RESEARCH



Genome-wide association study, population structure, and genetic diversity of the tea plant in Guizhou Plateau

Yihan Wang¹, Suzhen Niu^{2*}, Xinyue Deng³, Dingchen Bai¹, Zhengwu Chen^{4*}, Xiuling Deng¹ and Dejun Huang¹

Abstract

Background Guizhou Plateau, as one of the original centers of tea plant, has a profound multi-ethnic cultural heritage and abundant tea germplasm resources. However, the impact of indigenous community factors on the genetic diversity, population structure and geographical distribution of tea plant is still unclear.

Results Using the genotyping-by-sequencing (GBS) approach, we collected 415 tea plant accessions from the study sites, estimated genetic diversity, developed a core collection, and conducted a genome-wide association study (GWAS) based on 99,363 high-quality single-nucleotide polymorphisms (SNPs). A total of 415 tea accessions were clustered into six populations (GP01, GP02, GP03, GP04, GP05 and GP06), and the results showed that GP04 and GP05 had the highest and lowest genetic diversity (Pi=0.214 and Pi=0.145, respectively). Moreover, 136 tea accessions (33%) were selected to construct the core set that can represent the genetic diversity of the whole collection. By analyzing seven significant SNP markers associated with the traits such as the germination period of one bud and two leaves (OTL) and the germination period of one bud and three leaves (OtL), four candidate genes possibly related to OTL and OtL were identified.

Conclusions This study revealed the impact of indigenous communities on the population structure of 415 tea accessions, indicating the importance of cultural practices for protection and utilization of tea plant genetic resources. Four potential candidate genes associated with the OTL and OtL of tea plant were also identified, which will facilitate genetic research, germplasm conservation, and breeding.

Keywords Genotyping-by-sequencing, Tea plant, Population structure, Genetic diversity, Genome-wide association study, Guizhou plateau

*Correspondence: Suzhen Niu niusuzhen@163.com Zhengwu Chen zwchentea@163.com ¹College of Tea Science, Guizhou University, Guiyang, Guizhou Province 550025, China ²The Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region (Ministry of Education), Institute of Agro-Bioengineering, Guizhou University, Guiyang, Guizhou Province 550025, China ³School of Architecture, Guizhou university, Guiyang, Guizhou Province 550025, China ⁴Institute of Tea, Guizhou Academy of Agricultural Sciences, Guiyang, Guizhou Province 550006, China



© The Author(s) 2024. **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, 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/. The Creative Commons Dublic Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Biodiversity is the ecological complex formed by organisms, environmental factors and various related ecological processes [1, 2]. It is explored at three levels or types, including genetic diversity, species diversity and ecosystem diversity [3]. Biodiversity revitalizes the earth by supporting and sustaining healthy ecosystems, and lays the foundation for human survival and development [4, 5]. However, over-exploitation of resources, social transformation, climate change and pollution are posing serious threats to biodiversity [6–8]. Therefore, it is necessary to effectively protect biodiversity to maintain the ecosystem processes and ultimately sustain human life [9, 10].

The traditional culture of indigenous people and ecological knowledge of smallholder farmers contribute to biodiversity conservation and inhibit the rapid loss of biodiversity around the world [11, 12]. Previous studies have demonstrated that traditional management practices, customs and economic exchanges, including religious or ancestral worship ceremonies [13], dietary practices [14, 15], seed exchange systems [16] and marriage rituals [17] play positive roles in biodiversity conservation. Moreover, these practices and customs have also been related to preserving landraces and old trees with medicinal, nutritional and aesthetic values [18–20].

Tea (Camellia sinensis (L.) O. Kuntze), as one of the three most popular non-alcoholic beverages in the world, contains numerous secondary metabolites such as caffeine, amino acids, phenolics, terpenes and volatiles, which contribute to economic values, pleasant flavor and health benefits of tea [21]. It is known that the history of tea drinking in China can be traced back to the beginning of the 5th century AD until the end of the 6th century AD, and Chinese people consume tea for different medicinal and edible purposes and also as mixed or pure drinks [22]. Tea plant is an economic crop with tender buds and leaves in the lateral branches as its product organs [23]. The germination period of one bud and two leaves (OTL) and the germination period of one bud and three leaves (OtL) are the major agronomic traits of tea plant. Due to the relatively low temperature in early spring, early OTL and OtL tea plants can avoid a large amount of pests and diseases during the first round of harvesting, accumulating numerous natural compounds that can improve the quality of tea [23]. Early spring tea has higher economic value, therefore, breeding early OTL and OtL varieties is a valuable goal [24].

Tea plants are believed to be originated in southwestern China, which includes Guizhou Plateau and the adjacent provinces [25]. The Guizhou Plateau is a multiethnic area and home to six groups of indigenous people such as Miao (Miao Man, MM) and Gelao (Pu Ren, PR), who have the longest history of living in Guizhou Plateau, as well as Tujia and Yi (Di Qiang, DQ), Buyi (Bai Yue, BY) and Han (Han Ren, HR) [26]. Guizhou Plateau has a superior geographical location, profound cultural heritage and rich biodiversity. This region has preserved abundant tea germplasm resources, with wild type accessions, ancient landraces and modern landraces [27]. However, with the degradation of habitat, the reduction of genetic diversity, and the trend of monoculture farming [28, 29], it is increasingly important to understand the current population structure and genetic diversity of tea plant in the Guizhou Plateau. Although many studies have confirmed the importance of traditional agricultural management knowledge for biodiversity conservation [30], the complex interaction between the customs and the migration of indigenous people and the population structure of tea accessions has not been studied.

In this study, we integrated ethnobotanical and molecular genetic methods to explore the impact of migration, customs and economic exchanges of indigenous communities in the Guizhou Plateau on the population structure, geographical distribution of tea plant, and the protection of genetic diversity. We collected 415 tea germplasm resources from 33 regions in Guizhou Plateau, as well as Zhejiang, Hunan and Fujian Provinces. Based on the genotyping-by-sequencing (GBS) approach, the genetic diversity, population structure and linkage disequilibrium (LD) of 415 tea accessions were studied using large numbers of single-nucleotide polymorphisms (SNPs). Finally, we constructed a core collection of all tea accessions and conducted a genome-wide association study (GWAS) to identify candidate genes that control OTL and OtL of tea plant.

Results

Sequencing and variant discovery

A collection of 415 tea plants, including 159 wild type accessions, 174 ancient landraces and 82 modern landraces were used for this study (Additional file 1: Table S1). The geographical distribution of 410 accessions in the indigenous communities of Guizhou Plateau was shown in Fig. 1. The other five accessions were introduced from other provinces and cultivated in the tea garden of Guizhou Plateau. A GBS analysis was performed on all 415 tea accessions using Illumina HiSeq X Ten platform. After filtering low-quality reads, a total of 390.30 Gb clean data were retained, with an average of 0.94 Gb per accession (Additional file 2: Table S1). We mapped the clean reads to a tea reference genome sequence [31], a total of 29,393,327 initial SNPs were obtained from the 415 tea accessions using GATK (v. 3.7.0) software, and 99,363 high-quality SNPs were obtained after filtering, which ensured the reliability and accuracy of subsequent population structure and genetic diversity analyses. The average heterozygosity rate and the average missing rate for per accession were 6.98% and 13.81%, respectively



Fig. 1 Geographic distribution of the sampling locations. The different colored hexagonal stars represent the geographic distribution of sampling locations of five indigenous communities in Guizhou Plateau [26], including Han (HR) indigenous communities, Gelao (PR) indigenous communities, Tujia and Yi (DQ) indigenous communities, Buyi (BY) indigenous communities and Miao (MM) indigenous communities. The geographic data information used in the map came from field investigations and geographic information surveyed by the Natural Resources Department of Guizhou Province (https://zrzy.guizhou.gov.cn/wzgb/zwgk/zdlyxxgk/dlxxgl/), and the map was drawn using ArcGIS software

(Additional file 2: Table S2). The distribution of SNPs across chromosomes reveals the genomic landscape of SNPs in the tea population. The SNPs were unevenly distributed on 15 chromosomes, and the lowest and highest SNP densities were detected on chromosomes 5 and 1, respectively (Additional file 5: Fig. S4). At the genomic level, we found that more than half of the SNPs (78.05%) had base transitions, and the transition vs. transversion (ts/tv) ratio was 3.56. The C/G transversion and the C/T transition occurred at the lowest and highest frequencies, respectively. The frequencies of the C/G, G/T, A/T and A/C transversions were 4.10%, 5.60%, 6.79% and 5.46%, respectively, and the frequencies of two-type transitions were similar (39.39% for C/T and 38.66% for A/G) (Additional file 2: Table S3).

Genetic diversity analysis

The nucleotide diversity (Pi), minor allele frequency (MAF), observed heterozygosity (Ho) and inbreeding coefficient (Fis) are used as indicators of genetic diversity, which can evaluate the genetic structure and diversity of the population [32, 33]. Pi, MAF, Ho and Fis of 415 accessions were 0.226, 0.144, 0.073 and 0.691, respectively (Table 1). Further analysis of the genetic diversity of tea plant populations from three cultivation status showed that Pi, Ho, MAF and Fis of the wild type accessions

(WA) population were significantly higher than those of ancient landraces (AL) population and modern landraces (ML) population. Pi and MAF for the tea plant population in AL were significantly higher than those for the tea plant population in ML. We compared the genetic diversity of tea plant populations from MM, PR, BY, DQ and HR indigenous communities. Pi, Ho and MAF of the tea plant populations from MM and PR exhibited significantly higher compared with those of tea plant populations from BY, DO and HR. Pi, Ho and MAF were higher for the BY population compared with the DQ population. Analysis of the genetic diversity of tea plant populations from four different species revealed that Pi of the C. gymnogyna population was the highest, while Pi of the C. sinensis population was the lowest, and Fis of C. tachangensis population was significantly higher compared with that of C. sinensis, C. gymnogyna and near C. taliensis populations. Ho of the near C. taliensis population was significantly higher than that of *C. gymnogyna*, C. tachangensis and C. sinensis populations (Additional file 2: Table S4).

Tajima's D represents intraspecific polymorphism based on locus variation. The positive Tajima's D values of the evaluated tea populations indicated that they all underwent balancing selection or a bottleneck, and populations suddenly shrank [34] (Table 1 and Additional file

Туре		Number of accessions				Tajima's D	Pi	Но	MAF	Fis
		WA	AL	ML	Total					
Indigenous	MM	67	17	23	107	1.098	0.247a	0.074b	0.169a	0.714a
communities	PR	42	12	1	55	0.722	0.234b	0.092a	0.157b	0.617b
	BY	23	52	14	89	0.503	0.210c	0.071c	0.134c	0.676ab
	DQ	24	77	38	139	0.521	0.199d	0.070d	0.127d	0.657b
	HR	3	16	1	20	0.058	0.181e	0.053e	0.117e	0.706ab
Cultivation status	WA	-	-	-	159	1.240	0.246a	0.081a	0.170a	0.680a
	AL	-	-	-	174	0.529	0.181b	0.068b	0.118b	0.638b
	ML	-	-	-	82	0.325	0.167c	0.069b	0.109c	0.599b
All	All	159	174	82	415	1.326	0.226	0.073	0.144	0.691

Table 1 Genetic diversity of 415 tea accessions in Guizhou Plateau

Note nucleotide diversity (P_i), observed heterozygosity (H_o), minor allele frequency (MAF) and inbreeding coefficient (F_i s). In the same type and line, the different letters indicate a significant difference in p=0.05 levels by the T-test. Cultivation status includes wild type accessions (WA), ancient landraces (AL) and modern landraces (ML). Five indigenous communities include HR, DQ, BY, MM and PR indigenous communities



Fig. 2 Pairwise *Fst* and genetic distance among five indigenous communities of 415 accessions in Guizhou Plateau. (**A**) Inter-population genetic differentiation coefficient (*Fst*). The colors and numbers in the cells of the matrix represent the *Fst* values and the different letters indicate a significant difference in p = 0.05 levels by the T-test. (**B**) Pairwise genetic distance (*GD*), the lines in different colors represent the value of genetic distance, with red representing the shortest distance for the others, green representing the longest, other colors with darker colors representing the shorter distance, and lighter colors representing the longer distance. Source data underlying (**B**) is provided as an additional file: 2 Table S5. (**A and B**) HR: Han indigenous communities, DQ: Tujia and Yi indigenous communities, BY: Buyi indigenous communities, MM: Miao indigenous communities, and PR: Gelao indigenous communities

2: Table S4). The genetic differentiation coefficient (*Fst*) is used as a measure of population structure, and Fst of 0-0.05, 0.05-0.15, 0.15-0.25 and above 0.25 indicate small, moderate, large and great divergence, respectively [33]. In this study, Fst of indigenous community populations ranged from 0.006 to 0.068, with an average value of 0.041. The maximum Fst value (0.068) was observed in DQ vs. MM, indicating that there was moderate divergence between these two populations. In addition, the minimum Fst value (0.006) was observed in DQ vs. HR, indicating that there was small divergence between these two populations (Fig. 2A). Notably, the lowest Nm value was recorded in DQ vs. MM, whereas the highest Nm value was showed in DQ vs. HR. At a geographical level, the lowest pairwise genetic distance (GD) was for DQ vs. HR (Fig. 2B and Additional file 2: Table S5).

Population structure analysis

We revealed the ancestral genetic components of the 415 tea accessions by different methods and inferred the population structure. This study assumed ancestral populations (K) ranging from 2 to 9. The optimal ancestral population (K) value was determined according to the cross-validation (CV) error, and the optimal value obtained in this study was K=5 (Additional file 5: Fig. S5). Based on CV error, five ancestral genetic components were detected (Additional file 1: Table S2). Accessions with membership coefficients>0.60 were assigned to the pure population, while those with membership coefficients < 0.60 were assigned to the admixture population [33]. The first pure population (referred to as the ancient landraces of DQ or GP01 hereafter) was composed of 131 (74.9%) ancient landraces, 26 (14.9%) modern landraces, and 18 (10.3%) wild type accessions, of which 154 (88.0%) belonged to C. sinensis, mainly from the DQ indigenous

community. The second pure population (referred to as the wild type accessions of MM or GP02 hereafter) was composed of 51 wild-type C. tachangensis germplasms, mainly from the MM indigenous community. The third pure population (referred to as the ancient landraces of BY or GP03 hereafter) consisted of 18 (62.1%) ancient landraces, 3 (10.3%) modern landraces, and 8 (27.6%) wild type accessions, of which 21 (72.4%) belonging to C. sinensis were mainly from the BY indigenous community. The fourth pure population (referred to as the wild type accessions of PR or GP04 hereafter) consisted of 18 wild germplasms, including 12 (66.7%) C. gymnogyna and 6 (33.3%) C. tachangensis, mainly from the PR indigenous community. The fifth pure population (referred to as the modern landraces of BY and DQ or GP05 hereafter) was composed of 43 germplasms, of which 37 (86.0%) were modern landraces, mainly belonging to the C. sinensis from the BY and DQ integrate indigenous communities. Ninety-nine accessions (approximately 23.9% of a total of 415 samples) had admixture ancestry. In the admixture population (referred to as the waterway traffic population of multiethnic communities or GP06 hereafter) 64 (64.6%) were wild type accessions, 19 (19.2%) were ancient landraces, and 16 (16.2%) were modern landraces, including 33 (33.3%) *C. sinensis*, 31 (31.3%) *C. tachangensis*, 3 (3.0%) near *C. taliensis*, and 32 (32.3%) *C. gymnogyna* from the BY, HR, MM, PR and DQ indigenous communities. Based on the geographical location, among them, the origin of 78 (78.8%) accessions were near the rivers, while 21 (21.2%) accessions were far away from them, of which 14 (66.7%) accessions were near the ancient traffic post road, but 7 (33.3%) accessions were far away from them (Additional file 1: Table S3).

The collection of 415 accessions were analyzed by principal component analysis (PCA) and the Neighbor-Joining (NJ) phylogenetic tree to better understand the relationships between potential inferred populations. The result of PCA was highly consistent with that of the ADMIXTURE analysis (Fig. 3A, B and Additional file 5: Fig. S3). PCA revealed five main clusters, corresponding to five pure populations identified by the ADMIX-TURE software. The GP02 and GP04 clusters were more



Fig. 3 Phylogenetic relationships and population structure of 415 tea accessions. (**A**) Inferred population structure of 415 accessions. Bar plot of individual membership coefficients for the genetic clusters is inferred by using ADMIXTURE (K=5) base on 99,363 SNPs. Individual membership coefficients are sorted within each cluster. (**B**) Scatter diagram of the first three principal components. The first, second and third coordinates indicate PC1, PC2 and PC3. Each dot represents an accession. (**C**) A Neighbor-Joining (NJ) phylogenetic tree of all 415 tea accessions in Guizhou Plateau constructed using whole-genome SNP data based on the maximum composite likelihood (MCL) model. The NJ tree compared with indigenous communities (HR, DQ, BY, MM and PR indigenous communities) and ADMIXTURE software results (K=5). The lines with different colors represent the indigenous communities (the black blocks in the middle represent accessions imported from other provinces), and the squares with different colors nearby represent the populations we divided according to the results of ADMIXTURE. GP01, GP02, GP03, GP04 and GP05 are pure populations and GP06 is admixture population. (**D**) The LD decay plot of 415 accessions and six inferred populations

Ŀ.	GP01	0					
genet (<i>Fst</i>)	GP02	0.273b	0		_		
ation	GP03	0.035g	0.147d	0		_	
opula rentia	GP04	0.141d	0.137d	0.263b	0		
Inter-po diffe	GP05	0.062f	0.293a	0.079e	0.230c	0	
	GP06	0.052fg	0.145d	0.042g	0.077ef	0.065ef	0
	Tajima's D	0.470	0.619	0.295	0.563	0.334	0.770
Intra-	Pi	0.186c	0.198b	0.173d	0.214a	0.145e	0.190c
population genetic diversity	Но	0.071cd	0.085b	0.067d	0.104a	0.066d	0.077c
	MAF	0.121d	0.139b	0.115e	0.147a	0.099f	0.128c
	Fis	0.631a	0.657a	0.626ab	0.516b	0.553b	0.646a

Fig. 4 Genetic diversity of six inferred populations of 415 accessions in Guizhou Plateau. Inter-population genetic differentiation coefficient (*Fst*), nucleotide diversity (*P*), observed heterozygosity (*Ho*), minor allele frequency (*MAF*) and inbreeding coefficients (*Fis*). The colors and numbers in the cells of the matrix represent the *Fst* values. The colors and numbers in the cells below the *Fst* matrix represent the genetic diversity indices. The different letters indicate a significant difference in p = 0.05 levels by the T-test. GP01, GP02, GP03, GP04 and GP05 are pure populations and GP06 is admixture population base on ADMIXTURE software at K=5

scattered than GP01, GP03 and GP05, and the GP06 cluster was distributed between these five clusters along the PC1, PC2 and PC3 axes. The NJ tree also agreed with the ADMIXTURE analysis results (Fig. 3C). Furthermore, the results of the NJ tree were highly consistent with those of indigenous communities (MM, PR, BY, DQ and HR), the cultivation status (WA, AL and ML), and the species classification (*C. sinensis, C. tachangensis*, near *C. taliensis* and *C. gymnogyna*) (Additional file 5: Fig. S6).

Linkage disequilibrium analysis

In this study, 29,393,327 non linkage disequilibrium (LD) pruned SNPs were used to estimate the LD of 415 tea accessions. LD decay occurred rapidly with the increase in physical distance. The maximum r^2 value was 0.25 for the LD decay in all 415 tea accessions. As LD decayed to its half maximum (r^2 =0.13), the corresponding physical distance was approximately 8 Kb (Fig. 3D).

LD decay in these six inferred populations was estimated. The lowest LD decay was observed in GP05, as r^2 reached 0.13 (threshold) at approximately 40 Kb. On the contrary, LD decay was the highest in GP04, with r^2 =0.13 at the corresponding physical distance of approximately 12 Kb. The corresponding physical distances of GP02, GP06, GP01 and GP03 at r^2 =0.13 were approximately 15 Kb, 17 Kb, 26 Kb and 31 Kb, respectively.

Genetic differentiation analysis of the inferred populations

According to the results of population structure analysis, the genetic diversity indices were calculated for GP01, GP02, GP03, GP04, GP05 and GP06. *Pi*, *Ho* and *MAF* for GP04 were the highest, while those for GP05 were the lowest. We also found that the *Fis* was the highest in GP02 but the lowest in GP04. The Tajima's *D* values in all six inferred populations were positive, indicating the occurrence of balancing selection or population bottle-necks (Fig. 4).

The *Fst* values in inferred populations ranged from 0.035 to 0.293, with an average value of 0.136. (Fig. 4). The maximum *Fst* value (0.293) was observed in GP02 vs. GP05, indicating that there was the great divergence between these two populations. The minimum *Fst* value (0.035), however, was observed in GP01 vs. GP03, indicating that there was the small divergence between these two populations. Moreover, the highest *Nm* value was showed in GP01 vs. GP03, which corresponded to the lowest pairwise *GD* value (Additional file 2: Table S6).

Core collection development

Core collections provide GenBank curators and plant breeders with a way to reduce the size of their collections and populations while minimizing the impact on genetic diversity and allele frequency. This information can be used in molecular marker-assisted breeding, GWAS and other applications [35, 36]. We used a simple and flexible method, Core Hunter 3, to construct a core collection of 136 accessions (~1/3 of the original population size), which were selected to represent a set of 415 tea accessions and henceforth referred to as the 'core set' (Additional file1: Table S1).



Fig. 5 Construction a core collection of 415 tea accessions population. (**A**) The NJ trees of whole set and core set. (**B**) Genetic diversity of whole and core sets. Nucleotide diversity (*Pi*), observed heterozygosity (*Ho*), minor allele frequency (*MAF*) and average genetic distance (*AGD*). The different letters indicate a significant difference in p = 0.05 levels by the T-test

Table 2 Statistical analysis of OTL and OtL

Trait	Mean	Std.dev	Minimum	Maximum	Skewness	Skewness (Std.Err)	Kurtosis	Kurtosis (Std.Err)	CV (%)
OTL	4.325	1.522	1	7	-0.084	0.153	-0.319	0.306	35.2
OtL	5.944	1.370	2	9	-0.527	0.153	0.845	0.306	23.0

To evaluate the quality of the core set, we constructed the NJ trees and used the GD matrix to verify whether its backbone changed. We found that the core accessions were evenly distributed across the whole collection of 415 accessions that were consistently revealed by the NJ trees (Fig. 5A). We evaluated Pi, Ho, MAF and GD among the whole set and the core set. The core set had 100% of MAF, 99% of Pi, and 92% of Ho of the whole set. The average genetic distance (AGD) for the whole and core sets were 0.229 and 0.196, respectively (Fig. 5B). In the GD matrix of the whole set and the core set, the range of GD frequency (0.1 to 0.3) was the highest (Additional file 5: Fig. S7 A). These results showed that the core set included WA, AL and ML from five indigenous communities (Additional file 5: Fig. S7 B and C). Therefore, they could represent the genetic diversity of the whole set.

Phenotypic variation analysis of OTL and OtL

The OTL and OtL traits of 252 tea accessions were investigated. Table 2 showed that the average values of OTL and OtL were 4.325 and 5.944, with a coefficient of variation (CV) of 35.2% and 23.0%, respectively. Extensive genetic variation in OTL and OtL was investigated, and the phenotypic values of the OTL and OtL traits showed normal distribution or approximately normal distribution (Additional file 5: Fig. S8 A and B). These results revealed a broad trait diversity among the populations of 252 tea accessions.

GWAS analysis for OTL and OtL

The MLM and GLM methods were used for performing GWAS to detect and evaluate the genotypic variations in OTL and OtL of tea plant. In addition, the QQ plots were used to assess the extent of the difference between the observed and the expected P-values (Additional file 5: Fig. S9 A and B). The GLM method found a total of five SNP markers significantly associated with OTL $(-\log_{10}(P)=5.00, P=1/99,363),$ including S2_68972914 with $R^2 = 14.19\%$ and allele T/A, S6_11866408 with $R^2 = 12.12\%$ and allele C/G, $S12_{92443520}$ with $R^2 = 13.75\%$ and allele G/T, S6_137947647 with $R^2 = 11.94\%$ and allele C/T, and S2_68972933 with R^2 =12.02% and allele G/T. The MLM method found two SNP markers significantly associated with OtL $(-\log_{10}(P)=5.00)$, including S5_141340508 and S11_64920241 with $R^2 = 22.44\%$ and allele T/C, and R^2 =12.94% and allele C/T, respectively (Table 3). The Manhattan plots were shown in Fig. 6A.

Making further efforts to reveal the functions of the trait-associated SNPs, we searched for genes located within the 50 Kb regions surrounding the SNP loci using the BLAST analysis by the tea plant reference genome [31]. Finally, we found two genes located within the candidate regions probably associated with OTL, including a receptor-like protein kinase gene and a peroxidase gene. We also found two genes located within the candidate regions probably associated with OtL, including an L-tryptophan–pyruvate aminotransferase gene and a DNA methylation-related gene (Fig. 6B–E).

Table 3	SNP	loci significantl	y associated	l with OTL	. and OtL	using G	iLM and	MLM
---------	-----	-------------------	--------------	------------	-----------	---------	---------	-----

Trait	SNP Marker	Chromosome	Position	-log ₁₀ P	R ² (%)	Allele
OTL	S2_68972914	2	68,972,914	5.75	14.19%	T/A
	S6_11866408	6	11,866,408	5.41	12.12%	C/G
	S12_92443520	12	92,443,520	5.18	13.75%	G/T
	S6_137947647	6	137,947,647	5.17	11.94%	C/T
	S2_68972933	2	68,972,933	5.04	12.02%	G/T
OtL	S5_141340508	5	141,340,508	7.34	22.44%	T/C
	S11_64920241	11	64,920,241	5.26	12.94%	C/T

Note Significant SNP loci with $-\log_{10}(P) = 5.00$ (P = 1/n, n = 99,363). R^2 is the percentage of phenotypic variance explained by the SNP



Fig. 6 GWAS analysis and discovery of candidate genes for OTL and OtL. (**A**) Manhattan plot for the OTL of tea plant by the optimal model of GLM, and Manhattan plot for the OtL of tea plant by the optimal model of MLM. The red dashed horizontal line indicates the significance threshold $(-\log_{10}(P)$ is about equal to 5.00). (**B**, **C and D**) Genes within 50 Kb regions surrounding the significant trait-associated SNP loci according to BLAST analysis using the tea plant reference genome. (**B**) a is peroxidase gene, 1.79 Kb. (**C**) b is L-tryptophan–pyruvate aminotransferase gene, 8.30 Kb. c is DNA methylation-related gene, 2.39 Kb. (**D**) d is receptor-like protein kinase gene, 27.03 Kb. (**E**) Molecular function annotations of candidate genes

RT-qPCR analysis

To determine whether four potential candidate genes control OTL and OtL of tea plant, we measured the expression levels of the TEA028298.1 and TEA012503.1 genes associated with OTL, and the TEA027937.1 and TEA013853.1 genes associated with OtL using RT-qPCR approach. The results showed that the expression levels of these potential candidate genes were different in tea accessions at three germination stages. They all had the highest expression levels in accessions at the extremely early germination stage (EEGS) (Fig. 7A and B and Additional file 6).

Discussion

Genetic diversity analysis

GBS has been widely used to analyze the genetic diversity of rice [37], olive [38] and Chinese chestnut (*Castanea mollissima* Blume) [39], which is a convenient and costeffective approach [40, 41]. We used stringent filtering criteria to generate 390.30 Gb clean reads and identify 99,363 high-quality SNPs from 415 tea accessions. The number of SNPs identified in our result was higher than those found in previous studies [27, 42, 43], indicating that the GBS is a powerful method for the genetic diversity analysis of tea plant, and the sequencing data can be used for subsequent research.

In this study, the genetic diversity of the wild type accessions population was significantly higher compared



Fig. 7 RT-qPCR analysis of four candidate genes for OTL and OtL. (A) The expression level of TEA028298.1 and TEA012503.1 genes, two candidate genes associated with OTL. (B) The expression level of TEA027937.1 and TEA013853.1 genes, two candidate genes associated with OtL. The EEGS, IGS and ELGS represent the three stages of tea accessions germination period, including the extremely early germination stage (EEGS), intermediate germination stage (IGS), and extremely late germination stage (ELGS). Source data underlying is provided as an additional file 6

with that of the modern landraces population, because breeding practices tend to reduce genetic diversity to a greater extent than domestication [44, 45]. We also found that the genetic diversity of modern landraces population was significantly lower than that of ancient landraces population. A reasonable explanation for this result was that inbreeding or continuous directional selection during the breeding process narrows the genetic base of germplasm, reduces genetic diversity, and increases the possibility of genetic drift during the domestication process.

Interestingly, we observed that the genetic diversity was significantly higher for the tea populations in MM and PR indigenous communities than for the tea populations in BY, DQ and HR indigenous communities. The genetic diversity results of the inferred populations were basically consistent with the corresponding indigenous community populations. Among the five pure inferred populations, the genetic diversity of GP04 and GP02 was significantly higher than that of GP01, GP03 and GP05.

Previous studies have revealed that the custom, socioeconomic factors and management influenced the genetic diversity patterns of tea plant [6, 13]. We found that the culture of indigenous communities might be related to conservation practices of tea plant. For example, tea plant was important in the wedding ceremony of MM and they used it as the bride price to symbolize the fidelity and love between men and women [46]. Tea plant also played an important role in BY's daily life. In tea drinking customs of BY, the ancient tea drinking methods and experiences have been inherited and developed. After long-term migration and settlement, BY lived in harmony with other indigenous peoples and often made friends through tea [46]. In addition, tea plant was regarded as one of the most important offerings in the ceremony of deity worship by PR [47]. Moreover, tree worship, especially tea plant, was one of the most important forms of worship for DQ. Tea plants have played an important role in the life of Yi as a tribute to ancestors whom they worship, as well as for farming practices [48].

MM and PR were the oldest ethnic groups living on the Guizhou Plateau [26, 46]. They relied more on natural resources for survival and were not good at changing the natural environment. Therefore, the natural evolutionary habits of wild tea resources have been preserved, indicating that the wild tea populations (GP02 and GP04) have abundant genetic diversity, which will facilitate the genetic improvement of tea plant in future plant breeding programs. As the two ethnic groups gradually migrated, some relatively enlightened MM evolved into BY, while some relatively enlightened PR evolved into DQ [26, 49]. The BY and DQ have developed agricultural civilization to a certain extent and have made certain choices in the utilization of tea plant [46]. The tea plants distributed in these two indigenous communities were gradually evolving towards ancient landraces, reducing the genetic diversity of the tea populations (GP03 and GP01).

Since the Tang Dynasty, HR have migrated from the northern Guizhou Plateau to the central region, which has begun to influence the culture and management of PR, DQ, BY and MM indigenous communities [49]. The integration of HR with other indigenous peoples has further developed more advanced agricultural civilization, strengthening the targeted selection and utilization of tea plant. This historical event significantly reduced the genetic diversity of modern landraces population (GP05) in ethnic fusion region. According to the geographic distribution of sample sites, many tea plant accessions were distributed at the traffic nodes along the river, which promoted frequent gene exchange among tea populations in these five indigenous communities.

Population structure analysis

ADMIXTURE has been used to analyze the population structures of maize [50], wheat [51] and tea [52, 53]. It could effectively identify global clusters, and verified accuracy of its results via PCA and phylogenetic tree. The K value corresponding to the minimum CV error was deemed the optimal value for determining population structure. In this study, the clustering of 415 tea accessions correlated well with the distribution of indigenous communities at K=5 using the ADMIXTURE software. A total of 415 accessions were divided into six populations, including five pure populations (GP01, GP02, GP03, GP04 and GP05) and one admixture population (GP06). The NJ tree and PCA generated the same population structure output as ADMIXTURE. GP02 was C. tachangensis mainly from the MM indigenous community. GP04 was C. gymnogyna mainly from the PR indigenous community. GP01 represented ancient landraces, C. sinensis, mainly from the DQ indigenous community. GP03 was ancient landraces, C. sinensis, mainly from the BY indigenous community. GP05 was modern landraces, C. sinensis, mainly from the BY and DQ integrate indigenous communities, and GP06 was an admixture population mainly distributed at traffic nodes along the river.

Previous studies have reported that the *Fst*, pairwise *GD*, *Nm* and *Fis* were the key parameters in determining genetic differentiation between populations [33, 34]. In this study, we estimated the pairwise *GD*, *Fst*, *Fis* and *Nm* among six inferred populations. The higher pairwise *GD* and *Fst* and lower *Nm* were found in GP02 vs. GP01, GP04 vs. GP03, GP02 vs. GP05, GP04 vs. GP05, GP04 vs. GP01, and GP03 vs. GP02, while the smaller pairwise *GD* and *Fst* and higher *Nm* were observed in GP03 vs.GP01, GP05 vs.GP01, and GP05 vs. GP03.

A reasonable explanation was that ethnic cultural differences have to some extent influenced the genetic differentiation of tea populations over the long term. GP02 and GP04 were wild type accessions populations, distributed in the MM indigenous community and PR indigenous community, respectively. The MM and PR were the two oldest indigenous groups living in the Guizhou Plateau [26, 46]. They lived in the high mountain jungle and had less communication with other indigenous groups, which reduced gene exchange between the tea accessions distributed in these two indigenous groups and the outside world. These two indigenous groups have always utilized natural resources rather than altering them. Therefore, these two indigenous communities maintained the higher genetic diversity and retained the primary wild type accessions. With the migration of ethnic groups, a tribe of MM gradually evolved into BY, while a tribe of PR gradually evolved into DQ [26, 49]. BY and DQ had a certain agricultural civilization. They made tea accessions further evolved into a favorable direction for utilization by selecting, thus increasing the genetic differentiation between BY (GP03) vs. MM (GP02) and DQ (GP01) vs. PR (GP04). Afterwards, with the migration of HR and the integration with BY and DQ, the development of agricultural civilization in BY and DQ was further promoted, and the selection and utilization of tea accessions were strengthened. This result further increased the genetic differentiation between the modern landraces population distributed in the BY and DQ integrate indigenous communities (GP05) and the wild type accessions populations distributed in the MM and PR indigenous communities (GP02 and GP04).

Furthermore, we found that GP06 may be a product of the gene exchange between GP04, GP02, GP01 and GP03 through transportation hubs [54, 55] and may have formed earlier than GP05. The former may be gradually distributed and evolved into GP05 due to economic trade, development of the modern tea industry and ethnical cultural exchanges. Therefore, the GP06 may have a complex evolutionary history, and its utilization as a natural resource still needs further exploration.

Core collection development

Establishing a core collection that has the lowest level of genetic redundancy and preserves the maximum genetic diversity of the entire collection facilitates the identification of varieties suitable for GWAS [56]. Core collections can be evaluated based on different types of genetic markers, phenotypic traits and pairwise distance matrices [57]. Choosing the most appropriate evaluation method depends on the purpose of the core collection [58]. Preserving most alleles is an ideal way to conserve germplasm, while distance-based methods mainly maintain most allele combinations in specific genotypes, making them suitable for GWAS. Thus, we applied a precomputed distance matrix to GBS data to establish the core collection of tea accessions in the Guizhou plateau, including wild type accessions, ancient landraces and modern landraces from indigenous communities. The core collection was constructed based on the optimization of the average genetic distance between each accession and the nearest neighboring entry in the core collection (E-NE) model, which contained 33% of the total number of accessions. Ho, Pi and MAF of the core set accounted for more than 90% of the whole set, indicating that the core set can well represent 415 tea accessions for further research. This core collection will facilitate further development of GWAS and the breeding of tea varieties with excellent characteristics.

Genome-wide association study analysis

The choice of plant materials used for association mapping is very critical as they should have a wide range of diversity to capture the maximum number of historical recombination events. Tea plant is a self-incompatible and highly heterozygous woody perennial tree. Compared with self-pollinating plants, tea plants have more extensive genetic variation and higher diversity due to domestication through hybridization, long-term allogamy, and climate-based selection [59]. We used tea populations to perform GWAS analysis and discovered a set of SNP markers and genes associated with OTL and OtL, suggesting that the offspring with high genetic diversity produced through cross-pollination are suitable for association mapping. Moreover, we found that the majority of tea accessions from DQ indigenous community were at the earlier germination stage, which may be related to the local geographical location with mild climate and sufficient rainfall [60].

In this study, we conducted a GBS-based GWAS to discover five and two SNP loci associated with OTL and OtL through the optimal GLM and MLM models, respectively, which are the two most commonly used algorithms in GWAS [61]. These associated SNP loci could be considered as the candidate genetic markers affecting OTL and OtL traits. Using BLAST with the tea plant reference genome [31], we identified two potential candidate genes related to OTL, including a receptor-like protein kinase gene and a peroxidase gene, and two potential candidate genes related to OtL, including an L-tryptophan-pyruvate aminotransferase gene and a DNA methylationrelated gene within a 50 Kb region around the significant trait-associated SNP loci, which were validated using RT-qPCR approach. We found that LD became too low at distances greater than 50 Kb. Therefore, 50 Kb was selected as a reasonable distance that caused moderate LD between the gene and trait-associated SNPs [59].

Peroxidases are a superfamily of antioxidant enzymes that catalyze the oxidation of various substrates using H_2O_2 or other organic hydroperoxides enzymes and protect plant cells from oxidative stress damage. Evidence has proved that peroxidases not only are important components of plant growth and development but also can loosen plant cell walls, leading to dormancy break and bud germination [62]. We discovered one peroxidase gene located within 1.79 Kb downside of the significant trait-associated SNP, speculating that this gene was a potential candidate gene for the OTL trait in tea plant. The receptor-like protein kinase FERONIA is a key regulator of stress responses mediated by brassinosteroids (BRs). Previous study has shown an antagonistic relationship between ethylene and endogenous BRs for controlling hypocotyl elongation. The fer-2 mutant exhibited the loss of responses modulated by BRs, which led to an imbalance in this relationship and increased the effect of ethylene on growth inhibition [63]. Besides, the pathway related to the FER receptor kinase has been identified, which negatively regulated the ABA response through the activation of an A-type PP2C phosphatase, ABI2, promoting plant growth and leaf development [64]. In our research, one receptor-like protein kinase gene was located within 27.03 Kb upside of the significant trait-associated SNP.

L-tryptophan aminotransferase has been proved to be involved in the biosynthesis of auxin (IAA), and its first step is catalyzed in the IPA branch of the IAA biosynthesis pathway to initiate multiple growth changes in response to environmental and developmental cues. Maintaining appropriate IAA levels in roots requires TAA1 and TAR2, while the development of proper embryo patterning requires functions of TAA1, TAR1 and TAR2, which have an impact on plant growth and leaf development [65]. In our research, one L-tryptophanpyruvate aminotransferase gene was located within 8.30 Kb upside of the significant trait-associated SNP. RNAdirected DNA methylation (RdDM) is a process in which RNA molecules direct the addition of DNA methylation through RNA-DNA sequence interactions. Previous study has reported that the phenomenon of mRNA degradation directed by RdDM and siRNA in plants is the silencing of specific genes of matching sequence by RNA interference, which has a significant impact on the regulation of plant gene expression and development [66]. In our research, one DNA methylation-related gene was located within 2.39 Kb on the downside of the significant trait-associated SNP. These reports have confirmed our findings that the four potential candidate genes identified in this study are related to OTL and OtL of tea plant, which will be beneficial for future development of molecular markers and MAS breeding.

Conclusions

Genome-wide SNPs in tea accessions from the origin center, Guizhou Plateau, were identified in this study using GBS. These SNPs were used to analyze the population structure, genetic diversity and LD pattern of 415 accessions, which were divided into six populations, including five pure populations (ancient landraces of DQ, ancient landraces of BY, wild type accessions of MM, wild type accessions of PR and modern landraces of BY and DQ) and one admixture population (waterway traffic population of multiethnic communities). All 415 accessions exhibited high genetic diversity. The management and cultural practices of different indigenous communities in the Guizhou Plateau may play important roles in the protection and enhancement of the genetic diversity of tea plant. We developed a core set of 415 tea accessions and identified four potential candidate genes associated with OTL and OtL traits. The results of this study will help support genetic diversity conservation, the introduction of cultivars, and further research on tea plants that are valued by different indigenous peoples.

Methods

Plant materials

A collection of 415 tea accessions were used in this study. Based on the geographical distribution of indigenous peoples [26], the Guizhou Plateau included five indigenous communities. Miao living area is referred to as Miao Man (MM) indigenous community (included 107 accessions), Gelao living area is referred to as Pu Ren (PR) indigenous community (included 55 accessions), Buyi living area is referred to as Bai Yue (BY) indigenous community (included 89 accessions), Tujia and Yi living area is referred to as Di Qiang (DQ) indigenous community (included 139 accessions), and Han living area is referred to Han Ren (HR) indigenous community (included 20 accessions). The other five tea accessions were collected from Fujian, Zhejiang, Hunan and Guizhou Provinces and planted in the tea garden of Guiyang, China. Among the 415 tea accessions, 159 samples were identified as arbor and 256 samples were identified as shrub, according to the field investigation of tree types. Of these, 56 samples were large leaf, 204 samples were medium leaf, and 155 were small leaf. Based on the research of Niu et al. [27], wild tea plants older than 100 years and their natural offsprings are referred to as "wild type accessions (WA)", cultivated tea plants older than 100 years are referred to as "ancient landraces (AL)", while cultivated tea plants from tea gardens are referred to as "modern landraces (ML)". Of these, 159 accessions were WA, 174 accessions were AL, and 82 accessions were ML. According to the classification systems of Min [67] and Chen et al. [68], 251 C. sinensis, 100 C. tachangensis, 5 near C. taliensis and 59 C. gymnogyna were identified (Additional file 1: Table S1).

Geographic information was acquired using GPS, and based on field investigations during the sampling process and Yang et al.'s research [26], the distributions of tea accessions in the indigenous communities of Guizhou Plateau (MM, PR, DQ, BY and HR) were determined. The GIS technology was used for mapping. Professor Suzhen Niu from the Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region (Ministry of Education), Institute of Agro-Bioengineering, Guizhou University, associate professor Jie Yin and lecturer Qinfei Song from the College of Tea Science of Guizhou University, and researcher Zhengwu Chen from Institute of Tea, Guizhou Academy of Agricultural Sciences conducted morphological identification on all tea accessions, and stored them in the tea germplasm resource garden of Guizhou University. This study and experiment complied with local legislation, and national and international guidelines. The authors also complied with the Convention on the Trade in Endangered Species of Wild Fauna and Flora and Regulations of Guizhou Province on the protection of ancient tea plants.

DNA extraction, library preparation and sequencing

DNA was extracted from young leaves using a Plant Genomic DNA Rapid Extraction Kit (Beijing Biomed Gene Technology Co., Ltd., Beijing, China). The integrity and purity of DNA were tested by electrophoresis in 1% agarose gel and the Qubit Fluorometer (Invitrogen), and then the DNA samples were stored at -20 °C. The DNA content for library preparation was 100 ng, and the purified DNA was digested by the SacI and MseI restriction enzymes from New England Biolabs (NEB). Thereafter, the SacAD and MseAD adaptors were ligated to the end of digested DNA fragments for barcode ligation, gel DNA fragment selection, adaptor connection, and PCR amplification of fragments, and 500-550 bp amplified products were recovered. The paired-end 150 bp (PE150) sequencing was performed on the Illumina HiSeq X Ten platform [33, 34].

SNP calling and quality control

The adaptor was trimmed with the special Perl script, and the barcodes were used to demultiplex a set of raw sequence DNA reads. The reads with a base quality value below 5 were filtered out, and the remaining reads were used as clean data, which were then mapped to the reference genome (http://tpia.teaplants.cn/) [31] using BWA-MEM (v. 0.7.10) run with the default parameters [69]. GATK (v. 3.7.0) was used to perform SNP calling [70]. The criteria for filtering SNPs were as follows: (1) The variants (SNPs) must be biallelic. (2) GATK (v. 3.7.0) was used to filter SNPs with the parameters, including QUAL<50.0 || QD<2.0 || FS>60.0 || MQ<40.0 || Mapping Quality Rank Sum (MQRankSum) < -12.5 || Read-PosRankSum < -8.0. (3) The VCFtools program package (v. 0.1.15) was used to filter SNPs with minor allele frequencies (MAF)<0.05 or missing data rates>20% [27, 71]. The CMplot (v. 3.7.0) software was used to draw the SNP density plot [33]. A total of 99,363 high-quality SNPs (through LD pruned) from the 415 tea accessions were selected and subjected to the subsequent population structure and genetic diversity analysis.

Genetic diversity analysis

Genetic diversity indices included observed heterozygosity (*Ho*), minor allele frequency (*MAF*), inbreeding coefficient (*Fis*), nucleotide diversity (*Pi*), Tajima's *D*, genetic differentiation coefficient (*Fst*), gene flow (*Nm*) and genetic distance (*GD*). *MAF*, *Ho* and *Fis* values of each inferred population were calculated using Plink (v. 1.9) [72]. *Pi* and Tajima's *D* values of each inferred population and *Fst* values of the pairwise inferred population swere calculated using VCFtools (v. 0.1.15) [71]. *Nm* was calculated using the formula Nm = (1-Fst)/4Fst [33]. *GD* values of pairwise inferred populations were calculated by the MEGA (v. 10.2.4) software [73]. The significant differences between these indices were determined by SPSS Statistics (v. 26) software.

Population structure analysis

PLINK (v. 1.9) was used to convert the variant call format (VCF) files into the pedigree files [72]. Population structure was analyzed by the ADMIXTURE (v. 1.3.0) software, with the number of ancestral populations (K) ranging from 2 to 9 [74]. To determine the optimal population structure classification, the optimal K value was identified based on the minimum cross-validation (CV) error [34]. The threshold value for the membership coefficient was set to 0.6 to distinguish between admixture and pure populations [33]. The TASSEL (v. 5.2.72) software was used to perform PCA [75]. To clarify the genetic relationship among the inferred populations, phylogenetic trees were constructed. MEGA (v. 10.2.4) was used to calculate the distance matrix and construct the Neighbor-Joining (NJ) phylogenetic tree based on the maximum composite likelihood (MCL) model with 1000 bootstrap replicates [76].

Linkage disequilibrium analysis and core collection development

To estimate and compare the patterns of linkage disequilibrium (LD) between different inferred populations, the squared correlation coefficient (r^2) between genomewide unpruned SNPs was calculated and plotted using the PopLDdecay (v. 3.29) package [77].

We defined a core set of 415 tea accessions using the Core Hunter 3 program [57], according to the optimization of average genetic distance as described by Odong et al. [58] (Additional file 3).

Phenotypic trait evaluation and statistical analysis

We evaluated two traits related to the germination period, including the germination period of one bud and two leaves (OTL) and the germination period of one bud and three leaves (OtL). As shown in Additional file 4, each accession was examined for signs of germination from the beginning of March until beginning of May and scored from stage 1 to 9. The TASSEL (v. 5.2.72) software [75] was used to analyze phenotypic data. Analysis of statistics was estimated by SPSS Statistics (v. 26) software based on the trait means of each tea accession.

GWAS analysis for OTL and OtL

Based on the selected SNP markers developed from a total of 252 accessions, GWAS was performed for the OTL and OtL traits using the TASSEL (v. 5.2.72) software by two methods, namely the mixed linear model (MLM) and the general linear model (GLM) [59]. The red dotted line indicated the genome-wide control threshold $-\log_{10}(P$ -value)=5.00 (*P*-value was equivalent to 1/n,

where n was the total number of markers applied) [59]. Manhattan and Quantile–Quantile plots (QQ plots) were drawn using TASSEL (v. 5.2.72) software [75]. The trait-associated SNP markers were obtained using the optimal GLM and MLM models. Genes located within the 50 kb regions surrounding the trait-associated SNP loci were considered potential candidate genes.

RNA extraction for RT-qPCR analysis

We selected nine tea plant accessions, including three accessions at the extremely early germination stage (EEGS), three accessions at the intermediate germination stage (IGS), and three accessions at the extremely late germination stage (ELGS). Total RNA was extracted from each tea accession based on the traits (OTL and OtL) using the UNIQ-10 column Trizol Total RNA Purification Kit (Sangon Biotech Co., Ltd., Shanghai, China). The expression patterns of four potential candidate genes identified through GWAS analysis were verified using RT-qPCR, which was performed using the ChamQ Universal SYBR qPCR Master Mix Kit (Nanjing Novozan Biotechnology Co., Ltd., Nanjing, China). The results were analyzed using the $2^{-(\triangle \triangle Ct)}$ method [78] with GAPDH used as an internal reference gene. A total of three biological and three technical replicates were used.

Abbreviations

C. sinensis	Camellia sinensis (L.) O. Kuntze
C. gymnogyna	Camellia gymnogyna Chang
C. tachanaensis	Camellia tachangensis F.C. Zhang
C. taliensis	Camellia taliensis (W.W. Smith) Melchior
GBS	Genotypina-by-sequencina
GWAS	Genome-wide association study
SNPs	Single-nucleotide polymorphisms
OTL	Germination period of one bud and two leaves
OtL	Germination period of one bud and three leaves
MM	Miao Man
PR	Pu Ren
DQ	Di Qiang
BY	Bai Yue
HR	Han Ren
LD	Linkage disequilibrium
WA	Wild type accessions
AL	Ancient landraces
ML	Modern landraces
Но	Observed heterozygosity
MAF	Minor allele frequency
Fis	Inbreeding coefficient
Pi	nucleotide diversity
Fst	Genetic differentiation coefficient
Nm	Gene flow
GD	Genetic distance
PCA	Principal component analysis
CV error	Cross-validation error
NJ tree	Neighbor-Joining tree
MCL	Maximum composite likelihood
MLM	Mixed linear model
GLM	General linear model
EEGS	Extremely early germination stage
IGS	Intermediate germination stage
ELGS	Extremely late germination stage
AGD	Average genetic distance

E-NE RdDM The average genetic distance between each accession and the nearest neighboring entry in the core collection RNA-directed DNA methylation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-04761-x.

Additional file 1: table S1. Information of 415 tea accessions used in this study. Table S2. Statistics of the number and ratio of the accessions of cultivation status, species classification and indigenous communities in six inferred populations. Table S3. Statistics of the number and ratio of the accessions of near ancient hubs and near rivers in six inferred populations.

Additional file 2: table S1. The quality control data of 415 tea accessions. Table S2. Statistics of heterozygosity rate of 99,363 SNPs in 415 tea accessions. Table S3. Percentage of transition and transversion SNPs identified using genotyping-by-sequencing. Table S4. Genetic diversity of 415 tea accessions in Guizhou Plateau. Table S5. The Nm and pairwise genetic distance among five indigenous community populations. Table S6. The Nm and pairwise genetic distance among six inferred populations of 415 accessions.

Additional file 3: table S1. Precomputed pairwise genetic distance matrix of 415 accessions.

Additional file 4: table S1. Phenotypic data for OTL and OtL of tea plant.

Additional file 5: fig. S1. Geographic distribution of sampling sites. Geographical distribution of Ming Dynasty ancient transportation hubs and rivers in Guizhou Plateau. Fig. S2 Geographical distribution of tea accessions collected in this study. Geographical distribution of cultivation status (i.e., wild type accessions (WA), ancient landraces (AL) and modern landraces (ML)) and species classification (i.e., *C. sinensis, C. tachangensis,* near *C. taliensis* and *C. gymnogyna*) in Guizhou Plateau. Fig. S3 Geographic distribution of sampling sites. Geographical distribution of populations (i.e., GP01, GP02, GP03, GP04, GP05 and GP06) inferred by ADMIXTURE software (K=5) in Guizhou Plateau. Fig. S4. Distribution map of SNPs on 15 chromosomes graph. Fig. S5. Graph for CV error in the range of K=2–9 of 415 tea accessions. Fig. S6. Cluster analysis using NJ trees. Fig. S7. Summary of comparison information among core and whole sets. Fig. S8. Phenotype frequency distribution of OTL and OtL. Fig. S9.GWAS analysis for OTL and OtL.

Additional file 6: table S1. Analysis of TEA028298.1 potential candidate gene expression in 9 tea accessions. Table S2. Analysis of TEA012503.1 potential candidate gene expression in 9 tea accessions. Table S3. Analysis of TEA027937.1 potential candidate gene expression in 9 tea accessions. Table S4. Analysis of TEA013853.1 potential candidate gene expression in 9 tea accessions in 9 tea accessions.

Acknowledgements

We thank Chishui, Daozhen, Dejiang, Duyun, Guanling, Guian, Guiding, Huaxi, Huishui, Jinsha, Liping, Liuzhi, Nayong, Puan, Puding, Qinglong, Qixingguan, Renhuai, Sandu, Shiqian, Shuicheng, Tongzi, Wuchuan, Xingren, Xingyi, Xishui, Yanhe, Yinjiang, Yuqing, Zhenfeng, Zhengan and Zhijin for their help in tea plants collection. We appreciate the research facilities and computer facilities provided by the College of Tea Science, Guizhou University.

Author contributions

Y.H.W, S.Z.N and Z.W.C conceived and supervised the study. Y.H.W, X.Y.D and D.C.B analyzed and interpreted the genetic diversity, linkage disequilibrium and population structure. Y.H.W and S.Z.N wrote and reviewed the manuscript. Y.H.W draw all maps in the article and conduct GWAS analysis. X.L.D and D.J.H performed the DNA extraction and filtered the genotyping data. All authors read and approved the final version of the manuscript.

Funding

This work was funded by the Project of the National Science Foundation, in RP China (32060700), the National Guidance Foundation for Local Science and Technology Development of China ([2023]009), Guiyang Science and Technology Plan Project (Construction Technology Contract [2023] 48–21),

Integration and Demonstration of Quality Improvement and Efficiency Enhancement Technology for Wangmo Babu Tea (2021YFD1100307), and the Science and Technology Plan Project of Guizhou Province, in RP China ([2021] General 126) for design of the study and data analysis.

Data availability

The plant materials are growing in our resource nursery and are available from the corresponding author on reasonable request. The raw sequence data reported in this study have been deposited in the Genome Sequence Archive [79] in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number CRA001438 that is publicly accessible at http://bigd.big.ac.cn/gsa.

Declarations

Ethics approval and consent to participate

The collecting of these materials is allowed by the Convention on the Trade in Endangered Species of Wild Fauna and Flora and Regulations of Guizhou Province on the protection of ancient tea plants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 6 July 2023 / Accepted: 23 January 2024 Published online: 30 January 2024

References

- Shen S, Xu G, Li D, Clements DR, Zhang F, Jin G, et al. Agrobiodiversity and in situ conservation in ethnic minority communities of Xishuangbanna in Yunnan Province, Southwest China. J Ethnobiol Ethnomed. 2017;13(1):1–15.
- Liu U, Gianella M, Dávila Aranda P, Diazgranados M, Flores Ortíz CM, Lira-Saade R, et al. Conserving useful plants for a sustainable future: species coverage, spatial distribution, and conservation status within the millennium seed Bank collection. Biodivers Conserv. 2023;32:2791–839.
- Bolton RL, Mooney A, Pettit MT, Bolton AE, Morgan L, Drake GJ, et al. Resurrecting biodiversity: advanced assisted reproductive technologies and biobanking. Reprod Fertil. 2022;3(3):R121–46.
- Marselle MR, Lindley SJ, Cook PA, Bonn A. Biodiversity and health in the urban environment. Curr Environ Health Rep. 2021;8(2):146–56.
- Mace GM, Reyers B, Alkemade R, Biggs R, Chapin FS, Cornell SE, et al. Approaches to defining a planetary boundary for biodiversity. Global Environ Chang. 2014;28:289–97.
- Xin T, Huang W, De Riek J, Zhang S, Ahmed S, Van Huylenbroeck J, et al. Genetic diversity, population structure, and traditional culture of *Camellia reticulata*. Ecol Evol. 2017;7(21):8915–26.
- Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, et al. Biodiversity loss and its impact on humanity. Nature. 2012;486(7401):59–67.
- Garcia RA, Cabeza M, Rahbek C, Araujo MB. Multiple dimensions of climate change and their implications for biodiversity. Science. 2014;344(6183):1247579.
- Tisdell CA. Human values and biodiversity conservation: the survival of wild species. Cheltenham: Edward Elgar Publishing; 2014.
- Isbell F, Gonzalez A, Loreau M, Cowles J, Diaz S, Hector A, et al. Linking the influence and dependence of people on biodiversity across scales. Nature. 2017;546(7656):65–72.
- Bohn JL, Diemont SAW, Gibbs JP, Stehman SV, Mendoza Vega J. Implications of mayan agroforestry for biodiversity conservation in the Calakmul Biosphere Reserve, Mexico. Agrofor Syst. 2014;88(2):269–85.
- 12. Vigouroux Y, Barnaud A, Scarcelli N, Thuillet AC. Biodiversity, evolution and adaptation of cultivated crops. C R Biol. 2011;334(5–6):450–7.
- Xin T, de Riek J, Guo H, Jarvis D, Ma L, Long C. Impact of traditional culture on Camellia reticulata in Yunnan, China. J Ethnobiol Ethnomed. 2015;11(1):1–11.
- Lei QY, Zhou JJ, Xiong Y, Zhang WH, Luo J, Long CL. Genetic diversity evaluation and conservation of Kam fragrant glutinous rice (*Oryza sativa* L.) germplasm in southeast Guizhou, China. Plants. 2021;10(9):1898.

- Labeyrie V, Deu M, Barnaud A, Calatayud C, Buiron M, Wambugu P, et al. Influence of ethnolinguistic diversity on the sorghum genetic patterns in subsistence farming systems in eastern Kenya. PLoS ONE. 2014;9(3):e92178.
- Westengen OT, Okongo MA, Onek L, Berg T, Upadhyaya H, Birkeland S, et al. Ethnolinguistic structuring of sorghum genetic diversity in Africa and the role of local seed systems. Proc Natl Acad Sci U S A. 2014;111(39):14100–5.
- Deletre M, McKey DB, Hodkinson TR. Marriage exchanges, seed exchanges, and the dynamics of manioc diversity. Proc Natl Acad Sci U S A. 2011;108(45):18249–54.
- Luo B, Li F, Ahmed S, Long C. Diversity and use of medicinal plants for soup making in traditional diets of the Hakka in West Fujian, China. J Ethnobiol Ethnomed. 2019;15(1):60.
- Liu C, Cui D, Jiao A, Ma X, Li X, Han B, et al. Kam Ssweet rice (*Oryza sativa* L.) is a special ecotypic rice in southeast Guizhou, China as revealed by genetic diversity analysis. Front Plant Sci. 2022;13:830556.
- 20. Liu Y, Ahmed S, Liu B, Guo Z, Huang W, Wu X, et al. Ethnobotany of dye plants in Dong communities of China. J Ethnobiol Ethnomed. 2014;10(1):1–9.
- Zhang W, Zhang Y, Qiu H, Guo Y, Wan H, Zhang X, et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. Nat Commun. 2020;11(1):3719.
- 22. Willson KC, Clifford MN. Tea: cultivation to consumption. North Yorkshire: Springer Science & Business Media; 2012.
- Tang J, Chen Y, Huang C, Li C, Feng Y, Wang H, et al. Uncovering the complex regulatory network of spring bud sprouting in tea plants: insights from metabolic, hormonal, and oxidative stress pathways. Front Plant Sci. 2023;14:1263606.
- Yue C, Cao H, Hao X, Zeng J, Qian W, Guo Y, et al. Differential expression of gibberellin- and abscisic acid-related genes implies their roles in the bud activity-dormancy transition of tea plants. Plant Cell Rep. 2018;37(3):425–41.
- Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, et al. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci U S A. 2018;115(18):E4151–8.
- Yang C, Sun Z, Jin Y. Construction of ethnic relations in Guizhou. Guiyang: Guizhou People's Publishing House; 2010.
- 27. Niu S, Song Q, Koiwa H, Qiao D, Zhao D, Chen Z, et al. 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. 2019;19(1):328.
- Barrett CB, Travis AJ, Dasgupta P. On biodiversity conservation and poverty traps. Proc Natl Acad Sci U S A. 2011;108(34):13907–12.
- 29. Sage RF. Global change biology: a primer. Glob Chang Biol. 2020;26(1):3-30.
- Liu C, Wang Y, Ma X, Cui D, Han B, Xue D, et al. Traditional agricultural management of kam sweet rice (*Oryza sativa* L.) in southeast Guizhou Province, China. J Ethnobiol Ethnomed. 2022;18(1):30.
- Xia E, Tong W, Hou Y, An Y, Chen L, Wu Q, et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. Mol Plant. 2020;13(7):1013–26.
- 32. Greenbaum G, Templeton AR, Zarmi Y, Bar-David S. Allelic richness following population founding events–a stochastic modeling framework incorporating gene flow and genetic drift. PLoS ONE. 2014;9(12):e115203.
- Zhao Z, Song Q, Bai D, Niu S, He Y, Qiao D, et al. Population structure analysis to explore genetic diversity and geographical distribution characteristics of cultivated-type tea plant in Guizhou Plateau. BMC Plant Biol. 2022;22(1):55.
- Li W, Liu L, Wang Y, Zhang Q, Fan G, Zhang S, et al. Genetic diversity, population structure, and relationships of apricot (*Prunus*) based on restriction site-associated DNA sequencing. Hortic Res. 2020;7:69.
- Sertse D, You FM, Ravichandran S, Soto-Cerda BJ, Duguid S, Cloutier S. Loci harboring genes with important role in drought and related abiotic stress responses in flax revealed by multiple GWAS models. Theor Appl Genet. 2020;134(1):191–212.
- Mourad AMI, Belamkar V, Baenziger PS. Molecular genetic analysis of spring wheat core collection using genetic diversity, population structure, and linkage disequilibrium. BMC Genomics. 2020;21(1):434.
- 37. Kaboré KH, Kassankogno AI, Adreit H, Milazzo J, Guillou S, Blondin L, et al. Genetic diversity and structure of *Bipolaris oryzae* and *Exserohilum rostratum* populations causing brown spot of rice in Burkina Faso based on genotyping-by-sequencing. Front Plant Sci. 2022;13:1022348.
- Islam ASMF, Sanders D, Mishra AK, Joshi V. Genetic diversity and population structure analysis of the USDA olive germplasm using genotyping-bysequencing (GBS). Genes (Basel). 2021;12(12):2007.

- Jiang X, Fang Z, Lai J, Wu Q, Wu J, Gong B, et al. Genetic diversity and population structure of Chinese chestnut (*Castanea mollissima* Blume) cultivars revealed by GBS resequencing. Plants (Basel). 2022;11(24):3524.
- He J, Zhao X, Laroche A, Lu ZX, Liu H, Li Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci. 2014;5:484.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6(5):e19379.
- Pootakham W, Jomchai N, Ruang-Areerate P, Shearman JR, Sonthirod C, Sangsrakru D, et al. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics. 2015;105(5–6):288–95.
- Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, et al. Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3 (Bethesda). 2014;4(9):1681–7.
- Chen L, Yang Y, Yu F. Genetic diversity, relationship and molecular discrimination of elite tea germplasms [*camellia sinensis* (L.), O. Kuntze] revealed by rapd markers. Mol Plant Breeding. 2004;2:385–90.
- Varshney RK, Thudi M, Roorkiwal M, He W, Upadhyaya HD, Yang W, et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nat Genet. 2019;51(5):857–64.
- Zhang D. Exploration of Guizhou folk tea culture. Beijing: China Agricultural Press; 2019.
- 47. Wu S. Research on the geographical names of the Gelao people in the Gelao Culture Research Series. Beijing: China Yanshi Publishing House; 2015.
- 48. Lai Y. The value of Yi's tree culture and its inheritance and development. Chin Agric Sci Bull. 2016;32:26–32.
- 49. Li F, Yan Y. Migration history of ethnic groups living in Guizhou for generations. Guiyang: Guizhou People's Publishing House; 2011.
- 50. Shu G, Cao G, Li N, Wang A, Wei F, Li T, et al. Genetic variation and population structure in China summer maize germplasm. Sci Rep. 2021;11(1):8012.
- Chen H, Jiao C, Wang Y, Wang Y, Tian C, Yu H et al. Comparative population genomics of bread wheat (*Triticum aestivum*) reveals its cultivation and breeding history in China. BioRxiv. 2019;519587.
- 52. Huang F, Duan J, Lei Y, Liu Z, Kang Y, Luo Y, et al. Genetic diversity, population structure and core collection analysis of Hunan tea plant germplasm through genotyping-by-sequencing. Bev Plant Res. 2022;2:5.
- 53. Liu Z, Zhao Y, Yang P, Cheng Y, Huang F, Li S, et al. Population whole-genome resequencing reveals the phylogenetic relationships and population structure of four Hunan typical tea landraces. Bev Plant Res. 2023;3:9.
- 54. Zhu Q. Chinese water system atlas. Qingdao: Qingdao Publishing House; 2010.
- 55. Yan P. Information of Guizhou post road in the Ming Dynasty. Guiyang: Guizhou University Press; 2015.
- Ndjiondjop MN, Semagn K, Gouda AC, Kpeki SB, Dro Tia D, Sow M, et al. Genetic variation and population structure of *Oryza glaberrima* and development of a mini-core collection using DArTseq. Front Plant Sci. 2017;8:1748.
- 57. De Beukelaer H, Davenport GF, Fack V. Core Hunter 3: flexible core subset selection. BMC Bioinformatics. 2018;19(1):203.
- Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJ. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet. 2013;126(2):289–305.
- Wang RJ, Gao XF, Yang J, Kong XR. Genome-wide association study to identify favorable SNP allelic variations and candidate genes that control the timing of spring bud flush of tea (*Camellia sinensis*) using SLAF-seq. J Agric Food Chem. 2019;67(37):10380–91.
- Guizhou Provincial Water Resources Bulletin 2000–2020. Guizhou provincial department of water resources, guiyang. 2022. http://mwr.guizhou.gov.cn/ sjfb/slsj/202211/t20221117_77132741.html. Accessed 14 November 2022.
- Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 2016;12(2):e1005767.
- 62. Csiszar J, Galle A, Horvath E, Dancso P, Gombos M, Vary Z, et al. Different peroxidase activities and expression of abiotic stress-related peroxidases in apical root segments of wheat genotypes with different drought stress tolerance under osmotic stress. Plant Physiol Biochem. 2012;52:119–29.
- Deslauriers SD, Larsen PB. FERONIA is a key modulator of brassinosteroid and ethylene responsiveness in *Arabidopsis* hypocotyls. Mol Plant. 2010;3(3):626–40.

- 65. Stepanova AN, Robertson-Hoyt J, Yun J, Benavente LM, Xie DY, Dolezal K, et al. *TAA1*-mediated auxin biosynthesis is essential for hormone crosstalk and plant development. Cell. 2008;133(1):177–91.
- Fu L, Liu Z, Bai Y. Advances in the RNA-directed DNA methylation in plants. Chin J Biotechnol. 2006;22:891–6.
- 67. Min T. A revision of *Camellia* sect. Thea Acta Bot Yunnanica. 1992;14(2):115–32.
- Chen L, Yu F, Tong Q. Discussions on phylogenetic classification and evolution of sect. Thea J Tea Sci. 2000;20(2):89–94.
- 69. Jung Y, Han D, BWA-MEME:: BWA-MEM emulated with a machine learning approach. Bioinformatics. 2022;38(9):2404–13.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a map reduce framework for analyzing nextgeneration DNA sequencing data. Genome Res. 2010;20(9):1297–303.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–9.

- Zhao YP, Fan G, Yin PP, Sun S, Li N, Hong X, et al. Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. Nat Commun. 2019;10(1):4201.
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE. 2014;9(2):e90346.
- Stecher G, Tamura K, Kumar S. Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol. 2020;37(4):1237–9.
- Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8.
- Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods. 2001;25(4):402–8.
- 79. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: genome sequence archive. Genom Proteom Bioinf. 2017;15(1):14–8.

Publisher's Note

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