Data in Brief 67 (2026) 112894

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Complete chloroplast genome of *Camellia yokdonensis* Dung bis & Hakoda (Theaceae), a useful genomic data for elucidating evolution of Vietnamese endemic tea species

Nhat Nam Nguyen[a], Hoang Dang Khoa Do[b,c,*]

[a] School of Agriculture and Aquaculture, Tra Vinh University, Vinh Long 87000, Vietnam
[b] Functional Genomics Research Center, NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City 72820, Vietnam
[c] Nguyen Tat Thanh University Center for Hi-Tech Development, Saigon Hi-Tech Park, Ho Chi Minh City, Vietnam

ARTICLE INFO

Article history:
Received 6 May 2026
Revised 26 May 2026
Accepted 26 May 2026
Available online 28 May 2026

Keywords:
Comparative genomics
Plastid genome
Plastome evolution
Theaceae
Tea evolution

ABSTRACT

Various endemic *Camellia* L. species have been identified in Vietnam. However, most genomic data of these endemic tea taxa have not been characterized. In the current study, we employed next-generation sequencing method to sequence and complete chloroplast genome of *Camellia yokdonensis* Dung bis & Hakoda, an endemic species found in the central area of Vietnam. This circular genome was 157,090 bp in length and had 37.7% GC content. In addition, this quadripartite genome consisted of a large single-copy region (86,657 bp and 35.3% GC), a small single-copy region (18,273 bp and 30.6% GC), and two inverted repeat regions (26,080 bp and 43% GC each). The annotation result revealed 79 protein-coding genes, 30 transfer RNA genes, and four ribosomal RNA genes. Phylogenetic analysis pointed out that *C. yokdonensis* was placed as a basal clade to other examined *Camellia* species. This study provides initial genomic data for further

* Corresponding author at: Functional Genomics Research Center, NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City 72820, Vietnam.
E-mail address: dhdkhoa@ntt.edu.vn (H.D.K. Do).

https://doi.org/10.1016/j.dib.2026.112894

studies examining evolutionary history of *C. yokdonensis* and other related Theaceae taxa.

Specifications Table

Subject	Biology
Specific subject area	Genomics: Chloroplast genome
Type of data	Table, Figures. Raw, Analyzed.
Data collection	Fresh healthy leaves of *Camellia yokdonensis* were collected from natural habitat and used for total genomic DNA extraction using a commercial kit. The sequencing process was conducted using Onso system of PacBio. The raw reads were filtered and assembled using Geneious Prime v2024.0.1. The complete chloroplast genome was annotated and illustrated using Geseq and OGDRAW, respectively. Phylogenetic analysis of 30 *Camellia* species and three outgroups inferred from complete chloroplast genomes were conducted using IQ-TREE v2.4.0 after alignment step using MUSCLE v5. The phylogenetic tree was illustrated using Figtree v1.4.5.
Data source location	Location: Buon Don, DakLak, Vietnam (12°47′09.1"N 107°53′55.7"E, altitude: 227 m). Voucher information: The specimen of *Camellia yokdonensis* was deposited into NTT Hi-Tech Institute, Nguyen Tat Thanh University under voucher number FGRC_NTHI20250101_P037 (contact person: Hoang Dang Khoa Do, email: dhdkhoa@ntt.edu.vn).
Data accessibility	Repository name: NCBI Direct URL to data of complete chloroplast genome and raw data. 1/ https://www.ncbi.nlm.nih.gov/nuccore/PV612622 2/ https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1457648 3/ https://www.ncbi.nlm.nih.gov/biosample/SAMN57468249 4/ https://www.ncbi.nlm.nih.gov/sra/?term=SRR38263963 Repository name: Figshare Direct URL to data for phylogenetic analysis 10.6084/m9.figshare.32184297
Related research article	NONE

1. Value of the Data

- This dataset provides information about the complete chloroplast genome of *Camellia yokdonensis* regarding genome structure, genome size, gene content, and gene order. The complete chloroplast genome of *C. yokdonensis* provides initial genomic data for further studies examining the characterization of chloroplast genomes of other *C. yokdonensis* individuals and related *Camellia* species. The more available complete chloroplast genomes of *C. yokdonensis* are, the more comprehensive studies about its population genetics, evolutionary history, and molecular markers are conducted.
- Although different Vietnamese endemic *Camellia* chloroplast genomes have been published, these species are mainly distributed in Northern Vietnam. The *Camellia* species distributed in Southern Vietnam has not been reported in terms of genomic data, such as nuclear, mitochondrial, and chloroplast genomes. This genomic gap prevents further studies based on molecular data, such as divergence time estimation and biogeography of *Camellia* species. Therefore, the current genomic data of *C. yokdonensis* will be valuable as a gap-filling dataset for further genomic studies of *Camellia* and related species in Theaceae.
- Phylogenetic relationship analysis revealed a basal position of *Camellia yokdonensis*, which was not found in the previous studies due to the lack of the *C. yokdonensis* chloroplast genome. The phylogenetic position of *C. yokdonensis* suggested a possibility of unique spe-

ciation of *Camellia* taxa endemic to Southern Vietnam. However, the current phylogenetic analysis was inferred from chloroplast genome data, which may conflict with mitochondrial and nuclear genome datasets. Therefore, the current chloroplast genome data could be reused in further phylogenetic studies inferred from three genomes.

- The raw sequencing data of *Camellia yokdonensis* was generated from total genomic DNA, which included chloroplast, mitochondrial, and nuclear genomes. Therefore, the raw data could be used to explore the mitochondrial and nuclear genomes of *C. yokdonensis* for further comparative genomic analyses.

2. Background

Camellia L., containing 233 accepted species, is a genus of Theaceae and is native to East, South, and Southeast Asia [1]. This genus included various species which have ornamental and commercial values, especially *Camellia sinensis* (L.) Kuntze, an indispensable ingredient of tea [2]. Recently, more than 1300 nuclear genomes of *Camellia* species have been reported, providing new insights into metabolism of tea plants [3]. Vietnam is one of the original distributions of the *Camellia* species; therefore, new *Camellia* species have been identified recently [4–8]. However, the genomic data of Vietnamese endemic *Camellia* are limited. To date, some complete chloroplast genomes (cpDNA) of Vietnamese endemic *Camellia* species, including *Camellia huulungensis* Rosmann & Ninh, *Camellia flava* (Pit.) Sealy, and *Camellia tienii* Ninh, have been reported [9–11]. Therefore, in the current study, we employed next-generation sequencing method to complete and characterise cpDNA of *Camellia yokdonensis* Dung bis & Hakoda, an endemic *Camellia* in the central area of Vietnam (Fig. 1) [4]. The newly completed cpDNA will enlarge the genomic data as well as provide useful tools for further studies about species identification and phylogeny of *Camellia*.

Fig. 1. Photos of *Camellia yokdonensis*. A. Leaf. B and C. Flower buds. D. Flower. Small tree with oblong leaves. Terminal subsessile flower with orange/pink color. Flowers having 6-8 bracts, 7-10 petals, and over 100 stamens.

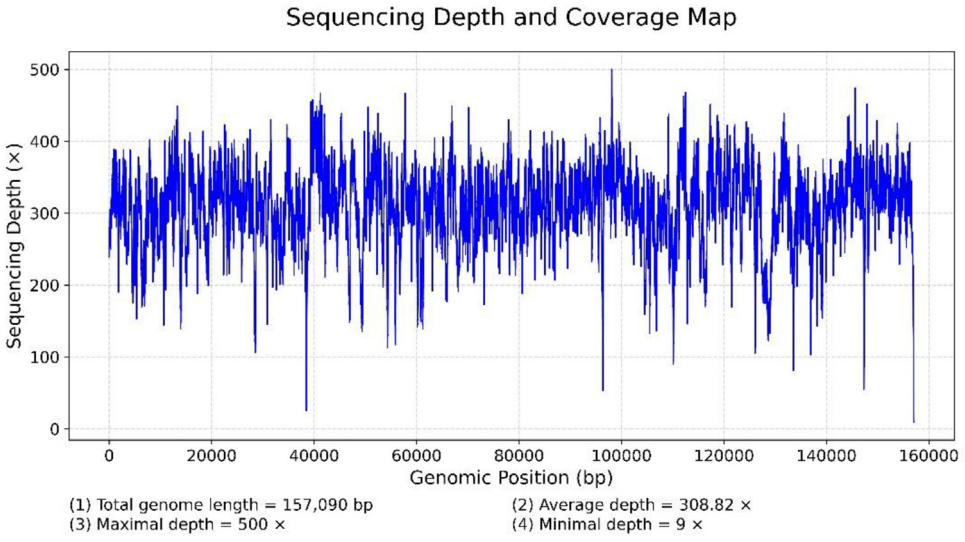

Fig. 2. Coverage depth of *Camellia yokdonensis* chloroplast genome.

3. Data Description

A total of 332,406 reads out of 59,974,172 raw reads which range from 100 bp to 150 bp were assembled to complete the cpDNA of *C. yokdonensis* with a minimum coverage of 9 ×, a maximum coverage of 500 ×, and an average coverage of 308 × (Fig. 2). The complete cpDNA of *C. yokdonensis* was 157,090 bp in length and had 37.7% GC content, which included a large single-copy region (86,657 bp and 35.3% GC), a small single-copy region (18,273 bp and 30.6% GC), and two inverted repeat regions (26,080 bp and 43% GC each) (Fig. 3). The annotation resulted in 113 unique coding regions including 79 protein-coding genes categorized in different functions, 30 transfer RNA genes, and four ribosomal RNA genes (Fig. 3, Table 1). Among the coding genes, *rps19, rpl2, rpl23, trnI_CAU, ycf2, trnL_CAA, ndhB, rps7, rps12, trnV_GAC, rrn16, trnI_GAU, trnA_UGC, rrn23, rrn4.5, rrn5, trnR_ACG, trnN_GUU*, and *ycf1* were duplicated in the inverted repeat region. However, *ycf1* and *rps19* exhibited incomplete duplication due to the expansion of the inverted repeat regions. Additionally, *rps16, atpF, rpoC1, petB, petD, rpl16, rpl2, ndhB, ndhA, trnK_UUU, trnI_GAU, trnA_UGC, trnG_UCC, trnL_UAA*, and *trnV_UAC* contained one intron, whereas *paf1* and *clpP1* had two introns. Notably, the *rps12* gene was trans-spliced with the locations of exon 2 and exon 3 in IR regions.

Phylogenetic analysis revealed the monophyletic status of *Camellia* species with high support value (bootstrap value = 100) (Fig. 4). However, the relationship among clades was quite weak although each clade had a strong bootstrap value. Notably, *C. yokdonensis* and another Vietnamese endemic *Camellia* species, *Camellia cattienensis* Orel, were placed as basal clades of other surveyed *Camellia* taxa.

4. Experimental Design, Materials and Methods

4.1. Plant sampling, DNA extraction, and next-generation sequencing

Healthy leaves of *C. yokdonensis* were collected at Buon Don, DakLak, Vietnam (12°47′09.1"N 107°53′55.7"E, altitude: 227 m) (Fig. 1). The specimen of *C. yokdonensis* was deposited into NTT Hi-Tech Institute, Nguyen Tat Thanh University under voucher number

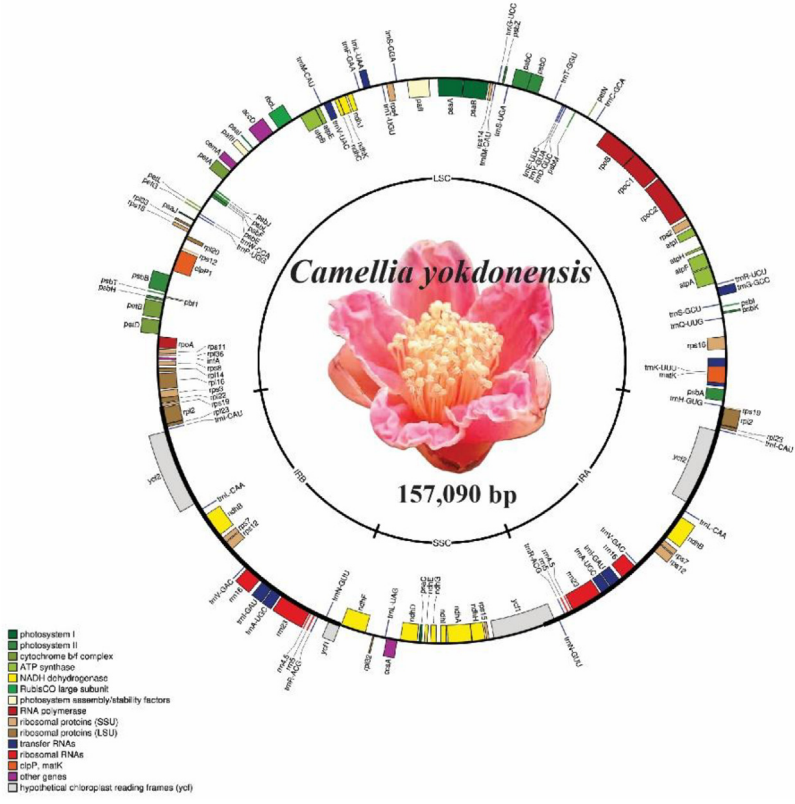

Fig. 3. The chloroplast genome map of *Camellia yokdonensis*. The inner and outer genes indicated the translation directions of clockwise and counterclockwise, respectively. Different colors indicate different functional groups of genes. The inner image is the flower of *C. yokdonensis*. The inner circle represents the locations of four regions in the chloroplast genome. LSC: large single-copy; SSC: small single-copy; IRA and IRB: inverted repeat region A and B, respectively.

FGRC_NTHI20250101_P037 (contact person: Hoang Dang Khoa Do, email: dhdkhoa@ntt.edu.vn). Then, the leaves were stored at −80°C for further experiments. The DNeasy Plant Pro Kit (Qiagen, USA) was used to extract total genomic DNA following the manufacturer's protocol. NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Fisher Scientific, USA) and 1% agarose gel electrophoresis were used to check the quality of the DNA sample with a minimum concentration of 50 ng/μL and a clear band on the gel. Then, the qualified DNA sample was applied for the Onso sequencing system (PacBio, USA) after library preparation following the manufacturer's instruction to generate paired-end reads of 150 bp with an insert size of 250 bp and 150 cycles, which were conducted at KTEST Co. Ltd. (Ho Chi Minh City, Vietnam).

4.2. Assembly and annotation of chloroplast genome

The raw reads (59,974,172 reads, ~8.81 Gb, ranging from 1 bp to 150 bp) were analysed using fastp v0.24.1 to identify and eliminate the reads which contained adapters and N bases, had a Qscore under 20, and were shorter than 100 bp [12]. The filtered reads were assembled to complete the cpDNA by following steps. First, the filtered reads were mapped to the coding region of *psbA, rpoB, psaA, petA, rpoA, ndhB, ndhD*, and *rps7* genes of *Camellia sinensis* (L.) Kuntze (PQ066596) which are distributed distantly within the cpDNA using Geneious

Table 1
Gene composition of *Camellia yokdonensis* chloroplast genome

Groups of genes	Name of genes
Ribosomal RNAs	*rrn4.5*[a], *rrn5*[a], *rrn16*[a], *rrn23*[a]
Transfer RNAs	*trnA_UGC*[a,b], *trnC_GCA*, *trnD_GUC*, *trnE_UUC*, *trnF_GAA*, *trnG_UCC*[b], *trnG_GCC*, *trnH_GUG*, *trnI_GAU*[a,b], *trnK_UUU*[b], *trnL_CAA*[a], *trnL_UAA*[b], *trnL_UAG*, *trnfM_CAU*, *trnI_CAU*[a], *trnM_CAU*, *trnN_GUU*[a], *trnP_UGG*, *trnQ_UUG*, *trnR_ACG*[a], *trnR_UCU*, *trnS_GCU*, *trnS_GGA*, *trnS_UGA*, *trnT_GGU*, *trnT_UGU*, *trnV_GAC*[a], *trnV_UAC*[b], *trnW_CCA*, *trnY_GUA*
Large units of ribosome	*rpl2*[a, b], *rpl14*, *rpl16*[b], *rpl20*, *rpl22*[a], *rpl23*[a], *rpl32*, *rpl33*, *rpl36*
Small units of ribosome	*rps2*, *rps3*, *rps4*, *rps7*[a], *rps8*, *rps11*, *rps12*[a], *rps14*, *rps15*, *rps16*[b], *rps18*, *rps19*[a]
RNA polymerase	*rpoA*, *rpoB*, *rpoC1*[b], *rpoC2*
Translational initiation factor	*infA*
Subunit of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *pafI*[c], *pafII*
Subunit of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *pbfI*, *psbM*, *psbT*, *psbZ*
Subunit of cytochrome	*petA*, *petB*[b], *petD*[b], *petG*, *petL*, *petN*
Subunit of ATP synthases	*atpA*, *atpB*, *atpE*, *atpF*[b], *atpH*, *atpI*
Large unit of Rubisco	*rbcL*
Subunit of NADH dehydrogenase	*ndhA*[b], *ndhB*[a, b], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Maturase	*matK*
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA	*accD*
C-type cytochrome synthesis gene	*ccsA*
ATP-dependent protease subunit P	*clpP1*[c]
Hypothetical proteins and conserved reading frames	*ycf1*[a], *ycf2*[a]

Note:
[a] Duplicated gene in IR region.
[b] Genes containing single intron.
[c] Genes containing two introns.

Prime v2024.0.1 via the Map to Reference function with default settings and 100 iterations. Then, the mapped reads were extracted and *de novo* assembled using Geneious Prime with default settings to complete the cpDNA of *C. yokdonenesis*. The coverage depth of the cpDNA was then generated using a previous protocol [13]. The cpDNA of *C. yokdonensis* was annotated using Geseq website (https://chlorobox.mpimp-golm.mpg.de/geseq.html) [14]. The annotations of protein-coding genes were verified using Geneious Prime v2024.0.1 for the start and stop codons. The map of the cpDNA was illustrated using OGDRAW v1.3.1 (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) with default settings [15].

4.3. Phylogenetic analysis

For reconstructing phylogenetic relationships, a total of 33 cpDNAs composed of 30 *Camellia* species and 3 outgroups of *Polyspora longicarpa* (Hung T.Chang) C.X.Ye ex B.M.Barthol. & T.L.Ming, *Tutcheria championii* Nakai, and *Schima superba* Gardner & Champ. of Theaceae were downloaded from the GenBank database. The complete cpDNAs were then aligned using MUSCLE v5 with default settings [16]. The aligned sequences were used to reconstruct the phylogenetic trees using the maximum likelihood (ML) method via IQ-TREE website [17]. The best substitution model of the aligned sequences, identified as GTR+I+G model, was evaluated using jModelTest2 [18]. For ML analysis, ultrafast bootstrap with 1000 bootstrap alignments, and the Sh-aLRT branch test with 1000 replicates were set up. The phylogenetic tree was illustrated and modified using Figtree v1.4.5 (https://github.com/rambaut/figtree/releases).

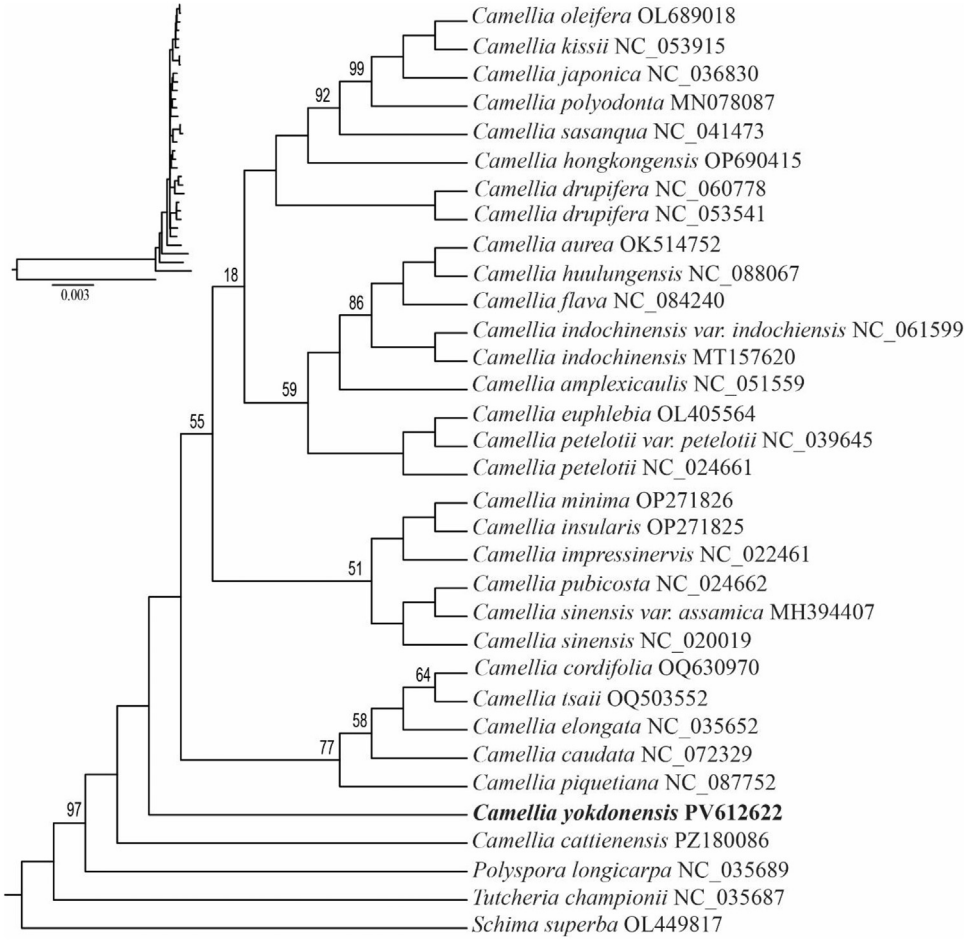

Fig. 4. Phylogenetic tree in cladogram format of *Camellia yokdonensis* and related species inferred from whole chloroplast genome using maximum likelihood method. Bold highlighted name indicates target species in the current study. Only bootstrap values under 100 were shown in the phylogenetic tree. The phylogram on the upper-left corner presents the original tree with nucleotide substitution rates. Numbers following the species names represent GenBank accession numbers

Limitations

Although this study reports the first complete chloroplast genome of *Camellia yokdonensis*, the dataset is generated from only one individual. Therefore, it only provides initial information of chloroplast genome of *C. yokdonensis*. More chloroplast genomic data should be explored to provide sufficient data for elucidating evolutionary history, population genetics, and molecular markers of *C. yokdonensis* and related species in Theaceae.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT Author Statement

NNN: Conceptualization, Data curation, Formal analysis, Writing – original draft. **HDKD:** Conceptualization, Methodology, Data curation, Formal analysis, Writing – review & editing.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We acknowledge Tra Vinh University for providing the opportunity to carry out this work. We acknowledge Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam for supporting this study, especially high-performance computing systems for data analysis.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Plants of the World Online, *Camellia* L. (2025). https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:39017-1 (accessed August 17, 2025).
[2] T.K. Mondal, Camellia, in: C. Kole (Ed.), Wild Crop Relatives: Genomic and Breeding Resources, 1st ed., Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 15–39, doi:10.1007/978-3-642-21201-7_2.
[3] W. Kong, X. Kong, Z. Xia, X. Li, F. Wang, R. Shan, Z. Chen, X. You, Y. Zhao, Y. Hu, S. Zheng, S. Zhong, S. Zhang, Y. Zhang, K. Fang, Y. Wang, H. Liu, Y. Zhang, X. Li, H. Wu, G.-B. Chen, X. Zhang, C. Chen, Genomic analysis of 1,325 *Camellia* accessions sheds light on agronomic and metabolic traits for tea plant improvement, Nat. Genet. 57 (2025) 997–1007, doi:10.1038/s41588-025-02135-z.
[4] N. Hakoda, S. Kirino, *Camellia yokdonensis* Dung bis & Hakoda, Int. Camellia J. 48 (2016) 120.
[5] N.Trung Thanh, L.Van Dung, D.Ngoc Dai, K.Soo Yoong, C.Anthony S, *Camellia maianhii* (Theaceae), a new species of red-flowered camellia from the North Central Coast Region of Vietnam, J. Bot. Res. 6 (2023), doi:10.36959/771/576.
[6] V.D. Luong, H.T. Luu, T.Q.T. Nguyen, Q.D. Nguyen, *Camellia* luteopallida (Theaceae), a new species from Vietnam, Ann. Bot. Fenn. 53 (2016) 135–138, doi:10.5735/085.053.0224.
[7] D.-H. Nguyễn, V.-D. Lương, T.-H. Lê, Q.-T. Trần, N.-Đ. Đỗ, N.-S. Lý, *Camellia puhoatensis* (Sect. Archecamellia – Theaceae), a new species from Vietnam, PhytoKeys 153 (2020) 1–11, doi:10.3897/phytokeys.153.49388.
[8] N.-D. Do, V.-D. Luong, T.-H. Le, D.-H. Nguyen, T.-N. Nguyen, N.-S. Ly, *Camellia ngheanensis* (Sect. Chrysantha Theaceae), a new species from Central Vietnam, Phytotaxa 452 (2020) 209–216, doi:10.11646/phytotaxa.452.3.3.
[9] Y. Tang, X. Zhou, M. Zhu, B. He, C. Jiang, G. Ding, The complete chloroplast genome of *Camellia flava* (Pitard) Sealy, a golden camellia of Vietnam, Mitochondrial DNA Part B 9 (2024) 1117–1121, doi:10.1080/23802359.2024.2392741.
[10] X. Zhou, F. Wang, Y. Xie, J. Ning, Y. Xiao, C. Jiang, G. Ding, Y. Tang, The complete chloroplast genome of *Camellia huulungensis* Rosmann et Ninh, a golden Camellia species endemic to Vietnam, Mitochondrial DNA Part B 9 (2024) 1365–1369, doi:10.1080/23802359.2024.2412227.
[11] M.P. Nguyen, V.T. Ho, Characterization of the complete chloroplast genome of *Camellia tienii* (Theaceae) and its relatives, Crop Breed. Appl. Biotechnol. 24 (2024) e486824415, doi:10.1590/1984-70332024v24n4a52.
[12] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890, doi:10.1093/bioinformatics/bty560.
[13] Y. Ni, J. Li, C. Zhang, C. Liu, Generating sequencing depth and coverage map for organelle genomes v1, (2023). 10.17504/protocols.io.4r3l27jkxg1y/v1.
[14] M. Tillich, P. Lehwark, T. Pellizzer, E.S. Ulbricht-Jones, A. Fischer, R. Bock, S. Greiner, GeSeq – versatile and accurate annotation of organelle genomes, Nucleic Acids Res. 45 (2017) W6–W11, doi:10.1093/nar/gkx391.
[15] S. Greiner, P. Lehwark, R. Bock, OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes, Nucleic Acids Res. 47 (2019) W59–W64, doi:10.1093/nar/gkz238.
[16] R.C. Edgar, Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny, Nat. Commun. 13 (2022) 6968, doi:10.1038/s41467-022-34630-w.
[17] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. Von Haeseler, R. Lanfear, E. Teeling, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37 (2020) 1530–1534, doi:10.1093/molbev/msaa015.
[18] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing, Nat. Methods 9 (2012) 772, doi:10.1038/nmeth.2109.