal values, botanists and evolutionary biologists have increasingly paid attention to this genus. As a result of frequent hybridization and polyploidization, *Camellia* is almost commonly regarded as one of the most taxonomically and phylogenetically difficult taxa in flowering plants[4]. Thus, it has long been problematic for the taxonomic classification of the *Camellia* species based on the morphological characteristics[5]. The chloroplast (cp) genomes are able to provide valuable information for taxonomic classification, tracing source populations[6,7] and the reconstruction of phylogeny to resolve complex evolutionary relationships[8–10] due to the conservation of genomic structure, maternal inheritance and a fairly low recombination rate. Genetically speaking, cp genomes are comparatively conserved than plant mitochondria (mt) genomes which are more heterogeneous in nature. However, the presence of NUPT (nuclear plastid DNA) into cp genomes argues that cp genomes assembled from WGS data may include the heterogeneity due to the nuclear cp DNA transferred to the nucleus, resulting in erroneous phylogenetic inferences[11]. It has long been acknowledged that mtDNA has the propensity to integrate DNA from various sources through intracellular and horizontal transfer[12–14]. Partially due to these reasons, the mt genomes vary from ~200 Kbp to ~11.3 Mbp in some living organisms[15–17]. The dynamic nature of mt genome structure has been recognized, and plant mt genomes can have a variety of different genomic configurations due to the recombination and differences in repeat content[18,19]. These characteristics make the plant mt genome a fascinating genetic system to investigate questions related to evolutionary biology. The first effort has been made to sequence the 13 representative *Camellia* chloroplast genomes using next-generation Illumina genome sequencing platform, which obtained novel insights into global patterns of structural variation across the *Camellia* cp genomes[4]. The reconstruction of phylogenetic relationships among these representative species of *Camellia* suggests that cp genomic resources are able to provide useful data to help to understand their

[1]Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou, 510642, China. [2]Affiliated Hospital, Qingdao University, Qingdao, 266003, China. [3]Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwestern China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650204, China. These authors contributed equally: Fen Zhang, Wei Li and Cheng-wen Gao. Correspondence and requests for materials should be addressed to L.-z.G. (email: Lgaogenomics@163.com)

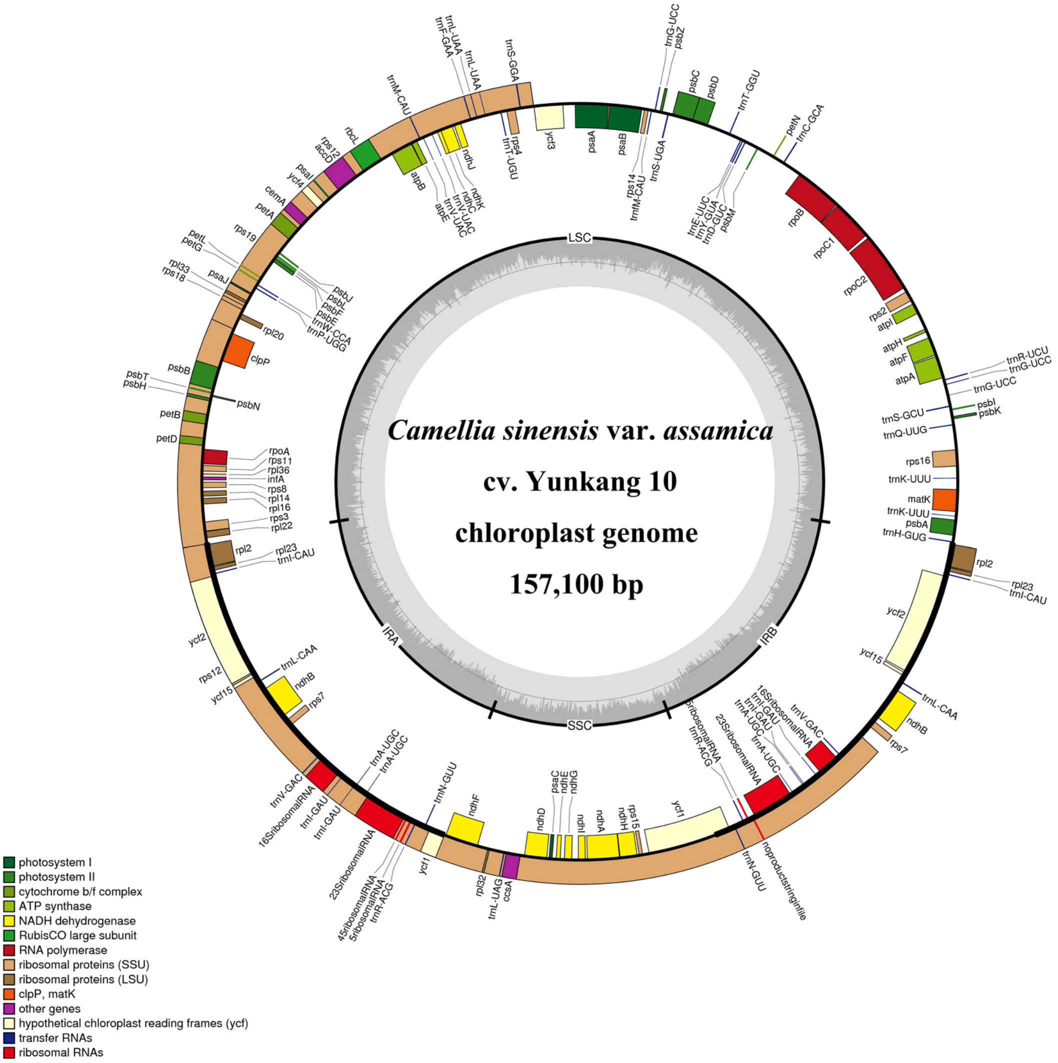

Fig. 1 Genome map of *C. sinensis* var. *assamica* cv. *Yunkang10*. Genes lying outside of the outer circle are transcribed in the clockwise direction whereas genes inside are transcribed in the counterclockwise direction. Genes belonging to different functional groups are color-coded. Area dashed darker gray in the inner circle indicates GC content while the lighter gray corresponds to AT content of the genome.

evolutionary relationships and classify the 'difficult taxa'. Increasing interest in the *Camellia* plants have made up to thirty-eight of cp genomes be sequenced up to date[20–37]. Recently, we decoded the first nuclear genome of *C. sinensis* var. *assamica* cv. *Yunkang10*, providing novel insights into genomic basis of tea flavors[38]. Besides the lack of the *C. sinensis* var. *assamica* cp genome among thirty-eight cp genomes that were sequenced in this genus[4,20–37], up to data, none of mt genome has been determined in the genus *Camellia*.

In this study, we filtered cpDNA and mtDNA reads from the WGS genome sequence project[38] and *de novo* assembled the mt genome and cp genome of *C. sinensis* var. *assamica*. The information of both cp and mt genomes will help to obtain a comprehensive understanding of the taxonomy and evolution of the genus *Camellia*. These genome sequences will also facilitate the genetic modification of these economically important plants, for example, through chloroplast genetic engineering technologies.

Methods

Plant materials, DNA extraction and genome sequencing. Young and healthy leaves of an individual plant of cultivar *Yunkang10* of *C. sinensis* var. *assamica* were collected for genome sequencing in April, 2009, from Menghai County, Yunnan 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 a modified CTAB method[39]. RNase A and proteinase K were separately used to remove RNA and protein contamination. The quality and quantity of the isolated DNA were separately checked by electrophoresis on a 0.8% agarose gel and a NanoDrop D-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). A total of eleven paired-end libraries, including four types of small-insert libraries (180 bp, 260 bp, 300 bp, 500 bp) and seven large-insert libraries (2 Kb, 3 Kb, 4 Kb, 5 Kb,

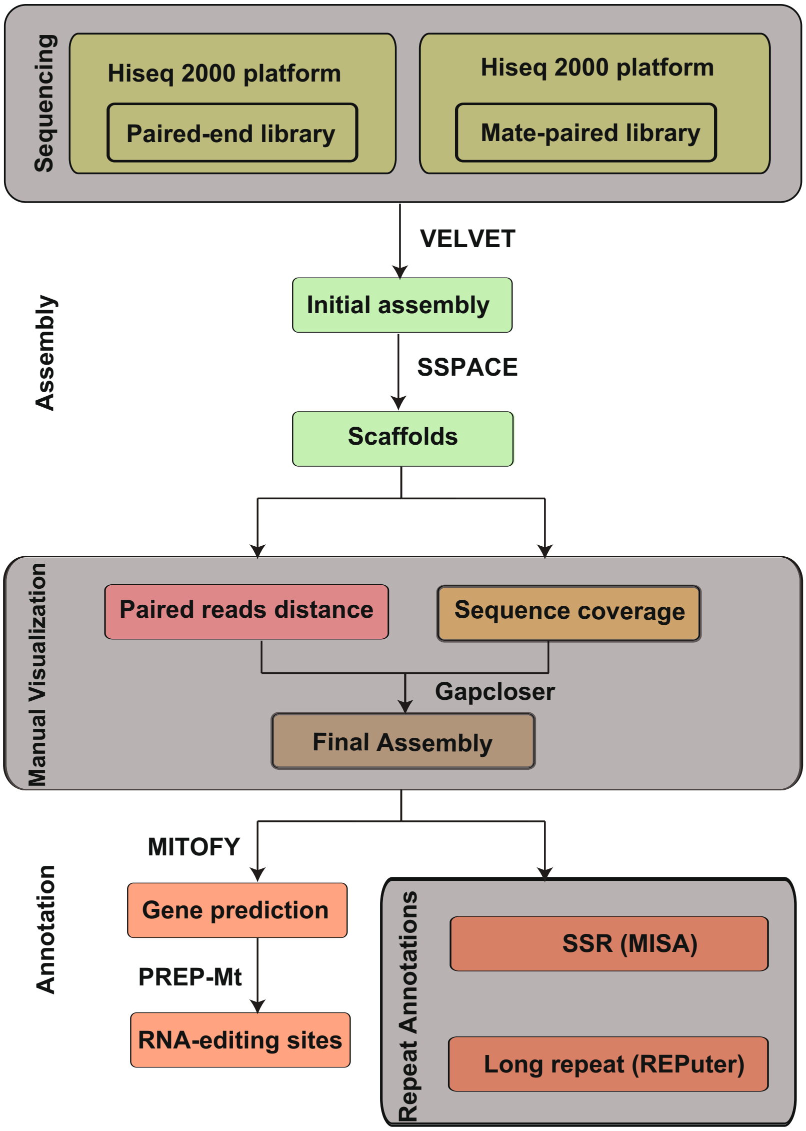

Fig. 2 The assembly and annotation pipeline of the tea tree mitochondrial genome.

6 Kb, 8 Kb, 20 Kb), were prepared following the Illumina's instructions, and sequenced using Illumina HiSeq. 2000 platform by following the standard Illumina protocols (Illumina, San Diego, CA). We totally generated ~707.88 Gb (~229.31×) of raw sequencing data[38]. Further reads quality control filtering processes yielded a total of ~492.15 Gb (~159.43×) high-quality data retained and used for subsequent genome assembly.

De novo chloroplast and mitochondria genome assemblies. The chloroplast reads were filtered from whole genome Illumina sequencing data of *C. sinensis* var. *assamica*, we mapped all the sequencing reads to the reference genomes[4] using bowtie2 (version 2.3.4.3)[40]. The mapped chloroplast reads were assembled into a circular contig of 157,100 bp in length with an overall GC content of 37.29% using CLC Genomics Workbench v. 3.6.1 (CLC Inc., Rarhus, Denmark) (Fig. 1). For mitochondria genome assembly, the PE and MP sequencing reads were used separately. Briefly, we first performed *de novo* assembly with VELVET v1.2.08[41], which was previously described[42,43]. Scaffolds were constructed using SSPACE v.3.0[44]. False connection was manually removed based on the coverage and distances of paired reads. Gaps between scaffolds were then filled with GapCloser (version 1.12)[45,46] using all pair-end reads. We obtained the two complete circular scaffolds (701719 bp and 177329 bp) of the *C. sinensis* var. *assamica* mt genome from the *de-novo* assembly of the filtered mitochondrial reads (Figs 2–4). The two scaffolds of the mt genome had overall GC contents of 45.63% and 45.81%, respectively. The completed chloroplast and mitochondria genomes are publicly available in NCBI GenBank under accession numbers MH019307, MK574876 and MK574877 and BIG Genome Warehouse WGS000271, WGS000272.

Genome annotation and visualization. The complete chloroplast genome of *C. sinensis* var. *assamica* was preliminarily annotated using the online program DOGMA[47] (Dual Organellar Genome Annotator) followed by manual correction. A total of 141 genes were annotated, of which 87 were protein-coding genes, 46

Category	Group	Genes
Photosynthesis related genes	Rubisco	rbcL
	Photosystem I	psaA, psaB, psaC, psaI, psaJ
	Assembly/stability of Photosystem I	ycf3
	Photosystem II	psbA, psbB, psbT, psbK, psbI, psbH, psbM, psbN, psbD, psbC, psbZ, psbJ, psbL, psbE, psbF
	ATP synthase	atpA, atpB, atpE, atpF, atpH, atpI
	Cytochrome b/f complex	petA, petB, petD, petN, petL, petG
	Cytochrome csynthesis	ccsA
	NADPH dehydrogenase	ndhA, ndhB (×2), ndhC, ndhD, ndhE, ndhF, ndhH, ndhG, ndhJ, ndhK, ndhI
Transcription and translation related genes	Transcription	rpoA, rpoC2, rpoC1, rpoB
	Ribosomal proteins	rps2, rps3, rps4, rps7 (×2), rps8, rps11, rps12, rps14, rps15, rps16, rps18, rps19, rpl2 (×2), rpl14, rpl16, rpl20, rpl22, rpl23 (×2), rpl32, rpl33, rpl36
	Translation initiation factor	infA
RNA genes	Ribosomal RNA	rrn16S (×2), rrn23S (×2), rrn4.5 (×2), rrn5 (×2)
	Transfer RNA	trnH-GUG, trnK-UUU (×2), trnQ-UUG, trnS-GCU, trnG-UCC (×2), trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-UCC, trnfM-CAU, trnS-GGA, trnT-UGU, trnL-UAA (×2), trnF-GAA, trnV-UAC (×2), trnM-CAU, trnW-CCA, trnP-UGG, trnI-CAU, trnL-CAA (×2), trnV-GAC, trnI-GAU (×3), trnA-UGC (×2), trnR-ACG (×2), trnN-GUU (×2), trnL-UAG, trnN-GUU, trnR-ACG, trnA-UGC (×2), trnV-GAC, trnI-CAU
Other genes	RNA processing	matK
	Carbon metabolism	cemA
	Fatty acid synthesis	accD
	Proteolysis	clpP
Genes of unknown function	Conserved ORFs	ycf1 (×2), cf2, ycf4, ycf2, ycf15 (×2)

Table 1. Gene annotation of the *C. sinensis* var. *assamica* cp genome.

	Name of genes	
Group of genes	**Scaffold 1**	**Scaffold 2**
Complex I	nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9 (×2)	nad1, nad2
Complex II	sdh3, sdh4	sdh3
Complex III		cob
Complex IV	cox1, cox2, cox3	
Complex V	atp1, atp4, atp6, atp8, atp9	atp9
Cytochrome c biogenesis	ccmFn, ccmB, ccmC	ccmFc
Ribosome large subunit	rpl2, rpl10, rpl16	rpl5
Ribosome small subunit	rps1, rps3, rps4, rps7, rps12, rps13, rps19	rps14, rps19
rRNA genes	rrn5, rrn18, rrn16	
tRNA genes	trnS(Ser), trnD(Asp), trnK(Lys), trnfM(Met) (×2), trnI(Ile)-cp, trnE(Glu), trnH(His)-cp, trnP(Pro), trnW(Trp)-cp, trnG(Gly), trnQ(Gln), trnC(Cys), trnD(Asp), trnS(Ser), trnV(Val)-cp	trnI(Ile), trnM(Met)-cp, trnC(Cys), trnN(Asn)-cp, trnY(Tyr), trnS(Ser), trnF(Phe), trnP(Pro)
chloroplast-derived genes	trnI(Ile)-cp, trnH(His)-cp, trnW(Trp)-cp, trnV(Val)-cp	trnM(Met)-cp, trnN(Asn)-cp
Other proteins	matR, mttB	

Table 2. Gene content of the *C. sinensis* var. *assamica* mt genome.

were tRNA genes and eight were rRNA genes (Table 1). MITOFY[15] was used to characterize the complement of protein-coding and rRNA genes in the mitochondrial genome. A tRNA gene search was carried out using the tRNA scan-SE software (version 1.3.1)[48]. We annotated a total of 71 genes, including 44 protein-coding genes, 24 tRNAs and 3 rRNAs (Table 2). Circular genome maps were drawn with OrganellarGenomeDRAW[49] (Figs 3–4).

Simple sequence repeats (SSRs) were identified and located using MISA (http://pgrc.ipk-gatersleben.de/misa/). All the annotated SSRs were classified by the size and copy number of their tandemly repeated: monomer (one nucleotide, $n \geq 8$), dimer (two nucleotides, $n \geq 4$), trimer (three nucleotides, $n \geq 4$), tetramer (four nucleotides, $n \geq 3$), pentamer (five nucleotides, $n \geq 3$), hexamer (six nucleotides, $n \geq 3$). A total of 214 SSRs were identified in cp genome with 74.42% of which were monomers, 19.07% of dimers, 0.47% of trimers, 4.65% of tetramers and 0.93% of hexamers (Table 3). There were no pentamers found in the cp genome. In mt genome, we obtained 665 SSRs distributed into monomers, dimers, trimers, pentamers, tetramers and hexamers with 31.53%, 45.35%, 4.95%, 15.17%, 2.70% and 0.15%, respectively (Table 3). Repeat sequences including forward and palindromic

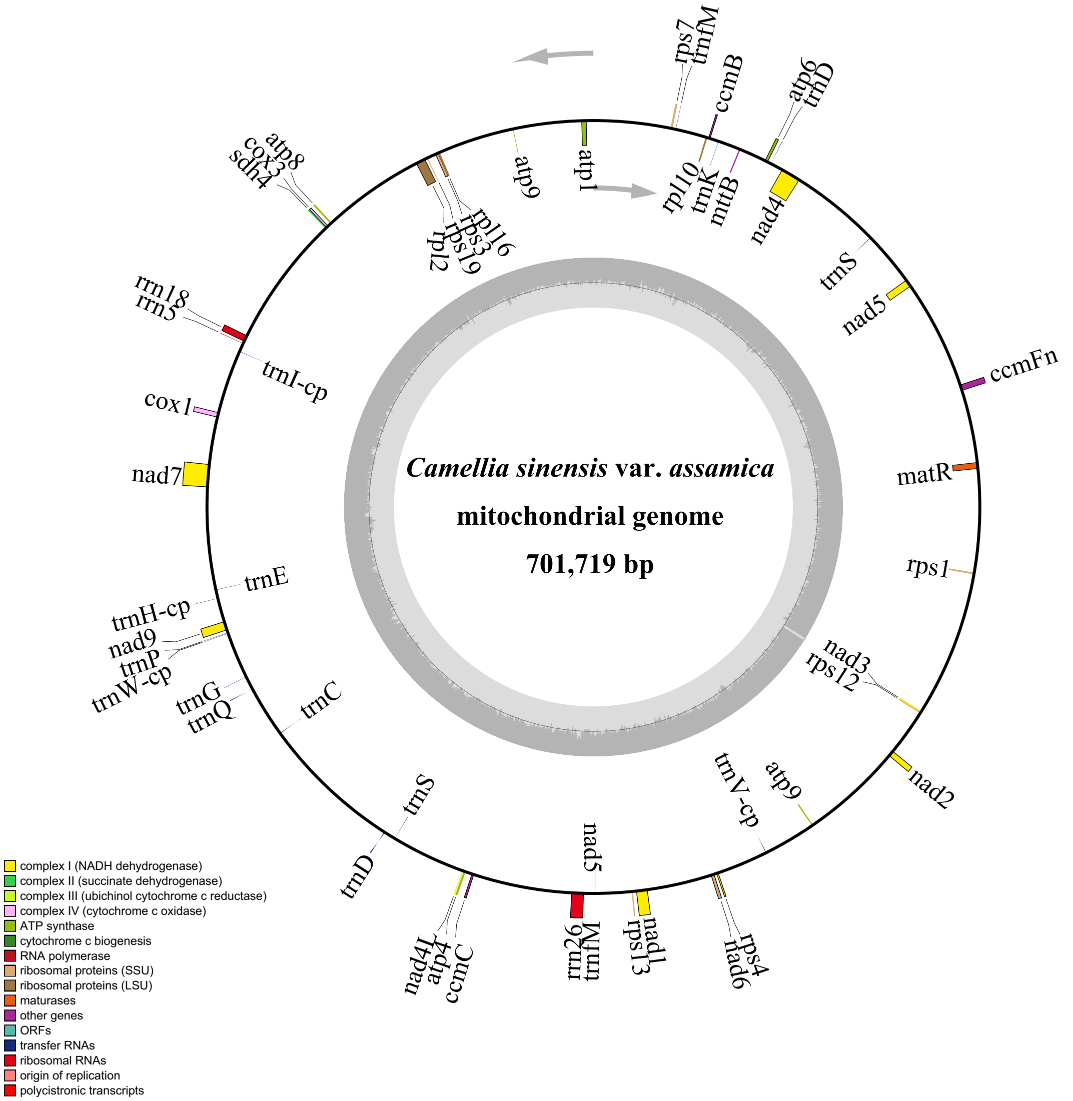

Fig. 3 Circular map of scaffold 1 in the *C. sinensis* var. *assamica* cv. *Yunkang10* mitochondrial genome. Gene map showing 54 annotated genes with different functional groups that are color-coded on outer circle as transcribed clock-wise (outside) and transcribed counter clock-wise (inside). The inner circle indicates the GC content as dark grey plot.

repeats, were also searched by REPuter[50] with the following parameters: minimal length 50 nt; mismatch 3 nt. Long repeat sequences (repeat unit > 50 bp) of forward and palindromic repeats were further annotated, resulting in 149 bp from 4 paired repeats in the cp genome (Table 4) and 37,878 bp from 58 paired repeats in the mt genome (Online-only Tables 1–2). Our repeat content analyses indicate that the mt genome is more abundant in repeat sequences and more variable than the cp genome of *C. sinensis* var. *assamica* (Table 4**;** Online-only Tables 1–2).

Prediction of RNA-editing sites. Putative RNA editing sites in protein-coding genes were predicted using the PREP-cp and PREP-mt Web-based program (http://prep.unl.edu/)[51,52]. To achieve a balanced trade-off between the number of false positive and false negative sites, the cutoff score (C-value) was set to 0.8 and 0.6, respectively[53].

Almost all transcripts of protein encoding genes in the plant mitochondria are subject to RNA editing except the *T-urf13* gene[54]. Our results showed that the extent of RNA editing varied by gene for both cp and mt genomes of *C. sinensis* var. *assamica*. In the *C. sinensis* var. *assamica* cp genome, we detected 54 RNA-editing sites in 21 protein-coding genes, ranging from one editing site in *atpF*, *atpI*, *petB*, *psaI*, *psbE*, *psbF*, *rpoA*, *rps2* and *rps8* to 8 editing sites in *ndhB* (Online-only Table 3). In the *C. sinensis* var. *assamica* mt genome, we predicted 478 RNA-editing sites in 42 protein-coding genes; they varied from two editing site in *atp9* (of scaffold2), *sdh3* (of scaffold1 and scaffold2, respectively) and *rps14* (of scaffold2) to 35 editing sites in *ccmFn* (of scaffold1) (Online-only Table 4–5).

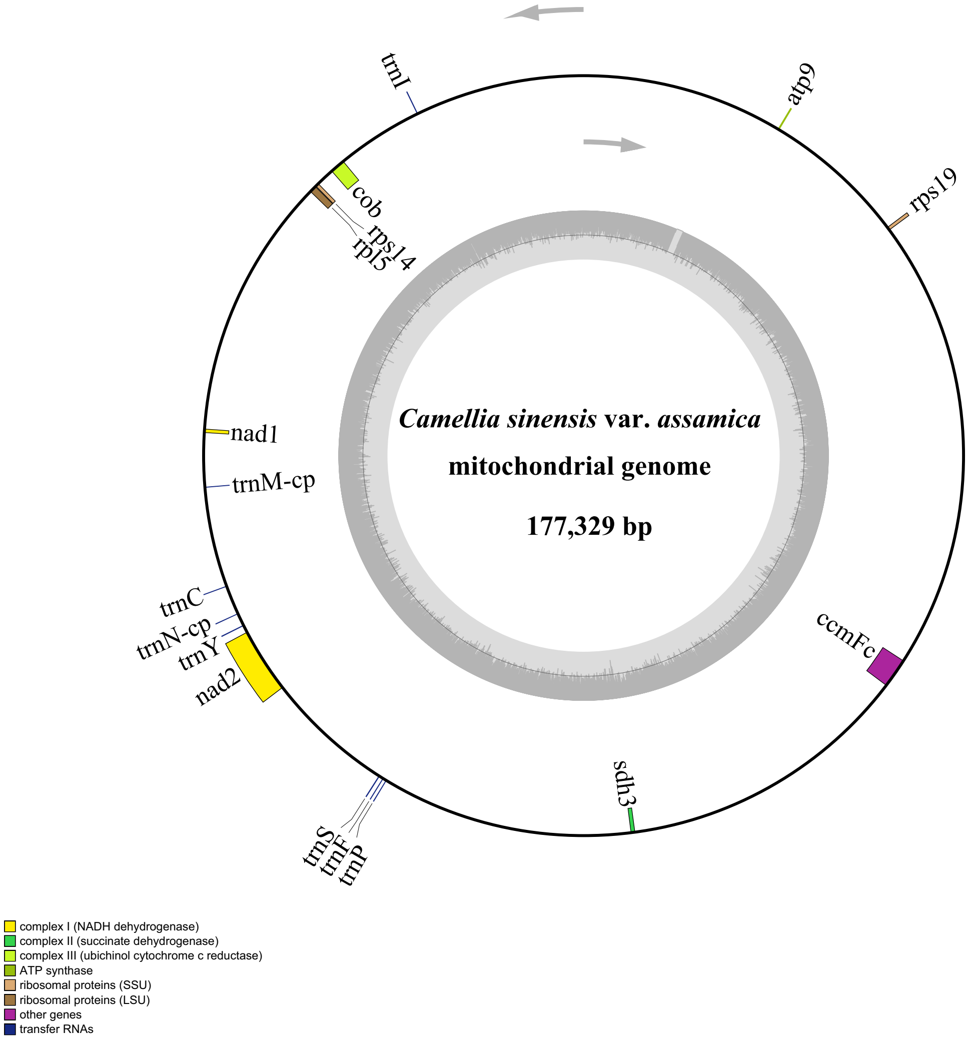

Fig. 4 Circular map of scaffold 2 in the *C. sinensis* var. *assamica* cv. *Yunkang10* mitochondrial genome. Gene map showing 17 annotated genes with different functional groups that are color-coded on outer circle as transcribed clock-wise (outside) and transcribed counter clock-wise (inside). The inner circle indicates the GC content as dark grey plot.

	mt Genome		cp Genome	
SSR-Motif	SSR Number	SSR %	SSR Number	SSR %
Monomer	210	31.53	160	74.42
Dimer	302	45.35	41	19.07
Trimer	33	4.95	1	0.47
Tetramer	101	15.17	10	4.65
Pentamer	18	2.70	0	0.00
Hexamer	1	0.15	2	0.93

Table 3. Statistics of SSR motifs in the *C. sinensis* var. *assamica* mt and cp genomes.

Phylogenetic analyses. To further determine the phylogenetic position of *C. sinensis* var. *assamica* we performed phylogenomic analysis of 20 complete cp genomes using the GTR + R + I model under the maximum likelihood (ML) inference in MEGA v.7.0[55]. Besides *C. sinensis* var. *assamica* cv. *Yunkang 10*, we selected cp genomes from the eighteen *Camelia* species (*C. oleifera*, *C. crapnelliana*, *C. szechuanensis*, *C. mairei*, *C. elongata*, *C. grandibracteata*, *C. leptophylla*, *C. petelotii*, *C. pubicosta*, *C. reticulata*, *C. azalea*, *C. japonica*, *C. cuspidata*, *C. danzaiensis*, *C. impressinervis*, *C. pitardii*, *C. yunnanensis* and *C. taliensis*) using *Apterosperm oblata* as outgroup. Our results showed that *C. sinensis* var. *assamica* was grouped with *C. grandibracteata* with 100% bootstrap support (Fig. 5).

Repeat Length	Type*	Start of Copy 1	Start of Copy 2
56	F	93938	93956
56	P	93938	149737
56	P	93956	149755
56	F	149737	149755

Table 4. Long repeats (repeat unit > 50 bp) in the *C. sinensis* var. *assamica* cp genome. *P indicates palindromic repeats; F indicates forward repeats. Overlapped repeats have been manually removed while calculating total length.

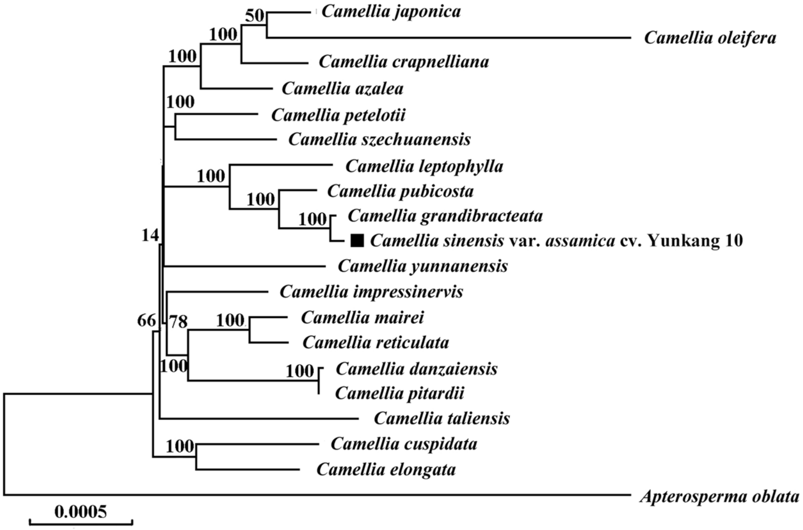

Fig. 5 Phylogenetic relationships of 20 complete chloroplast genomes. Maximum likelihood phylogenetic tree of *C. sinensis* var. *assamica* cv. *Yunkang 10* with 18 species in the genus *Camellia* based on complete chloroplast genome sequences. The chloroplast sequence of *Apterosperma oblata* was set as outgroup. The position of *C. sinensis* var. *assamica* cv. *Yunkang 10* is shown in bold and bootstrap values are shown for each node.

The same method was used for phylogenetic analysis with mt genome. A total of thirteen conserved mt protein-coding genes among *C. sinensis* var. *assamica* and 14 other plant species were individually aligned with ClustalW[56], and then concatenated to construct a contiguous sequence in the order of *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7* and *nad9*. The selected 14 species includes *Cycas taitungensis*, *Ginkgo biloba*, *Triticum aestivum*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Gossypium arboretum*, *G. barbadense*, *Carica papaya*, *Vitis vinifera*, *Hevea brasiliensis*, *Bupleurum falcatum*, *Glycine max* and *Salvia miltiorrhiza*. The alignment file was used for the construction of Neighbor-Joining Tree at 1000 bootstrap replicates with MEGA 7.0.26[55]. Our results showed that *C. sinensis* var. *assamica* is clearly grouped with other dicots that were separated from monocots of the angiosperms while the two gymnosperms (*Cycas taitungensis* and *Ginkgo biloba*) were formed the basal clade (Fig. 6).

Data Records

Raw reads from Illumina are deposited in the NCBI Sequence Read Archive (SRA)[57–62] and BIG Genome Warehouse[63]. Assembled cp genome sequences and accompanying gene annotations of *C. sinensis* var. *assamica* are deposited in the NCBI GenBank[64] and BIG Genome Warehouse[65]. The mt genome final assembly and accompanying gene annotations are deposited at NCBI GenBank[66,67] and BIG Genome Warehouse[68]. The alignment and tree files of the chloroplast genome and mitochondrial genome form the Camellia genus were deposited in Figshare database[69].

Technical Validation

Quality filtering of raw reads. The initially generated raw sequencing reads were evaluated in terms of the average quality score at each position, GC content distribution, quality distribution, base composition, and other metrics. Furthermore, the sequencing reads with low quality were also filtered out before the genome assembly and annotation of gene structure.

Assembly and validation. The chloroplast reads were filtered from whole genome Illumina sequencing data of *C. sinensis* var. *assamica*. We mapped all the cleaned reads to the reference chloroplast sequence[4] using bowtie2 (version 2.3.4.3)[40] with default parameters. The mapped chloroplast reads were *de novo* assembled into the complete chloroplast genome.

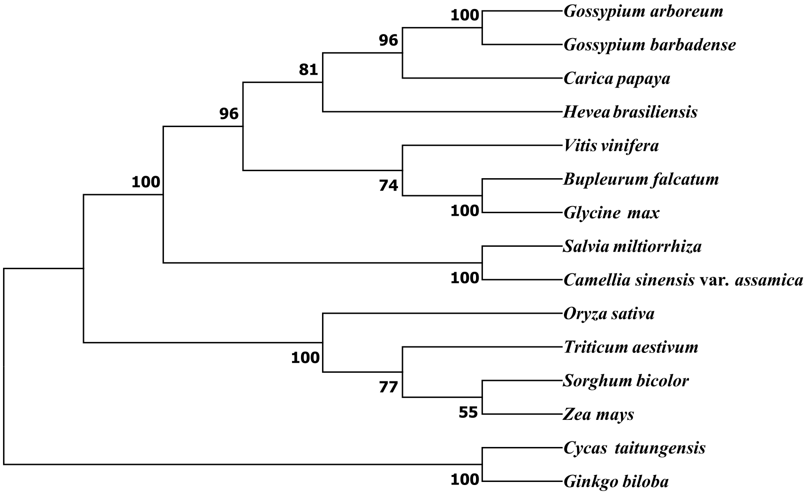

Fig. 6 Phylogeny inferred from 13 genes common in the 15 plant mitochondrial genomes. Neighbor-joining tree of *C. sinensis* var. *assamica* cv. *Yunkang 10* with other 14 species based on 13 conserved protein-coding gene sequences with bootstrap support values on each node. The mt sequence of *Cycas taitungensis* and *Ginkgo biloba* were set as outgroup.

For mitochondria genome assembly, the PE and MP sequencing reads were used separately. Briefly, we first performed *de novo* assembly with VELVET v1.2.08[41], which was previously described[42,43]. Scaffolds were constructed using SSPACE v.3.0[44]. False connection was manually removed based on the coverage and distances of paired reads. Gaps between scaffolds were then filled with GapCloser (version 1.12)[45,46] using all pair-end reads.

Code Availability

The following bioinformatic tools and versions were used for generating all results as described in the main text:

1. Bowtie2, version 2.3.4.3, was used for aligning sequencing reads to long reference sequences with default parameters: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
2. CLC Genomics Workbench, version 3.6.1, was used for genome assembly with default parameters: https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/
3. Velvet, version 1.2.08, was used for genome *de novo* assembly, which was previously described: https://www.ebi.ac.uk/~zerbino/velvet/
4. SSPACE, version 3.0, was used for genome scaffolds assembly with default parameters: https://www.baseclear.com/services/bioinformatics/basetools/sspace-standard/
5. GapCloser, version 1.12, was used to fill the gaps between scaffolds with default parameters: https://sourceforge.net/projects/soapdenovo2/files/GapCloser/
6. DOGMA (an online tool), accessed at 12/2018, was used for annotating cp genomes with default parameters: http://dogma.ccbb.utexas.edu/
7. Mitofy (an online tool), accessed at 12/2018, was used for annotating plant mt genomes with default parameters: http://dogma.ccbb.utexas.edu/mitofy/
8. tRNAscanSE, VERSION 1.3.1, was used to search tRNA with default parameters: http://lowelab.ucsc.edu/tRNAscan-SE/
9. Organellar Genome DRAW (an online tool), accessed at 12/2018, was used for creating high quality visual representation of cp gemome with default parameters: https://chlorobox.mpimp-golm.mpg.de/OGDraw.html
10. MISA,version 1.0, was used for annotating SSR with 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): http://pgrc.ipk-gatersleben.de/misa/misa.html
11. REPuter (an online tool), accessed at 1/2019, was used for annotating long repeated sequences with the following parameters: minimal length 50 nt; mis match 3 nt: https://bibiserv.cebitec.uni-bielefeld.de/reputer/
12. PREP-cp (an online tool), accessed at 1/2019, was used for predicting RNA editor for plant cp genes with the cutoff score (C-value) setting to 0.8: http://prep.unl.edu/
13. PREP-mt (an online tool), accessed at 1/2019, was used for predicting RNA editor for plant mt genes with the cutoff score (C-value) setting to 0.6: http://prep.unl.edu/
14. MEGA, version 7.0.26, was used for phylogenomics and phylomedicine at 1000 bootstrap: https://www.megasoftware.net/
15. ClustalW, version 2, was used for multiple sequence alignment with default parameters: https://www.ebi.ac.uk/Tools/msa/clustalw2/

References

1. Mondal, T. K., Bhattacharya, A., Laxmikumaran, M. & Singh Ahuja, P. Recent Advances of Tea (Camellia Sinensis) Biotechnology. *Plant Cell, Tissue and Organ Culture* **76**, 195–254 (2004).
2. Banerjee, B. Botanical Classification of Tea. (Chapman and Hall, London, 1992).
3. Ming, T. & Bartholomew, B. *Theaceae. In Flora of China*. (Beijing and St. Louis: Science Press and Missouri Botanical Garden, 2007).
4. 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 Evol Biol* **14**, 151 (2014).
5. 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* **7**, e29704 (2012).
6. Mccauley, D. E., Stevens, J. E., Peroni, P. A. & Raveill, J. A. The Spatial Distribution of Chloroplast DNA and Allozyme Polymorphisms within a Population of Silene alba (Caryophyllaceae). *American Journal of Botany* **83**, 727–731 (1996).
7. Small, R. L. & Wendel, R. C. C. J. Use of Nuclear Genes for Phylogeny Reconstruction in Plants. *Australian Systematic Botany* **17**, 145–170 (2004).
8. Jansen, R. K. *et al*. Analysis of 81 Genes From 64 Plastid Genomes Resolves Relationships in Angiosperms and Identifies Genome-Scale Evolutionary Patterns. *Proceedings of the National Academy of Sciences* **104**, 19369 (2007).
9. Parks, M., Cronn, R. & Liston, A. Increasing Phylogenetic Resolution at Low Taxonomic Levels Using Massively Parallel Sequencing of Chloroplast Genomes. *Bmc Biology* **7**, 84 (2009).
10. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic Analysis of 83 Plastid Genes Further Resolves the Early Diversification of Eudicots. *Proceedings of the National Academy of Sciences* **107**, 4623 (2010).
11. Richly, E. & Leister, D. NUPTs in Sequenced Eukaryotes and their Genomic Organization in Relation to NUMTs. *Molecular Biology and Evolution* **21**, 1972–1980 (2004).
12. Schuster, W. & Brennicke, A. Plastid, Nuclear and Reverse Transcriptase Sequences in the Mitochondrial Genome of Oenothera: Is Genetic Information Transferred Between Organelles Via RNA? *EMBO J* **6**, 2857–2863 (1987).
13. Stern, D. B. & Lonsdale, D. M. Mitochondrial and Chloroplast Genomes of Maize Have a 12-Kilobase DNA Sequence in Common. *Nature* **299**, 698–702 (1982).
14. Vaughn, J. C., Mason, M. T., Sper-Whitis, G. L., Kuhlman, P. & Palmer, J. D. Fungal Origin by Horizontal Transfer of a Plant Mitochondrial Group I Intron in the Chimeric CoxI Gene of Peperomia. *Journal of molecular evolution* **41**, 563 (1995).
15. Alverson, A. J. *et al*. Insights Into the Evolution of Mitochondrial Genome Size From Complete Sequences of Citrullus Lanatus and Cucurbita Pepo (Cucurbitaceae). *Mol Biol Evol* **27**, 1436–1448 (2010).
16. Ward, B. L., Anderson, R. S. & Bendich, A. J. The Mitochondrial Genome is Large and Variable in a Family of Plants (Cucurbitaceae). *Cell* **25**, 793–803 (1981).
17. Sloan, D. B. *et al*. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. *PLoS Biol* **10**, e1001241 (2012).
18. Palmer, J. D. & Herbon, L. A. Plant Mitochondrial DNA Evolves Rapidly in Structure, but Slowly in Sequence. *J Mol Evol* **28**, 87–97 (1988).
19. Marechal, A. & Brisson, N. Recombination and the Maintenance of Plant Organelle Genome Stability. *New Phytol* **186**, 299–317 (2010).
20. Zhang, Q. *et al*. The Complete Chloroplast Genome Sequence of Camellia Mingii (Theaceae), a Critically Endangered Yellow Camellia Species Endemic to China. *Mitochondrial DNA Part B* **4**, 1338–1340 (2019).
21. Lin, Y. *et al*. Characterization of the Complete Chloroplast Genome of Camellia Renshanxiangiae (Theaceae). *Mitochondrial DNA Part B* **4**, 1490–1491 (2019).
22. Li, W., Zhang, C., Guo, X., Liu, Q. & Wang, K. Complete Chloroplast Genome of Camellia Japonica Genome Structures, Comparative and Phylogenetic Analysis. *PLOS ONE* **14**, e216645 (2019).
23. Park, J. *et al*. The Complete Chloroplast Genome of Common Camellia Tree, Camellia Japonica L. (Theaceae), Adapted to Cold Environment in Korea. *Mitochondrial DNA Part B* **4**, 1038–1040 (2019).
24. Park, J. *et al*. The Complete Chloroplast Genome of Common Camellia Tree in Jeju Island, Korea, Camellia Japonica L. (Theaceae): Intraspecies Variations On Common Camellia Chloroplast Genomes. *Mitochondrial DNA Part B* **4**, 1292–1293 (2019).
25. Li, W. *et al*. Characterization of the Complete Chloroplast Genome of Camellia Granthamiana (Theaceae), a Vulnerable Species Endemic to China. *Mitochondrial DNA Part B* **3**, 1139–1140 (2018).
26. Liu, M. *et al*. Characterization of the Complete Chloroplast Genome of the Camellia Nitidissima, an Endangered and Medicinally Important Tree Species Endemic to Southwest China. Mitochondrial DNA Part B 3, 884, 886, 885, 887 (2018).
27. Liu, Y. & Han, Y. The Complete Chloroplast Genome Sequence of Endangered Camellias (Camellia Pubifurfuracea). *Conservation Genetics Resources* **10**, 843–845 (2018).
28. Dong, M. *et al*. The Complete Chloroplast Genome of an Economic Plant, Camellia Sinensis Cultivar Anhua, China. *Mitochondrial DNA Part B* **3**, 558–559 (2018).
29. Li, W., Xing, F., Ng, W. L., Zhou, Y. & Shi, X. The Complete Chloroplast Genome Sequence of Camellia Ptilophylla (Theaceae): A Natural Caffeine-Free Tea Plant Endemic to China. *Mitochondrial DNA Part B* **3**, 426–427 (2018).
30. Liu, Y. & Han, Y. The Complete Chloroplast Genome Sequence of Camellias (Camellia Fangchengensis). *Mitochondrial DNA Part B* **3**, 34–35 (2018).
31. Xu, X., Zheng, W. & Wen, J. The Complete Chloroplast Genome of the Long Blooming and Critically Endangered Camellia Azalea. *Conservation Genetics Resources* **10**, 5–7 (2018).
32. Zhang, W., Zhao, Y., Yang, G., Tang, Y. & Xu, Z. Characterization of the Complete Chloroplast Genome Sequence of Camellia Oleifera in Hainan, China. *Mitochondrial DNA Part B* **2**, 843–844 (2017).
33. Kim, S., Cho, C. H., Yang, M. & Kim, S. The Complete Chloroplast Genome Sequence of the Japanese Camellia (Camellia Japonica L.). *Mitochondrial DNA Part B* **2**, 583–584 (2017).
34. Wang, G., Luo, Y., Hou, N. & Deng, L. The Complete Chloroplast Genomes of Three Rare and Endangered Camellias (Camellia Huana, C. Liberofilamenta and C. Luteoflora) Endemic to Southwest China. *Conservation Genetics Resources* **9**, 583–585 (2017).
35. Tong, Y., Wu, C. & Gao, L. Characterization of Chloroplast Microsatellite Loci From Whole Chloroplast Genome of Camellia Taliensis and their Utilization for Evaluating Genetic Diversity of Camellia Reticulata (Theaceae). *Biochemical Systematics and Ecology* **50**, 207–211 (2013).
36. Yang, J. B., Yang, S. X., Li, H. T., Yang, J. & Li, D. Z. Comparative Chloroplast Genomes of Camellia Species. *PLoS One* **8**, e73053 (2013).
37. Kaundun, S. S. & Matsumoto, S. Molecular Evidence for Maternal Inheritance of the Chloroplast Genome in Tea, Camellia Sinensis (L.) O. Kuntze. *Journal of the Science of Food and Agriculture* **91**, 2660–2663 (2011).
38. Xia, E. *et al*. The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. *Molecular Plant* **10**, 866–877 (2017).
39. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA Extraction Protocol for Plants Containing High Polysaccharide and Polyphenol Components. *Plant Molecular Biology Reporter* **15**, 8–15 (1997).
40. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol* **10**, R25 (2009).
41. Zerbino, D. R. & Birney, E. Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. *Genome Res* **18**, 821–829 (2008).

42. Zhu, A., Guo, W., Jain, K. & Mower, J. P. Unprecedented Heterogeneity in the Synonymous Substitution Rate within a Plant Genome. *Mol Biol Evol* **31**, 1228–1236 (2014).
43. Grewe, F. *et al*. Comparative Analysis of 11 Brassicales Mitochondrial Genomes and the Mitochondrial Transcriptome of Brassica Oleracea. *Mitochondrion* **19 Pt B**, 135–143 (2014).
44. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding Pre-Assembled Contigs Using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
45. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: A De Novo Assembly Approach to Fill the Gap within Paired Reads. *BMC Bioinformatics* **13**(Suppl 14), S8 (2012).
46. Luo, R. *et al*. SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read De Novo Assembler. *Gigascience* **1**, 18 (2012).
47. Wyman, S. K., Jansen, R. K. & Boore, J. L. Automatic Annotation of Organellar Genomes with DOGMA. *Bioinformatics* **20**, 3252–3255 (2004).
48. Lowe, T. M. & Eddy, S. R. TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
49. Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW): A Tool for the Easy Generation of High-Quality Custom Graphical Maps of Plastid and Mitochondrial Genomes. *Curr Genet* **52**, 267–274 (2007).
50. Kurtz, S. *et al*. REPuter: The Manifold Applications of Repeat Analysis On a Genomic Scale. *Nucleic Acids Res* **29**, 4633–4642 (2001).
51. Mower, J. P. PREP-Mt: Predictive RNAEditor for Plant Mitochondrial Genes. *BMC Bioinformatics* **6**, 96 (2005).
52. Mower, J. P. The PREP Suite: Predictive RNA Editors for Plant Mitochondrial Genes, Chloroplast Genes and User-Defined Alignments. *Nucleic Acids Res* **37**, W253–W259 (2009).
53. Chaw, S. M. *et al*. The Mitochondrial Genome of the Gymnosperm Cycas Taitungensis Contains a Novel Family of Short Interspersed Elements, Bpu Sequences, and Abundant RNA Editing Sites. *Mol Biol Evol* **25**, 603–615 (2008).
54. Ward, G. C. & Levings, C. R. The Protein-Encoding Gene T-urf13 is Not Edited in Maize Mitochondria. *Plant Mol Biol* **17**, 1083–1088 (1991).
55. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870–1874 (2016).
56. Larkin, M. A. *et al*. Clustal W and Clustal X Version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
57. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708522 (2017).
58. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708523 (2017).
59. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708528 (2017).
60. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708529 (2017).
61. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708545 (2017).
62. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRX2708546 (2017).
63. *BIGD Genome Sequence Archive*, http://bigd.big.ac.cn/gsa/browse/CRA001582 (2019)
64. Gao, C-W. & Gao, L-Z. Camellia sinensis var. assamica cultivar Yunkang 10 plastid, complete genome. *GenBank*, https://identifiers.org/ncbi/insdc:MH019307 (2018).
65. *BIGD Genome Warehouse*, http://bigd.big.ac.cn/search?dbId=gwh&q=GWHAAIB00000000 (2019).
66. Zhang, F. Camellia sinensis var. assamica mitochondrion, complete genome. *GenBank*, https://identifiers.org/ncbi/insdc:MK574876 (2019).
67. Zhang, F. Camellia sinensis var. assamica mitochondrion, complete genome. *GenBank*, https://identifiers.org/ncbi/insdc:MK574877 (2019).
68. *BIGD Genome Warehouse*, http://bigd.big.ac.cn/search?dbId=gwh&q=GWHAAIC00000000 (2019).
69. Zhang, F. Deciphering tea tree chloroplast and mitochondrial genomes of Camellia sinensis var. assamica. *figshare*. https://doi.org/10.6084/m9.figshare.c.4420955 (2019).

Acknowledgements

We would thank Yunnan Tea Research Institute for providing tea plant materials in this study. We are grateful An-dan Zhu for technical support and anonymous reviewers for valuable comments on the manuscript. This work was supported by the Project of Innovation Team of Yunnan Province and Ten Thousands Talents Program of China (to L. Z. Gao).

Author Contributions

Li-zhi Gao designed the study; Fen Zhang, Wei Li and Dan Zhang assembled, annotated and analyzed the mt genome; Cheng-wen Gao assembled, annotated and analyzed the cp genome; Fen Zhang, Wei Li and Cheng-wen Gao drafted the manuscript; Li-zhi Gao revised the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.