

Characterization of the complete chloroplast genome sequence of *Camellia tetracocca* (Theaceae)

Lianghui Yi, Yanli Wang, Yunze Li, Dandan Zhang and Wei Tong

State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China

ABSTRACT

Camellia tetracocca H.T. Chang 1981 is an important wild relative of cultivated tea plants. Its leaves are widely used by local people to make tea, showing great economic and breeding values. We here report the complete chloroplast genome of *C. tetracocca* using Illumina sequencing technology. The complete chloroplast genome of *C. tetracocca* is 157,026 bp in length, and structurally contains a pair of inverted repeat regions (IRa and IRb, 26,052 bp) separated by a large single-copy region (LSC, 86,669 bp) and a small single-copy region (SSC, 18,253 bp). It is composed of 131 predicted genes, including 86 protein-coding genes, 37 transfer RNA genes, and eight ribosomal RNA genes. The overall GC content is 37.31%. Phylogenetic analysis among four *Camellia* species and 11 other close species reveals a close relationship between *C. tetracocca* and *C. gymnogyna*.

ARTICLE HISTORY

Received 24 August 2023
Accepted 4 February 2024

KEYWORDS

Camellia tetracocca;
chloroplast genome;
phylogenetic analysis

Introduction

Camellia tetracocca is a perennial wild tea plant belonging to the section *Thea* of genus *Camellia* (Chang 1981). It is mainly distributed in Pu'an County, southwest Guizhou Province of China, and grows in the mountains and forests of 1700–1950 m above sea level (Yuan and Qian 2009). The discovery of seed fossil of *C. tetracocca* in Mountain Yuntou of Pu'an County over 1 million years ago suggested a long history of *C. tetracocca* in southwest Guizhou, showing a precious historical and cultural value (Yan 2009). In addition, similar to the cultivated tea plants (*C. sinensis*), the leaves of *C. tetracocca* are also commonly used by local people to make tea – one of the three most worldwide popular nonalcoholic beverages with numerous good health effects. The tea products of *C. tetracocca* have significantly increased the income of local people; however, excessive harvesting has also seriously damaged its wild habitat. It is urgent to strengthen the scientific conservation and utilization of this valuable *Camellia* plants. Comparative chloroplast genomics is an effective tool for the study of plant phylogeny and conservation biology (Daniell et al. 2016). Although recent studies have released more than 55 chloroplast genomes of *Camellia* species (Huang et al. 2014; Xia et al. 2020), the complete chloroplast genome of *C. tetracocca* has yet to be sequenced.

In the present study, we reported the chloroplast genome sequence of *C. tetracocca*. Through comprehensive analysis, we aim to provide insights into the genome characterization and evolution of this important *Camellia* plant, which would help genetic breeding of cultivated tea plants in the future.

Materials and methods

Plant material collection and DNA extraction



The healthy and fresh leaves of *C. tetracocca* (Figure 1) were collected from Pu'an County, Guizhou Province, China (104°95'E, 25°78'N) in 13 April 2018. The specimens were preserved in the Laboratory of State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University under the specimen number SQ#119 (<https://tealab.ahau.edu.cn>; Dr. Wei Tong, wtong@ahau.edu.cn). The genomic DNA was isolated by using a modified CTAB method.


Library construction and sequencing

Libraries were constructed according to the Illumina's protocol, and then sequenced using Illumina NovaSeq 6000 platform. The raw sequencing data were processed using SolexaQA (Cox et al. 2010) to remove adaptors, low-quality base (Phred score <20), short read (length <30 bp), and potential contaminations.

Sequence assembly and annotation

The whole genome sequencing reads were aligned against 55 publicly available *Camellia* chloroplast Refseq genomes using BWA (Li and Richard 2010) to extract the candidate reads for chloroplast genome sequences of *C. tetracocca*. Those candidate chloroplast reads were then assembled using GetOrganelle with parameter '-R 30 -k 65,85,105,115 -F

CONTACT Wei Tong  tw7649116@gmail.com  State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei 230036, China

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/23802359.2024.2316067>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.



Figure 1. Plant individual of *C. tetracocca*. (A) Whole plant of the collected individual. (B) Leaf margin morphology of mature leaves. (C, D) Buds, young, and mature leaves of *C. tetracocca* (imaged by Lianghui Yi).

embplant_pt' (Jin et al. 2020). Annotation of the chloroplast genome of *C. tetracocca* was performed using GeSeq (Tillich et al. 2017), followed by manual adjustments according to the homologous genes of cultivated tea plants *C. sinensis* (MW148820.1). The resulted GenBank annotation file was used to visualize chloroplast genome map by CPGView (<http://www.1kmpg.cn/cpgview/>). The simple sequence repeat (SSR) of *C. tetracocca* chloroplast genome was identified using MISA package with default parameters 'unit size/minimum number of repeats = (1/10), (2/5), (3/4), (4/3), (5/3), (6/3), Maximal number of bases interrupting 2 SSRs = 100' (Beier et al. 2017).

Phylogenetic tree construction

A total of 396 chloroplast genomes belong to 125 *Camellia* species were collected from NCBI GenBank database, such as *C. sinensis*, *C. chrysanthoides*, *C. gymnogyna*, *C. japonica*, *C. oleifera*, *C. ptilophylla*, *C. reticulata*, *C. sasanqua*, *C. tachangensis*, *C. taliensis*, and *C. tetracocca* (Supplemental Table 1). The complete sequences of those genomes were further multiple aligned using MAFFT (Kato et al. 2002). The phylogenetic tree was then reconstructed using IQ-TREE after estimated the best suitable model for phylogeny construction (Minh et al. 2020). Phylogenetic tree was visualized using Figtree (<https://github.com/rambaut/figtree>).

Results

We sequenced the whole genome sequence of *C. tetracocca* using Illumina NovaSeq 6000 sequencing platform. As a result, a total of 17,863,628 paired-end chloroplast-derived

reads of *C. tetracocca*, accounting for $11,378\times$ of coverage, were obtained for subsequent assembly using GetOrganelle (Jin et al. 2020). Annotation of the chloroplast genome was initially performed using GeSeq (Tillich et al. 2017) and further manually adjusted by comparison with homologous genes of *C. sinensis* chloroplast genome (MW148820.1) (Chen et al. 2021). The annotated chloroplast genome of *C. tetracocca* has been deposited into NCBI GenBank database under the accession number OK405020.1.

The complete chloroplast genome of *C. tetracocca* was 157,026 bp in size, with a G + C content of 37.31% (Figure 2). It contains a typical quadrant structure that composed of a large single-copy (LSC) region (86,669 bp), an small single-copy (SSC) region (18,253 bp), and a pair of inverted repeat (IR) regions (26,052 bp). A total of 131 genes are predicted, which include 86 protein-coding genes, 37 tRNA genes, and eight rRNA genes (Supplemental Figures 1 and 2). Annotation of the chloroplast genome further identified a total of 71 SSRs. Of them, mono-nucleotide repeats (54; 76.1%) are the most abundant SSR type, followed by tetra-nucleotide repeats (11; 15.5%), di-nucleotides repeats (4; 5.6%), and tri-nucleotides repeats (2; 2.8%) (Supplemental Table 2 and Figure 3).

To further investigate the phylogenetic relationship of *C. tetracocca* with other *Camellia* plants, a total of 395 additional chloroplast genomes from 125 *Camellia* species were obtained from NCBI database for phylogenetic analysis. The maximum-likelihood phylogenetic tree of those chloroplast genomes was then constructed using IQ-TREE. Results showed that most of the species from same sections of genus *Camellia* could be clustered together, which is basically consistent with the phylogenetic treatments in Ming and

Camellia tetracocca

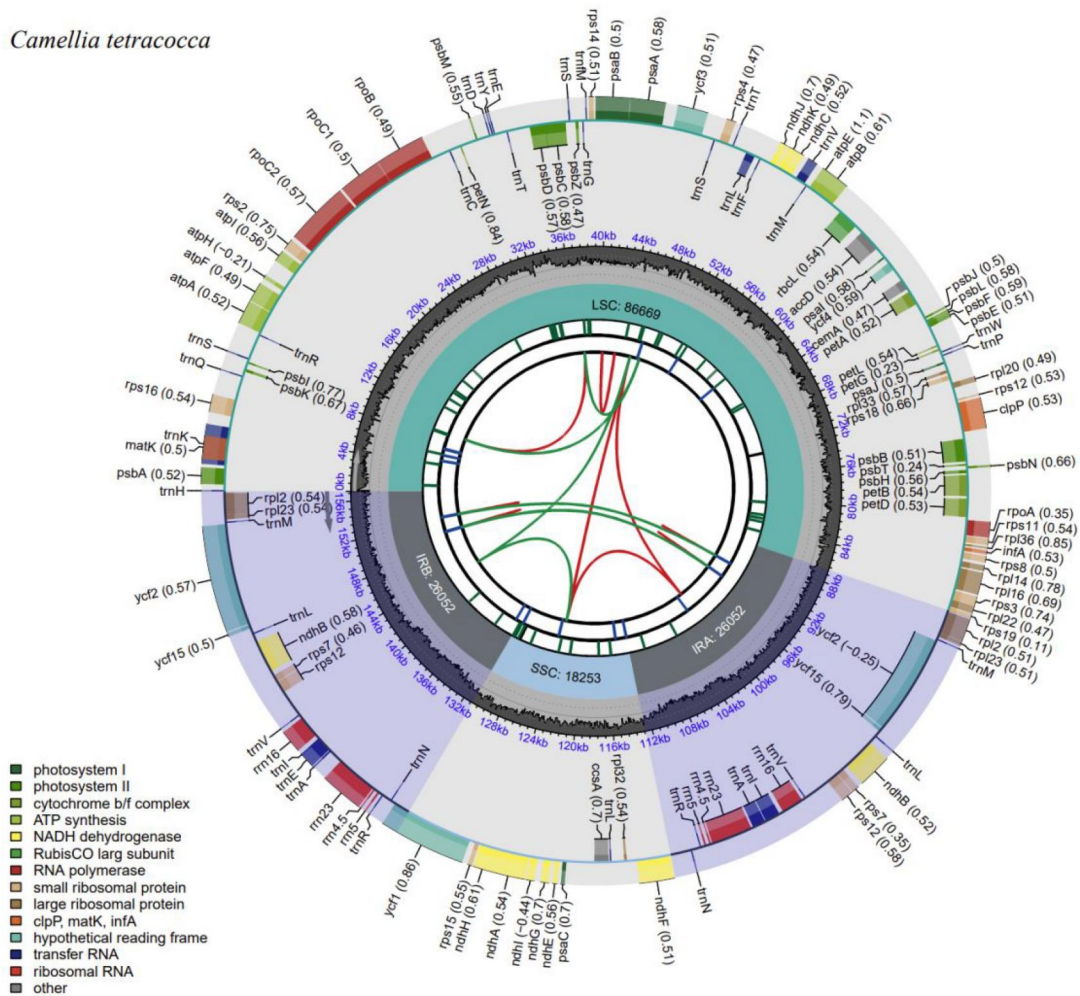


Figure 2. The complete chloroplast genome map of *C. tetracocca*. Arrangement of 131 genes represented in the map, including 86 protein coding genes, 37 tRNA genes, and eight rRNA genes. The GC% along the chloroplast is represented by the inner circle.

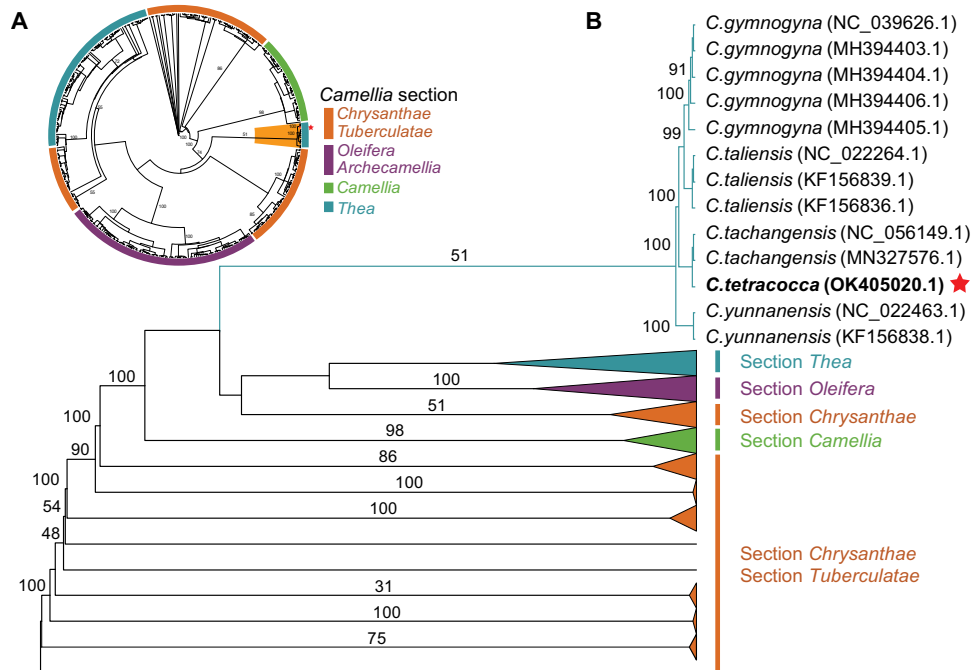


Figure 3. Phylogenetic relationship between *C. tetracocca* and 395 other *Camellia* plants using complete chloroplast genome sequences. (A) The overall phylogeny of all the selected *Camellia* species. A detailed phylogenetic relationship of those 396 *Camellia* species is illustrated in Supplemental Figure 4. (B) A clear relationship of *C. tetracocca* with other several species in same branch. Bootstrap value for some nodes was shown on the branch. The red filled asterisk represents *C. tetracocca* reported in this study.

Chang's classification system of *Camellia* (Chang 1981; Ming and Bartholomew 2007). We found that *C. tetracocca* show closer relationships with *C. tachangensis*, followed by *C. gymnogyna* and *C. taliensis* among all the examined species (Figure 3). This observation is highly supported by the latest taxonomical combination of *C. tetracocca* with *C. tachangensis* in Ming's classification system of *Camellia* (Ming and Bartholomew 2007).

Discussion and conclusions

In this study, the chloroplast genome sequence of *C. tetracocca* was reported for the first time. Read mapping shows that all genomic regions of *C. tetracocca* chloroplast genome is well covered by sequencing reads, suggesting a high accuracy of the genome assembly (Supplemental Figure 5). Results show that *C. tetracocca* was closely related to *C. tachangensis*, *C. gymnogyna*, and *C. taliensis*. The overall obtained results and findings will not only provide valuable genetic markers for tea breeding programs, but also offer vital phylogenetic framework for future population genetic studies and conservations in *C. tetracocca*. In the future, we will further sequence the whole nuclear genome of *C. tetracocca* and investigate the evolution patterns of tea quality related secondary metabolites to identify key genes associated with tea quality and disease resistance, which will help the genetic research and breeding progress of cultivated tea plants in the future.

Acknowledgements

We would appreciate the anonymous reviewers for their insightful suggestions and comments on the manuscript.

Author contributions

W.T. involved in the conception and design. L.Y. analyzed the data and wrote the article. L.Y., Y.W., Y. L., D.Z., and W.T. interpreted the data; all authors revised the manuscript and agreed to be accountable for all aspects of the work.

Ethical approval

This work does not require Ethical approval or specific permissions according to the recommendations of the Research Ethic Committee of Anhui Agricultural University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China under Grant numbers [32172626] and [32002086]; the Natural

Science Research Project of University in Anhui Province under Grant number [2022AH050867].

Data availability statement

The assembled chloroplast genome sequence data that support the findings of this study are freely available in GenBank of NCBI (<https://www.ncbi.nlm.nih.gov/>) under the accession number of OK405020.1. The raw sequencing data of *C. tetracocca* have also been deposited into the NCBI Bio-Project and SRA database with the accession number of PRJNA1003289 and SRR25569055, respectively. The information of the plant material was deposited into NCBI Bio-Sample database with the accession number of SAMN36886831.

References

- Beier S, Thiel T, Münch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. 33(16):2583–2585. doi:10.1093/bioinformatics/btx198.
- Chang HT. 1981. A taxonomy of the genus *Camellia*. *Acta Sci Nat Univ Sunyatseni*. 1:1–180.
- Chen S, Li RY, Ma YY, Lei SR, Ming R, Zhang XT. 2021. The complete chloroplast genome sequence of *Camellia sinensis* var. *sinensis* cultivar Tieguanyin (Theaceae). *Mitochondrial DNA B Resour*. 6(2):395–396. doi:10.1080/23802359.2020.1869615.
- Cox M, Peterson D, Biggs P. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*. 11(1):485. doi:10.1186/1471-2105-11-485.
- Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol*. 17(1):134. doi:10.1186/s13059-016-1004-2.
- Huang H, Shi C, Liu YL, Mao SY, Gao LZ. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing genome structure and phylogenetic relationships. *BMC Evol Biol*. 14(1):151.
- Jin JJ, Yu WB, Yang JB, Song Y, Claude WD, Yi TS, Li DZ. 2020. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol*. 21(1):241. doi:10.1186/s13059-020-02154-5.
- Katoh K, Kazuharu M, Kuma K, Takashi M. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30(14):3059–3066. doi:10.1093/nar/gkf436.
- Li H, Richard D. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 26(5):589–595. doi:10.1093/bioinformatics/btp698.
- Ming T, Bartholomew B. 2007. Theaceae. In: Wu Z, Raven P, editors. *Flora of China*, Vol. 12. Beijing: Science Press; p. 366–478.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 37(5):1530–1534. doi:10.1093/molbev/msaa015.
- Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*. 45(W1):W6–W11. doi:10.1093/nar/gkx391.
- Xia E, Tong W, Wu Q, Wei S, Zhao J, Zhang Z, Wei C, Wan X. 2020. Tea plant genomics: achievements, challenges and perspectives. *Hortic Res*. 7(1):7. doi:10.1038/s41438-019-0225-4.
- Yan D. 2009. Research progress on tea germplasm resources and investigation of wile tea resource in Guizhou. *Guizhou Agric Sci*. 37(7):184–187.
- Yuan M, Qian C. 2009. Study on endemic plant *Camellia tetracocca* in Pu'an Country, Guizhou. *Guizhou Sci*. 27(2):80–85.