Molecular Biology Reports (2024) 51:715
https://doi.org/10.1007/s11033-024-09632-0

ORIGINAL ARTICLE

Analysis of population structure and genetic diversity of *Camellia tachangensis* in Guizhou based on SNP markers

Dejun Huang[1] · Suzhen Niu[1,2] · Dingchen Bai[1] · Zhifei Zhao[1] · Caiyun Li[1] · Xiuling Deng[1] · Yihan Wang[1]

Received: 4 December 2023 / Accepted: 10 May 2024 / Published online: 1 June 2024
© The Author(s) 2024

Abstract

Background *Camellia tachangensis* F. C. Zhang is a five-compartment species in the ovary of tea group plants, which represents the original germline of early differentiation of some tea group plants.

Methods and results In this study, we analyzed single-nucleotide polymorphisms (SNPs) at the genome level, constructed a phylogenetic tree, analyzed the genetic diversity, and further investigated the population structure of 100 *C. tachangensis* accessions using the genotyping-by-sequencing (GBS) method. A total of 91,959 high-quality SNPs were obtained. Population structure analysis showed that the 100 *C. tachangensis* accessions clustered into three groups: YQ-1 (Village Group), YQ-2 (Forest Group) and YQ-3 (Transition Group), which was further consistent with the results of phylogenetic analysis and principal component analyses (PCA). In addition, a comparative analysis of the genetic diversity among the three populations (Forest, Village, and Transition Groups) detected the highest genetic diversity in the Transition Group and the highest differentiation between Forest and Village Groups.

Conclusions *C. tachangensis* plants growing in the forest had different genetic backgrounds from those growing in villages. This study provides a basis for the effective protection and utilization of *C. tachangensis* populations and lays a foundation for future *C. tachangensis* breeding.

Keywords Tea · *Camellia tachangensis* · Guizhou Plateau · Population structure · SNP

Introduction

Tea is one of the three major non-alcoholic beverage crops in the world and has important health, economic, and ecological value. The Guizhou Plateau is one of the core origins of tea plants and has a rich tea germplasm due to its distinct geographical location and suitable ecological environment [1]. The germplasm of wild tea plants includes valuable genetic resources for studying their domestication and breeding. The Theaceae family member *Camellia tachangensis* is a wild tea plant that was discovered and named by Zhang Fangci. *C. tachangensis* is only distributed in the border area of Yunnan, Guangxi, and Guizhou, where it is widely distributed [2–4]. *C. tachangensis* reprents a type of tea tree that has not been domesticated in tea gardens and is mostly wildly grown. The majority of *C. tachangensis* plants grow in forests or around their perimeter, and a small part grows in villages or farmlands, although few are used to make tea for drinking. Some are used to protect terraces or as landscape trees or sacrificial trees in front of graves. Recently, with advances in *C. tachangensis* research, its germplasm has been found to containkey nutritional components that meet various human needs. and display excellent comprehensive properties for the development of its tea industry [4]. The significance and potential of the *C. tachangensi* germplasm can be further elucidated by investigating its origin and evolution. Molecular markers allow the detection of variations or polymorphisms in specific regions of DNA among individuals in a population [5]. Examples include restriction fragment length polymorphisms, random amplified polymorphic DNA (RAPD), amplified fragment length polymorphisms, simple sequence repeats (SSRs), and single-nucleotide polymorphisms (SNPs). SNPs, the most common type of genetic variation, have been widely

✉ Suzhen Niu
niusuzhen@163.com; 914348065@163.com

[1] Institute of Tea, Guizhou university, Jiaxiu South Road, Guiyang, Guizhou, China

[2] Institute of Agro-Bioengineering, Guizhou university, Xueshi Road, Guiyang, Guizhou, China

used in research on natural populations to investigate the genetic diversity and population structure of agricultural and horticultural crops such as maize [6], wheat [7], pepper [8], potatoes [9], apples [10], and tea. For example, Pang et al. (2020) identified 327,609 SNPs among 768 wheat cultivars using genotyping-by-sequencing (GBS). Feng et al. (2020) identified 38,395 SNPs among 112 cultivated and wild peppers at a genome-wide level using the GBS method and provided genetic evidence of multiple species splitting events or multiple lineage splitting events. A total of 112,072 SNPs were analyzed by Zhao et al. (2022), who also performed phylogenetic analysis, PCA and population structure analysis were carried out to explore the genetic diversity and geographical distribution characteristics of cultivated tea trees in Guizhou Plateau. The results showed that there were significant differences between the ancient tea cultivars in the Yangtze River basin and the Pearl River Basin [32]. . They further revealed cluster relationships and verified three inferred populations. Niu et al.(2019) [4], analyzed the population structure of 415 tea accessions with 79,016 SNPs and identified four groups: pure wild type, admixed wild type, ancient landraces, and modern landraces.

GBS, a reduced-representation genome sequencing technique, can be used to identify SNPs to perform genotyping studies for germplasm identification and analysis [11], genetic linkage map construction [12], genome-wide association analysis [13], and molecular marker-assisted breeding [14]. For example, Babu et al. (2020) used GBS to conduct a genome-wide association study (GWAS) that identified candidate genes for traits related to yield and oil-yield in palm. Kaur et al. (2021) used GBS to develop a genetic linkage map of *Momordica charantia* L. containing 3,144 SNP markers, 15 linkage groups, and spanning 2,415.2 cM with an average marker distance of 0.7 cM. However, the genetic diversity and population structure analysis for *C. tachangensis* population have not yet been reported. In our previous research [4], *C. tachangensis* was found to cluster into two different groups, although the reason for this phenomenon remained unclear.

In this study, 100 samples of *C. tachangensis* were selected for genetic, phylogenetic and population structure analysis using GBS sequencing technology. 100 samples of *C. tachangensis* were clustered into 3 inferred populations: village group (YQ-1), forest group (YQ-2) and transition group (YQ-3), and the heritage distance of the 3 populations was analyzed. The results showed that the genetic diversity of village group was lowest, that of transition group was highest, and that of forest group was medium. This study provides a theoretical basis for elucidating the population structure of *C. tachangensis* in Guizhou Province and further protecting and utilizing *C. tachangensis*. The present research will lay a foundation for future *C. tachangensis* variety breeding and molecular marker development.

Materials and methods

Plant materials

The study used 100 *C. tachangensis* accessions based on Niu's method [4]. Base on the growth location of *C. tachangensis*, the 100 *C. tachangensis* accessions were divided into two types: village population (32 accessions in village and 6 accessions in farmland) and forest population (27 accessions near forest and 35 accessions in forest). The 100 *C. tachangensis* accessions comprised five accessions from Northwest Guizhou, 35 accessions from Southern Guizhou, 54 accessions from Southwest Guizhou and 6 accessions from Northern Guizhou (Fig. 1 Table S1, S2).

DNA extraction, library construction, and sequencing

We used the Plant genomic DNA Rapid Extraction kit (Biomed Gene Technology) to isolate genomic DNA from samples. DNA integrity was detected on 1% agarose gel, and DNA purity was detected and quantified using Qubit Fluorometer (Invitrogen). In a reaction system of 25μL, we digested 100 NGDNA with 5U SacI and MseI(NEB) and 1 times the restriction enzyme buffer. After digestion, SacAD and MseAD splicers were attached to the digested DNA fragments. Twelve samples were mixed in equal volume and purified using the QIAick PCR purification kit (Qiagen). The purified DNA fragments were then amplified using a mixture of a PCR substrate cocktail and a PCR parent. Amplified fragments of 500–550 bp (including 120 bp aptamers) were obtained by 2% agarose Gel electrophoresis and purified using the QIAick Gel EXTRACTION Kit. The average length of DNA fragments was determined using Agilent DNA 12,000 kit and 2100 BioAnalyzer System(Agilent), and the resulting DNA libraries were quantified using TaqMan probe real-time fluorescent polymerase chain reaction. The paired terminal 150(PE150) sequencing strategy was used on the Illumina HiSeq X Ten platform [4, 15].

Sequence alignment and SNP mining

We have demultiplexed the raw DNA sequence reads based on the barcodes, and trimmed the adapters with a custom Perl script. Any read with a poor quality run of 5 was eliminated. A BWA-MEM (v. 0.7.10)(https://sourceforge.net/projects/bio-bwa/files/) was used to map the reads to the reference tea plant (*Camellia sinensis var. sinensis*)

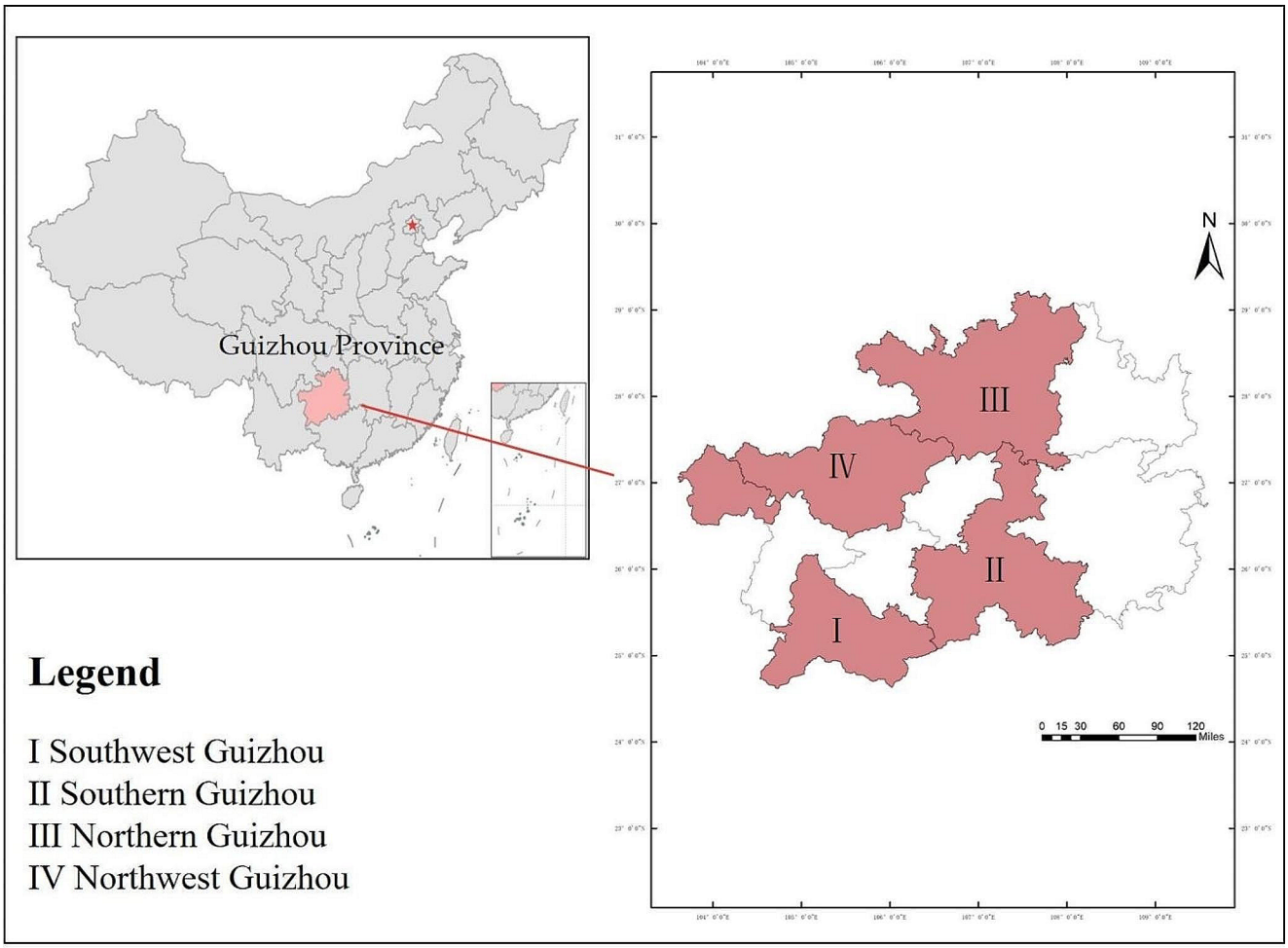


Fig. 1 Geographie distribution of tea accession analyzed in this study according to thecollection

genome [16] (tpdb.shengxin.ren) with its default parameters (Xia et al.2020). According to Niu'study [4]. SNPs were filtered according to several criteria. (1) Biallelic SNPs were selected. (2) "QUAL < 50.0 || QD < 2.0 || FS > 60.0 || MQ < 40.0 || Mapping Quality Rank Sum < -12.5 || Read Pos Rank Sum < -8.0" were used in GATK v. 3.7.0 (https://github.com/broadinstitute/gatk/releases) to filter the SNPs (Mckenna et al.2010). (3) VCFtools (v.0.1.160) (https://github.com/vcftools/vcftools) conserved SNPs with minor allele frequency (MAF) > 0.05 and missing data rates < 20% [17]. The SNP density plot was drawn in CMplot (v.3.7.0) (https://rdrr.io/cran/CMplot/) [18]. The analysis was conducted on 91,959 SNPs from 100 tea accessions.

Population structure and linkage disequilibrium (LD)

We calculated LD (linkage disequilibrium) using the correlation coefficients (r2) for pairwise SNPs across the genome using PopLDdecay. (v. 3.29) (https://github.com/BGI-shenzhen/PopLDdecay) with its default parameters. A VCF file was converted to a ped file with the assistance of VCFtools [17]. Population structure analysis was performed using ADMIXTURE (v1.3.0) (http://dalexander.github.io/admixture/download.html) software and applying the Bayesian model-based clustering method with maximum likelihood estimation. The number of clusters (K value) were assessed by determining cross-validation errors (CV errors) which were tested from 1 to 9, running ten iterations for each K value. The optimal K value was identified according to the lowest CV error [19]. To distinguish between pure and admixture subgroups, the membership coefficient was set at 0.8 [18]. PCA was performed in TASSEL (v. 5.2.72) (https://tassel.bitbucket.io) [20]. In MEGA (v. 10.2.4) (https://www.megasoftware.net/dload_win_gui), a Neighbor-Joining tree was constructed [21]. The tree was visualized and colored using iTOL (https://itol.embl.de/) [22] .

Genetic diversity

The inbreeding coefficient (Fis), observed heterozygosity (Ho) and MAF of each inferred population were calculated using Plink v. 1.90 (https://www.cog-genomics.org/plink2/) [23]. VCFtools was used to determine Tajima's D, nucleotide diversity (Pi), genetic differentiation coefficient (Fst), and Tajima's D for each inferred population. The formula Nm =(1-Fst) / 4Fst was used to compute Nm (gene flow) [23]. The pairwise inferred populations' GD was computed using MEGA. In SPSS version 25, significant variations between these indices were found (IBM Corp., Armonk, NY, USA) [24] .

Results

Mining SNPs data of *C. tachangensis* germplasm from Guizhou Plateau

A total of 29,393,327 SNP markers were obtained through the GBS of 100 *C. tachangensis* accessions from the Guizhou Plateau. After filtering and screening, a total of 91,959 high-quality SNP markers were used for further analysis. Among the 15 chromosomes, Chr1 had the highest number (7,900) of SNPs and Chr15 had the lowest (3,993). The greatest average SNP density was 35.5 SNPs per 1 Mb on Chr1 and the smallest was 30.5 SNPs per 1 Mb on Chr12. (Fig. S1, Table S3) Transitions accounted for 78.38% (AG 39.17%, CT 39.21%) of SNPs, and transversions accounted for 21.62% (AT 6.76%, AC 5.39%, CG 3.89%, GT 5.58%) (Table 1).

Figure 2 The distributions of SNPs on 15 chromosomes. The different color on each chromosome represented the number of SNPs within 1 Mb window size.

Analysis of genetic distance and genetic diversity

Genetic distance is the basic parameter used in genetic diversity research, reflecting the phyletic evolution of groups. In this study, the genetic distance matrix between all 100 *C. tachangensis* accessions was calculated using MEGA software. Among the 4,950 combinations, the genetic distance ranged from 0.053 to 0.512, with an average of 0.26. The lowest genetic distance (0.053) was detected between accessions s394 and s397, both from Tongzi County (Northern Guizhou). The highest genetic distance (0.512) was found between accessions s279 and s348, from the Xingyi (Southwest Guizhou) and Sandu (Southern Guizhou) Counties of Guizhou province, respectively (Table S6). Among the 4,950 genetic distance combinations, 51.76%, 36.73%, and 6.79% were mainly distributed in the range of 0.250–0.350, 0.150–0.250, and 0.350–0.450 cM, respectively. Furthermore, 2.51% of the 4,950 genetic distance combinations were greater than 0.450 cM, and 2.20% were less than 0.150 cM (Fig. S2). In the genetic diversity analysis of *C. tachangensis* in Guizhou, we divided the 100 accessions into forest (62 accessions) and village (38 accessions) populations according to their growing area. The results show that the Tajima's D values of the two populations were positive and the Ho, MAF, and Pi of the village population were 0.08, 0.14, and 0.22, respectively. The Ho, MAF, and Pi of the forest population were 0.09, 0.15 and 0.23, respectively. There was no significant difference in Fis between the forest and village populations (Table 2).

Population structure analysis of *C. tachangensis* accessions in Guizhou Plateau

Changes in the population structure of 100 *C. tachangensis* accessions were further assessed under different K values based on 91,959 SNPs using ADMIXTURE software. Analysis of CV error revealed that K = 2 exhibited the minimum CV error. (Fig. 2a). The population structure at K = 2 was mainly clustered into three subgroups: YQ-1–3 (Fig. 2b). The membership function threshold was set to 0.8 and used to differentiate between pure and mixed subgroups. Thus, the 100 *C. tachangensis* accessions were divided into three subgroups: two pure subgroups (YQ-1 and YQ-2) and one mixed subgroup (YQ-3) (Fig. 2c). To investigate the phylogenetic relationships among the 100 *C. tachangensis* accessions, a phylogenetic tree was constructed based on the 91,959 SNPs using the neighbor-joining (NJ) method in MEGA software. We then used iTOL to visualize and colorize the obtained tree (Fig. 3). The YQ-1 subgroup contained 10 *C. tachangensis* accessions. Among them, 9(90%) grew in villages and 1(10%) in farmland, with all being distributed in Southwestern Guizhou. Thus, YQ-1 was designated the "Village Group".The YQ-2 subgroup contained 50 accessions, composed of 38(76%) that grew in forests and 12(24%) that grew in villages. Of the former, 35(92%) grew in forests and 3(8%) grew near forests.

Table 1 Percentage of transformation and transversion SNPs identified using genotyping-by-sequencing

	Transformation		Transversion			
	AG	CT	AT	AC	CG	GT
Number of sites	266,725	267,010	46,046	36,680	26,492	37,986
Percentage	39.17%	39.21%	6.76%	5.39%	3.89%	5.58%
Total	78.38%		21.62%			

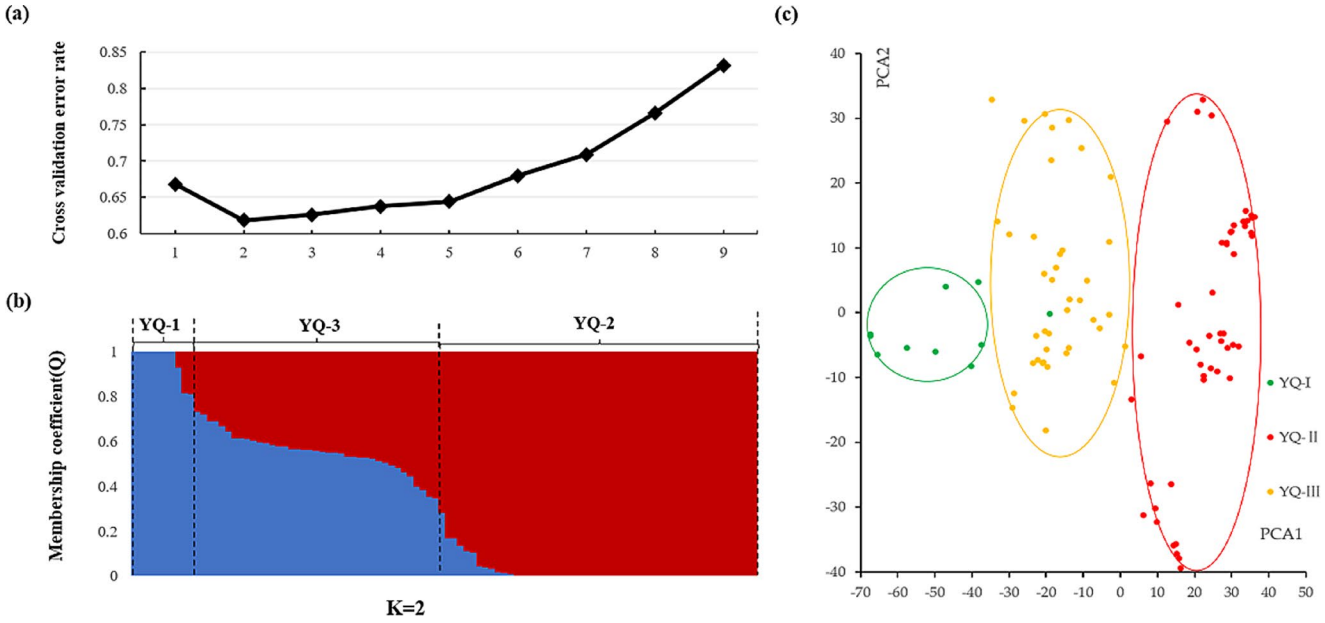


Fig. 2 (**a**) Cross validation error rates corresponding to different k values (**b**) Inferred population structure of 100 accessions. Bar plot of individual membership coefficients for the genetic clusters inferred using ADMIXITURE (K = 2) base on 91,959 SNPs. Individual membership coefficients (Q) were sorted within each cluster. YQ-1 and YQ-2 are shown in blue and red, respectively (**c**) Principal component analysis (PCA). The three PCA scatter diagram was made by the first and second principal components

Among the 50 accessions, 8% were distributed in Northern, 6% in Northwestern, 46% in Southern, and 40% in Southwestern GuizhouTherefore, YQ-2 was designated the "Forest Group".

The YQ-3 subgroup contained 40 accessions, 24(60%) of which were distributed in the Forest Group and 16(40%) of which were distributed in the Village Groups. Among the former, 24(100%) grew near forests. Among the latter, 3(19%) grew in farmland and 13(81%) grew in villages. We regard the YQ-1 pure group from the region where were growing in village and the YQ-2 pure group from the region where were growing in forest, and the YQ-3 mixed group from the region where were growing near forest. Henceforth, YQ-3 was designated the "Transition Group." Among its 40 accessions, 24(60%) were distributed in Southwestern, 30% in Southern, 5% in Northwestern, and 5% in Northern Guizhou.

Analysis of genetic diversity and genetic composition of inferred population

The genetic diversity of three inferred populations was analyzed by Plink and VCFtools. The results show positive Tajima's D values for the three inferred populations. The Ho, MAF, and Pi values of inferred population YQ-1 were 0.05, 0.11, and 0.16, respectively, being significantly lower than those of YQ-2, YQ-3, and the total population. The Ho, MAF, and Pi values of YQ-2 were 0.08, 0.14, and 0.20 (Table 2), representing significantly lower values than those of the total population and YQ-3. Genetic diversity parameters in YQ-3 were not significantly different from those in the total population. The three inferred populations did not differ significantly in terms of Fis.

The Fst and genetic distance among the three inferred populations were analyzed by VCFtools and MEGA. The results revealed significant differences in Fst among the three inferred populations. The Fst and genetic distance between YQ-1 and YQ-2 were the largest, followed by those between YQ-2 and YQ-3, while those between YQ-1 and YQ-3 were the lowest. The Nm between YQ-1 and YQ-3 was the largest, while that between YQ-1 and YQ-2 was the smallest (Table S7, Fig. 4).

LD analysis

LD analysis was used to explain domestication and mating history [6, 25]. In a population of 100 accessions, we used 29,393,327 non-LD-pruned SNPs to determine the level of LD. The LD decreased rapidly as genetic distance increased or with increasing distance between SNPs. The highest r^2 values for LD decline among all 100 accessions were 0.25. As r^2 decayed to half its maximum value (0.125), the equivalent physical distance was 0.2 Kb (Fig. 5, Table S8).

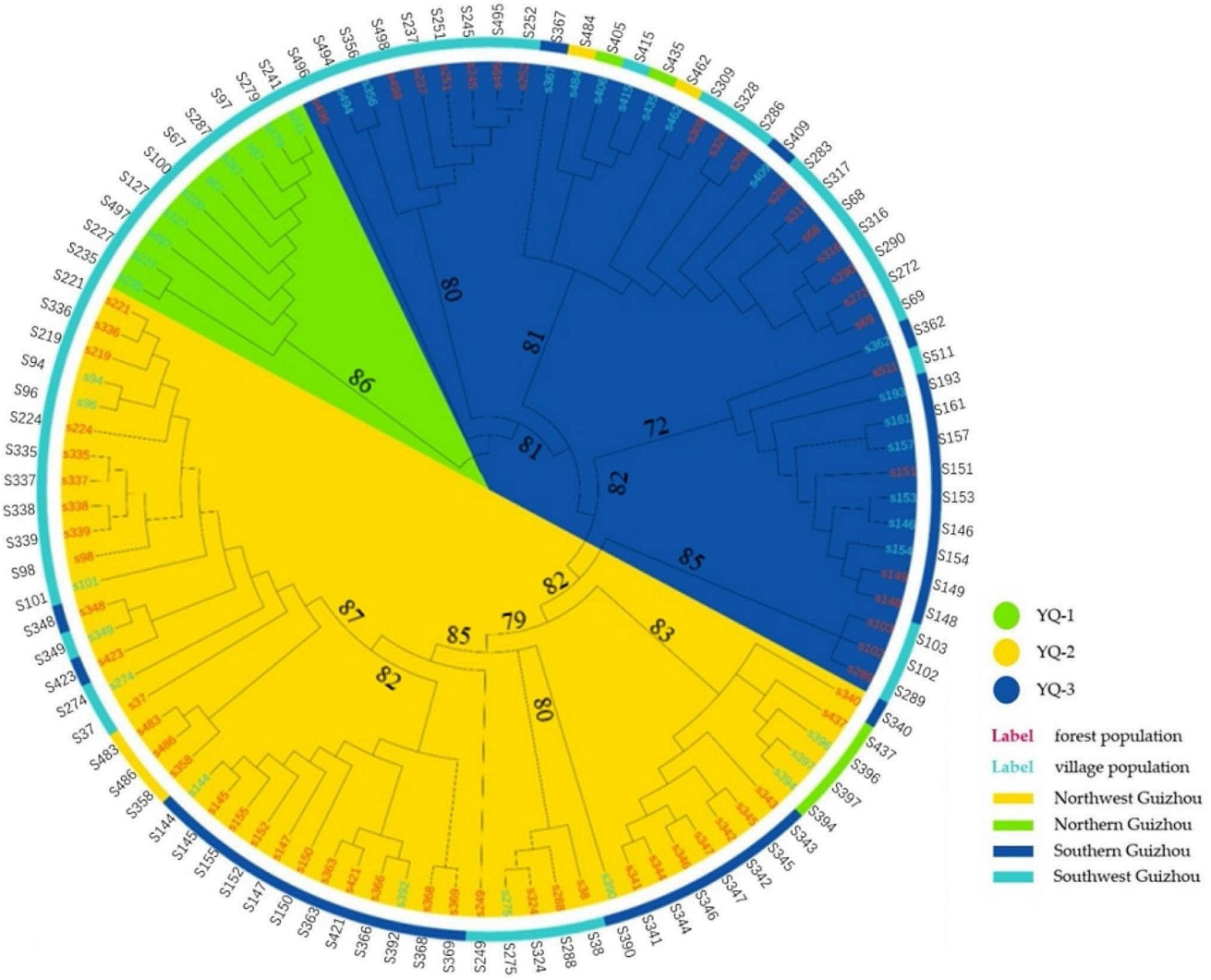


Fig. 3 Phylogenetic analysis of 100 *C.tachangensis* accessions in Guizhou Plateau. Unroot Neighbor-Joining (NJ) phylogenetic tree was constructed using MEGA software with default parameter and used iTOL to visualize and color the tree. The different colored sector represents the different inferred populations, the leaf with different color represents the different population, and outer ring strip represents the different position

Table 2 Genetic diversity analysis of *Camellia tachangensis* population

Group	Number	Ho	MAF	Pi	Fis	Tajima's D
All	100	0.09a	0.15a	0.23a	0.65a	0.57
Forest	62	0.09a	0.15a	0.23a	0.62a	0.47
Village	38	0.08b	0.14b	0.22b	0.66a	0.46
YQ-1	10	0.05c	0.11c	0.16c	0.70a	0.55
YQ-2	50	0.08b	0.14b	0.20b	0.58a	0.58
YQ-3	40	0.09a	0.15a	0.22a	0.62a	0.47

Note Ho observed heterozygosity; *MAF* Minor Allele Frequency; *Pi* nucleotide diversity; *Fis* Inbreeding coefficient, significant difference analysis of abc representation

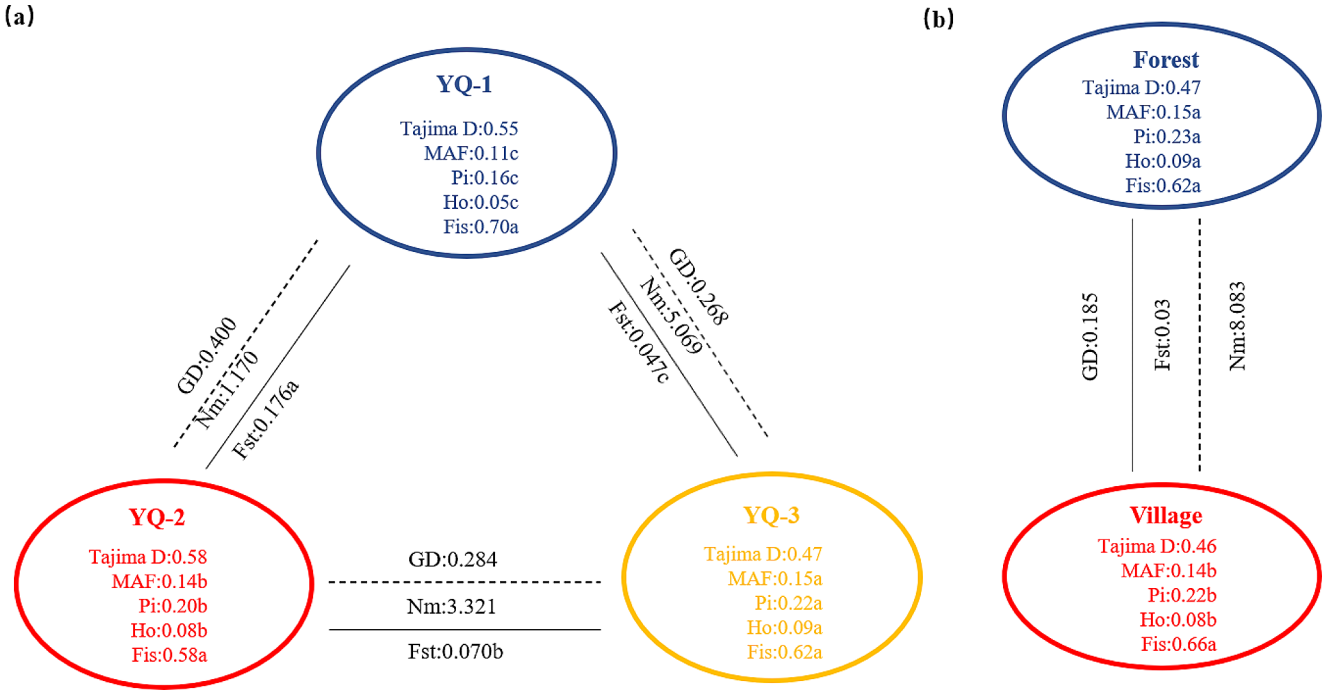


Fig. 4 Genetic diversity of three inferred populations of 100 accessions. Pi nucleotide diversity, Ho observed heterozygosity, MAF minor allele frequency, Fis inbreeding coefficient, GD genetic distance, Fst differentiation coefficient. The different letters indicate a significant difference in $p = 0.05$ levels by the T-test. (**a**) YQ-1, YQ-2 are pure subgroups and the YQ-3 is the admixture subgroup base on ADMIXTURE software at K = 2 (**b**) forest population and village population base on accessions growth place

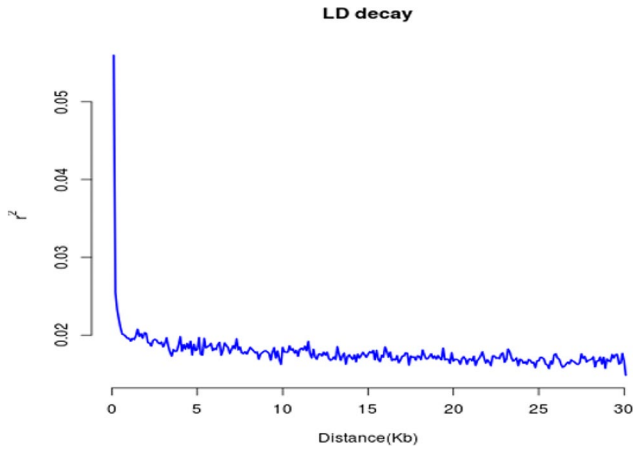


Fig. 5 The Linkage disequilibrium decay plot of 100 *C. tachangensis* within 30KB. the x-axis represents the distance in KB and the y-axis represents the r^2

Discussion

Genetic distance and genetic diversity

Previous studies have demonstrated that the GBS method is effective for the genetic diversity analysis of tea plants [26, 27]. This study was the first to analyze the evolutionary relationships and genetic diversity of the *C. tachangensis* population. A total of 91,959 SNPs from 100 *C. tachangensis* accessions were analyzed using the GBS method. Genetic distance analysis was used to determine the genetic difference between a pair of specific accessions or populations, and usually shares many alleles with low genetic distance [28]. In the present study, the lowest genetic distance was beteween accessions from Northern Guizhou, belonging to YQ-2. The largest genetic distance was between accessions from Southwestern and Southern Guizhou, belonging to YQ-1 and YQ-2, respectively. The 10 pairs of accessions with minimum genetic distance were from the same inferred population and the same position. In 10 pairs of accessions with the larger pairwise genetic distance, there were 9 pairs of accessions from inferred population YQ-1 and YQ-2 respectively, and 2 pairs of accessions with the largest pairwise genetic distance from the different location, Southern and Southwest Guizhou respectively, which might be because *C. tachangensis* displays substantial genetic variation in the process of spreading to new areas.

The positive Tajima's D values for all inferred populations suggested that they all underwent population bottlenecks and/or balancing selection. Moreover, the Ho, MAF, and Pi of the Forest Group were significantly higher than those of the Village Group. This indicates that the Village Group has lower genetic diversity due to the destruction of the surrounding natural environment and human selection, and that the Forest Group has not underwent excessive artificial selection and has retained high genetic diversity. The

results of Zhao et al. (2014), who found that wild populations display greater diversity than recently domesticated populations, are similar to ours [29].

Population structure

With the continuous development of high-throughput sequencing, the mining of SNP molecular markers through sequencing has been widely applied to tea plant kinship identification, population genetic diversity analyses [30], and genetic variation research [31]. Zhao et al. (2022) selected a total of 112,072 SNPs from the 253 tea accessions and performed a GBS analysis [32]. In our study, GBS technology was used to analyze the SNPs of 100 *C. tachangensis* accessions from the Guizhou Plateau. Three different methods were used to analyze their population structure based on the 91,959 high-quality SNPs, which further verified the accuracy of the population structure. Population genetic structure refers to a nonrandom distribution of genetic variation in species or populations [33, 34]. Through population structure analysis, the classification of relationships between individuals can be attained, populations can be divided into several inferred populations, and the occurrence of gene exchange between populations and the degree of hybridization of each individual can be elucidated [25, 33]. In this study, 100 *C. tachangensis* accessions were divided into three inferred populations according to their growth location: the Village Group (YQ-1), the Forest Group (YQ-2), and the Transition Group (YQ-3).

In this research, the genetic distance among the three inferred populations was consistent with the results of Fst and Nm. We observed the highest genetic differentiation and genetic distance and the lowest gene flow between Village and Forest Groups, the lowest genetic differentiation and genetic distance and the highest gene flow between Village and Transition Groups, and moderate genetic differentiation, genetic distance, and gene flow between Forest and Transition Groups. The lowest genetic diversity was detected in the Village Group, the highest was found in the Transition Group, and a moderate genetic diversity was detected in the Forest Group. These results support the notion that *C. tachangensis* evolution was related to human activity in the Guizhou Plateau. The Forest Group featured the most ancestral accessions originating in the forest without human intervention until present, and its retained group purity was owing to the limited pollen transmission distance in the forest, leading to plants that can only cross or self-cross within a small range. The Transition Group accessions were likely preserved because this plant is evergreen, grows slowly, and has no use value in the process of human long-term use of forests, deforestation and invasion of forests. Without the hindrance posed by forests, pollen among tea trees spreads farther, and tea trees with large genetic background differences were more likely to cross [35], which significantly increased the diversity of the Transition Group among all inferred groups. The Village Group underwent artificial selection because in our sampling process, we found that its accessions were located at the edge of terraces inside the village, playing the role of a retaining wall, or planted in front of villagers' houses and used for brewing and drinking. In terms of evolutionary direction, the Forest Group was inclined to display high survival and the Village Group was inclined to produce higher yields and improved secondary metabolites for human consumption.The growth environments of the Village Group and Forest Group were separated by the Transition Group, which makes it difficult for them to have gene exchange. The reason for frequent gene exchange between the Village and Transition Groups might be that in recent years, with the expansion of human villages, the growth environment of the Transition Group was closer to that of the Village Group than that of the Forest Group. It might that village expanding into forest led the gene exchange between Forest Group and Village Group with increase of population and formed greater genetic diversity Transition Group. Overall, human activities have played an important role in the evolution of wild-type *C. tachangensis*. Although these accessions were not used for industrialization, a differentiation between *C. tachangensis* growing in villages and forests exists. Future research could focus on the protection of the Forest and Village Groups and the utilization of the Transition Group. The expansion of villages into forests may have led to gene exchange between Forest and Village Groups, leading to the formation of a Transition Group with a larger population and greater genetic diversity.

Conclusions

In this study, 100 *C. tachangensis* accessions were collected, and SNP molecular data was used to perform a population structure analysis, which clustered them into three inferred populations: the Village Group (YQ-1), the Forest Group (YQ-2), and the Transition Group (YQ-3). The highest genetic differentiation and genetic distance and the lowest gene flow was between Village and Forest Groups, the lowest genetic differentiation and genetic distance and the highest gene flow was between Village and Transition Groups, and there was moderate genetic differentiation, genetic distance, and moderate gene flow between Forest and Transition Groups. The growth environment of the Village and Forest Groups was separated by the Transition Group, which decreased their gene exchange. The reason for the frequent gene exchange between the Village and Transition

Groups might be that in recent years, with the expansion of human villages, the growth environment of the Transition Group was closer to that of the Village Group than that of the Forests Group. Additionally, the lowest genetic diversity was found in the Village Group, the highest genetic diversity was observed in the Transition Group, and a moderate genetic diversity was observed in the Forest Group. Without the hindrance of forests, pollen among tea trees spread farther and tea trees with large genetic background differences were more likely to cross, which significantly increased the diversity of the Transition Group among all inferred groups. These results support the notion that *C. tachangensis* evolution was related to human activity in the Guizhou Plateau. Future research could focus on the protection of Forest and Village Groups and the utilization of the Transition Group. The latter could play an important role in the future breeding of new varieties, and human activities will affect the genetic diversity of the *C. tachangensis* population.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11033-024-09632-0.

Acknowledgements We thank College of tea science of Guizhou University for providing research facilities and computing facilities. We thank B. X for his guidance and software suggestions in data processing.

Author contributions Conceptualization, SZ N. and DJ H.; methodology SZ N. and DJ H.; software, DC B.; validation, XL D.; formal analysis, SZ N.; investigation, ZF Z and CY Li; resources, SZ N.; data curation, YH W.; writing—original draft preparation, Dejun Huang.; writing—review and editing, Suzhen Niu.; visualization, Suzhen Niu.; supervision, Suzhen Niu.; project administration, Suzhen Niu.; funding acquisition, Suzhen Niu. All authors have read and agreed to the published version of the manuscript.

Funding This work funded by Project of the National key R & D plan (2021YFD1200203-1), Projectofthe National Science Foundation, in PR China· (32060700) for design of the study, the National Guidance Foundation for Local Science and Technology Development of China (2023009), Guiyang Science and Technology Plan Project (Construction Technology Contract [2023] ·48 – 21). Project of the key filed project of Natural Science Foundation of Guizhou Provincial Department of education (KY [20211·042) for data analysis Science and Technology Plan Project of Guizhou province, in PR China ([2021] General 126) for design of the study.

Data availability The raw sequence data reported in this study have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA001438 that is publicly accessible at Genome Sequence Archive (cncb.ac.cn).

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Fang W, Cheng H, Duan Y, Jiang X, Li X (2012) Genetic diversity and relationship of clonal tea (Camellia sinensis) cultivars in China as revealed by SSR markers. Plant Syst Evol 298:469–483. https://doi.org/10.1007/s00606-011-0559-3
2. Chen L, Yu F, Tong Q (2000) Discussions on phylogenetic classification and evolution of section Thea. J Tea Sci 89–94. https://doi.org/10.13305/j.cnki.jts.2000.02.003
3. Zhang F (1980) Two new species of Camellia from Yunnan. Yunnan Bot Res 103–106
4. Niu S, Song Q, Koiwa H et al (2019) Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (Camellia sinensis) from an origin center, Guizhou plateau, using genome-wide SNPs developed by genotyping-by-sequencing. BMC Plant Biol 19(1):1–12. https://doi.org/10.1186/s12870-019-1917-5
5. Wachira FN (2017) Molecular markers - New tools for an old science - The case of tea
6. Khadka K, Torkamaneh D, Kaviani M, Belzile F, Raizada MN, Navabi A (2020) Population structure of Nepali spring wheat (Triticum aestivum L.) germplasm. BMC Plant Biol 20(1):1–12. https://doi.org/10.1186/s12870-020-02722-8
7. Pang Y, Liu C, Wang D et al (2020) High-resolution genome-wide association study identifies genomic regions and candidate genes for important agronomic traits in wheat. Mol Plant 13(9):1311–1327. https://doi.org/10.1016/j.molp.2020.07.008
8. Feng S, Liu Z, Hu Y, Tian J, Yang T, Wei A (2020) Genomic analysis reveals the genetic diversity, population structure, evolutionary history and relationships of Chinese pepper. c 7:158. https://doi.org/10.1038/s41438-020-00376-z
9. Yamakawa H, Haque E, Tanaka M et al (2021) Polyploid QTL: eq towards rapid development of tightly linked DNA markers for potato and sweet potato breeding through whole genome resequencing. Plant Biotechnol J 19:2040–2051. https://doi.org/10.1111/pbi.13633
10. Luo F, Evans, Norelli JL et al (2020) Prospects for achieving durable disease resistance with elite fruit quality in apple breeding. Tree Genet Genomes 16. https://doi.org/10.1007/s11295-020-1414-x
11. Lu L, Chen H, Wang X et al (2021) Genome-level diversification of eight ancient tea populations in the Guizhou and Yunnan regions identifies candidate genes for core agronomic traits. Hortic Res (English) 8:10. https://doi.org/10.1038/S41438-021-00617-9

12. Kaur G, Pathak M, Singla D et al (2021) High-density GBS-Based genetic linkage Map Construction and QTL Identification Associated with Yellow Mosaic Disease Resistance in bitter Gourd (Momordica charantia L). Front Plant Sci 12:671620. https://doi.org/10.3389/FPLS.2021.671620
13. Babu BK, Mathur RK, Ravichandran G et al (2020) Genome wide association study (GWAS) and identification of candidate genes for yield and oil yield related traits in oil palm (Eleaeis guineensis) using SNPs by genotyping-based sequencing. Genomics 112:1011–1020. https://doi.org/10.1016/j.ygeno.2019.06.018
14. Chu (2021) A new PCR/LDR-Based multiplex functional molecular marker for marker-assisted breeding in Rice. Rice Sci 28. https://doi.org/10.1016/J.RSCI.2020.11.002
15. Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379. https://doi.org/10.1371/journal.pone.0019379
16. Xia E, Tong W, Hou Y et al (2020) The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. Mol Plant 13(7):1013–1026. https://doi.org/10.1016/j.molp.2020.04.010
17. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al (2011) The variant call format and VCFtools. Bioinf (Oxford England) 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330
18. Delfini J, Moda-Cirino V, Dos Santos Neto J et al (2021) Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. Sci Rep 11(1):2964. https://doi.org/10.1038/s41598-021-82437-4
19. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–1664. https://doi.org/10.1101/gr.094052.109
20. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635. https://doi.org/10.1093/bioinformatics/btm308
21. Stecher G, Tamura K, Kumar S (2020) Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol 37(4):1237–1239. https://doi.org/10.1093/molbev/msz312
22. Slifer SH (2018) Curr protocols Hum Genet 97(1):e59. https://doi.org/10.1002/cphg.59. PLINK: key functions for data analysis
23. Cheng J, Kao H, Dong S (2020) Population genetic structure and gene flow of rare and endangered tetraena mongolica Maxim. Revealed by reduced representation sequencing. BMC Plant Biol 20(1):1–13. https://doi.org/10.1186/s12870-020-02594-y
24. Evans BA, Rozen DE (2013) Significant variation in transformation frequency in Streptococcus pneumoniae. ISME J 7(4):791–799. https://doi.org/10.1038/ismej.2012.170
25. Zhang H, Zhang J, Xu Q et al (2020) Identification of candidate tolerance genes to low-temperature during maize germination by GWAS and RNA-seq approaches. BMC Plant Biol 20:1–17. https://doi.org/10.1186/s12870-020-02543-9
26. Li C, Ge Y, Yan M et al (2021) Genetic diversity and population structure analysis of 32 tea local population species resources. Tea Sci 41(5):12
27. Pascual L, Ruiz M, López-Fernández M et al (2020) Genomic analysis of Spanish wheat landraces reveals their variability and potential for breeding. BMC Genomics 21:1–17
28. Ndjiondjop MN, Semagn K, Gouda AC et al (2017) Genetic variation and population structure of Oryza glaberrima and development of a mini-core collection using DArTseq. Front Plant Sci 8:1748. https://doi.org/10.3389/fpls.2017.01748
29. Zhao DW, Yang JB, Yang SX, Kato K, Luo JP (2014) Genetic diversity and domestication origin of tea plant Camellia taliensis (Theaceae) as revealed by microsatellite markers. BMC Plant Biol 14:1–12. https://doi.org/10.1186/1471-2229-14-14
30. Jiang Y, Zhang C, Cheng H et al (2016) Study on SSR identification of different varieties in tea seed farm. J tea 57(03):105–112. https://doi.org/10.3969/j.issn.1007-4872.2016.03.001
31. Li H, Yang D, Wang L et al (2020) ISSR Analysis of Genetic Diversity of Tea Germplasm resources in Dongshan. Suzhou Subtropical Plant Sci 49(3):5. https://doi.org/10.3969/j.issn.1009-7791.2020.03.001
32. Zhao Z, Song Q, Bai D et al (2022) Population structure analysis to explore genetic diversity and geographical distribution characteristics of cultivated-type tea plant in Guizhou Plateau. BMC Plant Biol 22:1–14. https://doi.org/10.1186/S12870-022-03438-7
33. Luo L, Huang Y, Zeng Z, Zhou M, Xie M, Yan C (2020) Distant hybridization of tea plant F_ study on genetic variation of leaf phenotype of the first generation. Tea Communication 47(4):8
34. Gopalan P, Hao W, Blei DM, Storey JD (2016) Scaling probabilistic models of genetic variation to millions of humans. Nat Genet 48(12):1587–1590. https://doi.org/10.1038/ng.3710
35. Wang Y, Jiang CJ, Zhang HY (2008) Observation on the self-incompatibility of pollen tubes in self-pollination of tea plant in style in vivo. J Tea Sci 28(6):429–435. https://doi.org/10.13305/j.cnki.jts.2008.06.008

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