BMC Plant Biology

RESEARCH

Open Access

Genome-wide association study of leaf-related traits in tea plant in Guizhou based on genotyping-by-sequencing

Yanjun Chen[1], Suzhen Niu[1,2*], Xinyue Deng[3], Qinfei Song[1], Limin He[1], Dingchen Bai[1] and Yingqin He[1]

Abstract

Background Studying the genetic characteristics of tea plant (*Camellia* spp.) leaf traits is essential for improving yield and quality through breeding and selection. Guizhou Plateau, an important part of the original center of tea plants, has rich genetic resources. However, few studies have explored the associations between tea plant leaf traits and single nucleotide polymorphism (SNP) markers in Guizhou.

Results In this study, we used the genotyping-by-sequencing (GBS) method to identify 100,829 SNP markers from 338 accessions of tea germplasm in Guizhou Plateau, a region with rich genetic resources. We assessed population structure based on high-quality SNPs, constructed phylogenetic relationships, and performed genome-wide association studies (GWASs). Four inferred pure groups (G-I, G-II, G-III, and G-IV) and one inferred admixture group (G-V), were identified by a population structure analysis, and verified by principal component analyses and phylogenetic analyses. Through GWAS, we identified six candidate genes associated with four leaf traits, including mature leaf size, texture, color and shape. Specifically, two candidate genes, located on chromosomes 1 and 9, were significantly associated with mature leaf size, while two genes, located on chromosomes 8 and 11, were significantly associated with mature leaf texture. Additionally, two candidate genes, located on chromosomes 1 and 2 were identified as being associated with mature leaf color and mature leaf shape, respectively. We verified the expression level of two candidate genes was verified using reverse transcription quantitative polymerase chain reaction (RT-qPCR) and designed a derived cleaved amplified polymorphism (dCAPS) marker that co-segregated with mature leaf size, which could be used for marker-assisted selection (MAS) breeding in *Camellia sinensis*.

Conclusions In the present study, by using GWAS approaches with the 338 tea accessions population in Guizhou, we revealed a list of SNPs markers and candidate genes that were significantly associated with four leaf traits. This work provides theoretical and practical basis for the genetic breeding of related traits in tea plant leaves.

Keywords Tea plant, Leaf traits, Genome-wide association study (GWAS), Single nucleotide polymorphism (SNP), Candidate genes, Genotyping-by-sequencing (GBS)

*Correspondence:
Suzhen Niu
niusuzhen@163.com
Full list of author information is available at the end of the article

Background

Tea, made from the fresh and soft leaves of tea plants, has become a popular healthy beverage worldwide with high nutritional and medicinal value [1–3]. Tea plants are highly heterozygous and cross-pollinated. Conventional breeding approaches in tea depend on phenotypic selection and have been time-consuming because of tea's relatively long juvenile stage. The mature leaf morphological traits of tea plants, such as leaf size, texture, shape, and color, are used regularly to determine genetic diversity among germplasm accessions, which play important roles in botanical classification, and tea yield and quality. The Guizhou Plateau, an important part of the original center of tea plants, has rich tea germplasms with strong genetic variation in morphological traits, especially in the leaf traits because of the three-dimensional ecological environment [4, 5]. Although some studies have addressed the leaf traits of tea germplasm in the Guizhou Plateau [6, 7], the genetic architecture and molecular bases of numerous leaf traits remain largely unknown. The functional genes of the leaf trait and the molecular markers in the functional genes have not yet been identified.

The tea plant is a major economic crop with buds and tender leaves as its product organ. The leaf traits are among the most important agronomic traits of the tea plant. The leaf size, texture, shape, and color play an important role in the yield, appearance, and quality of dried tea [8–10]. Breeding high-yield and suitable shape cultivars is a valuable goal for tea plant breeders. The variation of leaf traits is a complex mechanism arising from plants adapting to their environment. External factors (such as environment temperature, nutrient condition, and water supply) and internal factors (including cell division and cell expansion) contribute to the final leaf traits [11, 12]. Our comprehension of the genetic determinants for the mature leaf trait in tea plants remains primarily at the level of the genes found and their expression. Although some quantitative trait loci (QTLs) and single nucleotide polymorphisms (SNPs) associated with mature leaf size have been reported, the finding of genes and the development of functional molecular markers related to leaf traits remains insufficient for breeding [6, 7, 13, 14].

QTL mapping can be an effective tool for identifying genes underlying major leaf traits [15, 16]. Nevertheless, accurate QTL mapping depends on a large enough mapping population, which is not only a time-consuming process because of the long juvenile stage but the QTLs captured are limited because of the use of biparental hybridization in tea plants [2, 17]. Genome-wide association studies (GWASs), an association mapping method, have become a powerful tool for identifying multiple-related candidate genes that regulate major crop traits. GWASs are based on obtaining the genotype data of the research population and using high-density SNP markers with extensive genetic variation to predict the number of causative genes. It is based on the natural population to perform association analysis. The natural population has naturally hybridized for many generations and has sufficient recombination. This gives GWASs higher positioning accuracy [18]. GWASs have been successfully carried out in crops [19–21] and woody plants [22, 23]. Wang et al. [2] carried out GWAS of the timing of Spring Bud flush (TBF) on 151 tea plant germplasm resources, identified 26 SNPs and "Thymine. Thymine." favorable allele variants, and designed a cleavage amplification polymorphism (dCAPS) marker associated with TBF. Yamashita et al. [24] conducted a GWAS on five metabolites related to tea quality from 150 tea plant materials and identified potential candidate genes controlling epicatechin (57 genes), epicatechin gallate (97 genes), epigallocatechin gallate (64 genes), total catechin (80 genes), and caffeine (83 genes). However, few studies have addressed the association analysis of leaf characteristics using the diverse tea germplasms and GBS sequencing method in Guizhou Province.

Molecular markers are essential for breeding major crops today, and many molecular marker techniques have been developed [25–27]. Among them, restriction fragment length polymorphism (RFLP), simple sequence repeat (SSR), amplified fragment length polymorphism (AFLP), and single nucleotide polymorphisms (SNPs) are the most effective marker systems for the determination of polymorphism, which are functionally responsible for specific traits or used to trace the evolutionary history of a species. SNPs are the most common source of genetic variation in eukaryotic species and have become an important marker for the genetic studies of plants. For example, 3.6 million SNPs were identified on 517 rice varieties using next-generation sequencing (NGS) with $19\times$ coverage. Taranto et al. [28] performed the genotyping-by-sequencing (GBS) for the genome-wide identification of SNPs in a collection of Capsicum spp. accessions. The results showed that 32,950 high-quality SNPs were identified on 222 C. annuum accession. Niu et al. [4] identified 79,016 high-quality SNPs at the genome level of 415 tea accessions using the GBS method.

The reduced cost and rapid progress in NGS and related bioinformatics resources promoted the large-scale discovery of SNPs in manly plants. GBS is the cost-effective NGS method to simultaneously discover and score segregating markers in populations of

interest. GBS holds the potential to close the genotyping gap between references of broad interest and mapping/breeding populations of local or specific interest. The multiplexing of samples in GBS protocols keeps molecular biology costs low while the resultant next-generation sequencing data has immediate applications in many different research areas, ranging from gene discovery to genomic-assisted breeding [29].

In this study, 338 tea accessions were used from the Guizhou Plateau to identify high-quality SNPs for analyzing the population structure, linkage disequilibrium, and genetic diversity. The phenotype of the mature leaf trait and the genotype of 338 tea accessions were used to find the candidate genes and develop the molecular markers by GWAS. Our results provide useful information for future molecular marker-assisted breeding.

Table 1 Statistical analysis of mature leaf size (MLZ) in the three years

Year	Min (c m ^{2})	Max (c m ^{2})	Mean ± SD (c m ^{2})	CV (%)
2019	4.37	54.17	23.16 ± 11.24	48.50
2020	4.14	56.71	23.51 ± 10.95	46.60
2021	6.14	55.34	23.41 ± 10.87	46.40

Results

Statistical analysis of phenotype variation

To further understand the differences between different individuals for the same trait, we investigated four leaf traits of 338 tea plants in three years, including one quantitative trait (MLZ) and three qualitative traits (MLC, MLT, and MLS) [30]. We found that the four leaf traits of 338 tea plants displayed broad variation. For example, in the three years, the variation range of MLZ was 4.37 to 54.17 c m ^{2}, 4.14 to 56.71 c m ^{2}, and 6.14 to 55.34 c m ^{2}, respectively, while the CV of MLZ were 48.5%, 46.6%, and 46.4%, respectively (Table 1). The 3-year frequency distribution of MLZ showed a normal distribution based on the best linear unbiased prediction (BLUP) (Fig. 1A-C). According to the descriptors for tea germplasm resources (NY/T 2943–2016), the three qualitative traits (MLC, MLT, and MLS) were mainly displayed by morphological characteristics in frequency distribution (Fig. 1D-F). The phenotypic characteristics of the three qualitative traits identified in this study were MLC (yellow green, light green, green and dark green), MLT (soft, medium and hard), and MLS (round, oval, long oval and lanceolate), which respectively account for 100%, 100% and 80% of the corresponding phenotypic characteristics in descriptors for tea germplasm resources (NY/T 2943–2016)

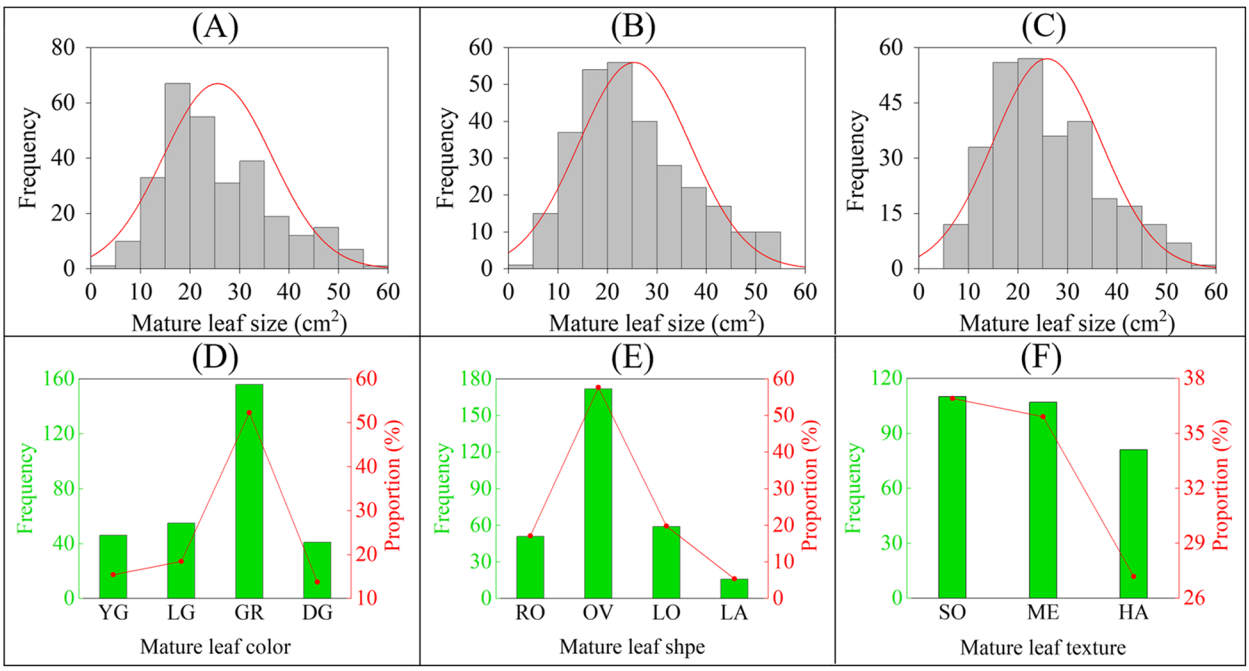

Fig. 1 The frequency distribution of MLZ (in three years) and three qualitative traits (MLC, MLS, MLT), the frequency distribution of the three qualitative traits are mainly displayed by morphological characteristics. **A** Frequency distribution of MLZ in 2019. **B** Frequency distribution of MLZ in 2020. **C** Frequency distribution of MLZ in 2021. **D** Frequency distribution of MLC. Y.G.: yellow-green; L.G.: light green; G.R.: green; D.G.: dark green. **E** Frequency distribution of MLS. R.O.: round; O.V.: oval; L.O.: long oval; L.A.: lanceolate. **F** Frequency distribution of MLT. SO: Soft; ME: Medium; HA: Hard

Table 2 Phenotypic identification statistical analysis of three qualitative traits

Traits	K		K′		PCP%	
Mature leaf color	4	11	4	12	100	91.67
Mature leaf texture	3		3		100	
Mature leaf shape	4		5		80	

K Phenotypic types identified, *K′* Phenotype types in NY/T 2943–2016, *PCP% = K/K′* The proportion of phenotypic types determined in this study which in descriptors for tea germplasm resources (NY/T 2943–2016)

Table 3 Variance analysis and broad-sense heritability (h_B^2) for mature leaf size (MLZ) of the 338 tea accessions

Source of variation	DF	ANOVA SS	Mean Square	F value	h_B^2
genotype (G)	287	104,095.08	356.70	127.14**	97.68%
environment (E)	2	25.71	13.05	4.65**	
error	574	1610.42	2.81		

** Represent $p < 0.05$

(Table 2). These results suggested a broad diversity of the four leaf traits in 338 tea plants population.

The effects of genotype (G) and environment (E) on the MLZ of 338 accessions in the three years were studied by ANOVA. The difference in MLZ among 338 genotypes and within three years was significant ($p < 0.05$), indicating that the environment extensively influenced MLZ. The h_B^2 of MLZ was evaluated as 97.68%, and the stable inheritance was higher than 90% (Table 3).

Sequencing and variant discovery

GBS was performed on 338 tea accessions using Illumina HiSeq X ten. After the filtering step, 217 Gb clean data were generated with an average of 0.64 Gb per accession, and an average coverage depth was about 10 × (Table S1). Moreover, all clean reads were mapped on the released genome of the C. sinensis var. sinensis to identify the high-quality SNPs. In this study, a total of 29,393,327 SNPs were obtained from the 338 tea accessions, and 100,829 high-quality SNPs with a minor allele frequency (MAF) > 0.05 and missing data rate < 20% were conserved for subsequent analysis (Tables S2 and S3). Moreover, the distribution of high-quality SNPs was further investigated, and the results showed that 100,829 high-quality SNPs were roughly evenly distributed in 15 chromosomes of the tea genome (Fig. 2, Table S4). Chromosome 1 contained the largest number of high-quality SNPs (9140 SNPs), followed by chromosome 4 (7613 SNPs), and chromosome 15 had the smallest number of high-quality SNPs (4140 SNPs) (Table 4). The average

Fig. 2 Distribution of single nucleotide polymorphisms (SNPs) on 15 chromosomes of the tea plant. The horizontal axis shows the chromosome length

Table 4 SNPs density of chromosome based on GBS

Chromosome	Length(bp)	SNP	Density(kb)
1	222,690,619	9140	41.04
2	212,893,553	7355	34.55
3	187,305,764	6286	33.56
4	196,128,943	7613	38.82
5	194,778,845	6165	31.65
6	180,870,292	6642	36.72
7	187,027,217	6354	33.97
8	162,894,436	6889	42.29
9	165,397,305	5645	34.13
10	166,753,223	5624	33.73
11	123,582,105	4771	38.61
12	162,686,597	5563	34.19
13	136,080,369	5054	37.14
14	133,792,454	5007	37.42
15	119,038,877	4140	34.78
Contig	2,540,239,288	8581	38.70
Total	3,021,230,785	100,829	38.24

number of high-quality SNPs on each chromosome was 6150. The lowest (32 SNPs/Mb) and highest (42 SNPs/Mb) SNP densities were detected on chromosomes 5 and chromosomes 8, respectively (Table 4).

Genetic diversity estimation

Nucleotide diversity (*Pi*), minor allele frequency (*MAF*), observed heterozygosity (*Ho*), expected heterozygosity (*He*), and inbreeding coefficient (*Fis*) were used as indicators of genetic diversity. The *Pi*, *MAF, Ho*, *He* and *Fis* of 338 tea accession populations were 0.2268, 0.1448, 0.0721 0.2264 and 0.6936, respectively (Table 5). We compared the genetic diversity of five inferred populations of 338 Guizhou tea accession. *Pi, Ho, He* and *MAF* of G-V tea population were significantly higher than those of other tea populations, while *Pi*, *Ho*, *He* and *MAF* of G-III tea population were lower than those of other tea populations. *Fis* was higher for the tea population in G-III was higher than that of other tea populations (Table 5).

Previous studies showed that when a positive Tajima's D value was determined for a population, the population was in a population bottleneck and/or balancing selection [31, 32]. The positive Tajima's D values of these five tea populations here suggest that they all underwent population bottlenecks and/or balancing selection (Table 5).

Population structure, PCA, and Phylogenetic analysis

To further explore the relationship of 338 tea plant populations, a total of 100,829 high-quality SNPs were used to perform the population structure analysis, PCA, and phylogenetic analysis. Dynamic changes in population structure were further evaluated under different K values (K = 2–9) (Figure. S1). Analysis of cross-validation error (CV error) under different K values revealed that the CV error was the smallest when K was equal to 4 (Fig. 3D). The membership coefficient 0.8 was used to distinguish pure ancestral and admixture groups. Accessions with the membership coefficient > 0.80 were assigned to the corresponding ancestral groups. Those with the membership coefficient < 0.80 were assigned to the admixture group (Fig. 3B). The 338 tea plants populations were further classified into five groups, four ancestral groups and one admixture group. The first ancestral group ('*C. tachangensis* group' or 'G-I') contained 34 *Camellia tachangensis* accessions (Tables S5 and S6). The second ancestral group (referred to as 'Modern landraces group' or 'G-II' from now on) contained 24 accessions, including 23 *C. sinensis* accessions mainly from the modern landraces (82.6%) and one *C. tachangensis* plant (Tables S5 and S6). The third ancestral group (referred to as 'Transitional group' or 'G-III' from now on) contained three *C. remotiserrata* plants and three *C. tachangensis* plants (Tables S5 and S6). The fourthancestral group (referred to as 'Ancient landrace group or 'G-IV' from now on) contained 113 tea accessions, including 108 *C. sinensis* plants (91 tea accessions from the ancient landrace and 17 accessions from the modern landrace) and five *C. remotiserrata* plants (Tables S5 and S6). The admixture group (G-V)

Table 5 Genetic diversity parameters of 338 tea accessions in Guizhou Plateau

Group	Number	Tajima's D	Pi	Ho	He	MAF	Fis
G-I	34	0.8588	0.1508c	0.0647c	0.1479c	0.1090c	0.5984a
G-II	24	0.2908	0.1402d	0.0596d	0.1340d	0.0932d	0.6287a
G-III	6	0.4549	0.1373e	0.0565e	0.1257e	0.0923d	0.6553a
G-IV	113	0.5016	0.1754b	0.0671b	0.1745b	0.1158b	0.5580a
G-V	161	0.8812	0.2254a	0.0814a	0.2245a	0.1448a	0.4724a
All	338	1.2618	0.2268	0.0721	0.2264	0.1448	0.6936

Pi nucleotide diversity, *Ho* observed heterozygosity, *He* expected heterozygosity, *MAF* minor allele frequency, *Fis* inbreeding coefficient; In the same type and line, the different letters indicate a significant difference in $p = 0.05$ levels by the T-test

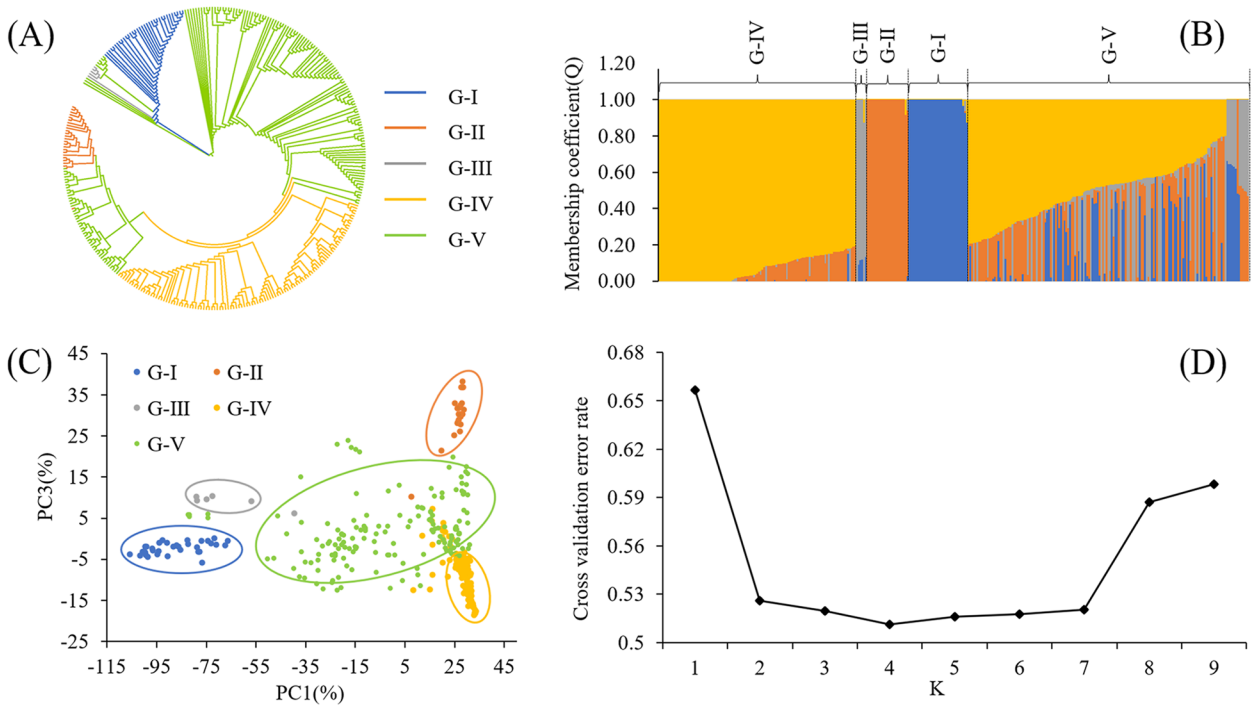

Fig. 3 Population structure analysis, principal component analysis (PCA), and phylogenetic tree establishment of the 338 tea accessions. **A** The neighbor-joining tree was established by 100,829 high-quality SNPs. **B** Population structure separates the accessions into four subgroups (K = 4). **C** PCA of 338 tea accessions. And first and third components of the PCA analyses were shown; each dot represents an individual. **D** When K = 4, the CV error value was the smallest (0.51127)

had 161 tea accessions, including 44 wild *C. remotiserrata* plants, 50 wild *C. tachangensis* plants, 63 cultivated *C. sinensis* plants, and four uncertain wild species (Tables S5 and S6).

To explore the cluster relationships and verify the stability of the potential population structure, principal component analysis (PCA) and Neighbor-Joining tree (NJ tree) was carried out on 100,829 SNPs of 338 tea accessions. According to the topology of the phylogenetic tree, the 338 tea accessions were mainly clustered into five groups: Group I (G-I), Group II (G-II), Group III (G-III), Group IV (G-IV), and Group V (G-V) (Fig. 3A). Among them, G-I contained the tea plants from Bijie, Zunyi, Qiannan Buyi and Miao, and Southwest Guizhou Autonomous Prefectures; G-II contained tea plants from Bijie, Guiyang, Liupanshui, Tongren, and Zunyi, Qiannan Buyi and Miao Autonomous Prefecture; G-III contained the members of tea accessions all from Zunyi City; G-IV contained tea accessions from Aushun, Bijie, Guiyang, Liupanshui, Tongren, Zunyi, Qiandongnan Miao and Dong, Qiannan Buyi and Miao, and Southwest Guizhou Autonomous Prefectures; G-V contained tea accessions from Guiyang, Tongren, Bijie, Zunyi, Qiandongnan Miao and Dong, and Southwest Guizhou Autonomous Prefectures. PCA disclosed five groups corresponding to G-I, G-II, G-III, G-IV and G-V (Fig. 3C). Therefore, the accuracy of the population structure was mutually verified.

Linkage disequilibrium analysis

Linkage disequilibrium data can be used to detect associations between common variants and important traits. Linkage disequilibrium (LD, indicated by r^2) analysis indicated that the 338 tea plants' genomes had a relatively short r^2 distance (~7 kb) and rapid r^2 decay (Fig. 4). The r^2 decreased to half its maximum value, at 7 kb. Therefore, we consider that 7 k b can be used as a search range of late candidate genes. As the average marker density was 30.7 kb (3.1 g/100,829) per SNP, we conclude that these selected SNPs were sufficiently dense to perform genome-wide association study.

Genome-wide association study and candidate genes prediction

Plant domestication conducted over several millennia had modified specific plant traits, especially leaf traits. To further identify the molecular mechanism involved in regulating four leaf traits, six linear regression models (GLM-Q, GLM-P, MLM-Q + K, MLM-P + K, cMLM-Q + K, and cMLM-P + K) were used to perform the GWAS using TASSEL5.0 software (Figure. S2). Comparing

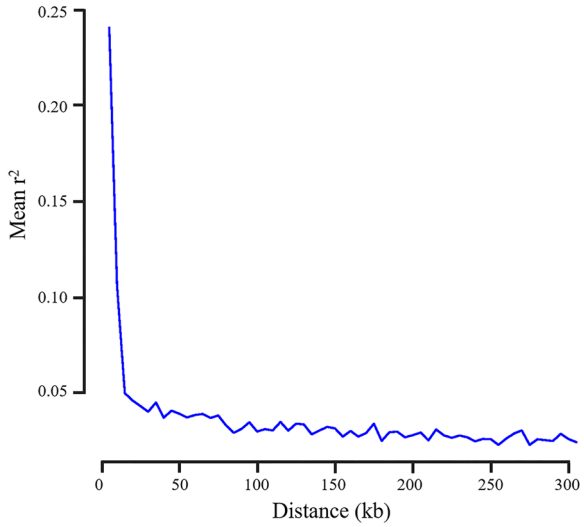

Fig. 4 L.D. decay of 338 tea accessions

the Q-Q plots of the output of the six GWAS models, the best model fitting the curve of the expected value to the observed value was entered into the later analysis as the optimal model for each trait. Our result showed that cMLM-P + K, GLM-Q, GLM-P, and cMLM-P + K models were suitable for GWAS analysis for MLZ, MLC, MLT, and MLS, respectively (Fig. 5). A total of 59 high-quality SNPs were significantly associated with four leaf traits ($-\log_{10}^{(P)} \geq 4.0$) (Tables 6 and 7). Among them, 9, 18, and 10 high-quality SNPs were significantly associated with MLC, MLT, and MLS, respectively (Table 6). And 22 high-quality SNPs were significantly associated with MLZ in three years (Table 7). Only one SNP (P-1076408) of nine high-quality SNPs was discovered in the coding region for MLC. Functional annotation of the coding sequence where the SNP is located revealed that the TEA027527.1 gene, encoding RAB geranylgeranyl transferase type 2 subunit β1, was involved in the biosynthesis of carotenoids and chlorophyll based on gene function annotation and KEGG analysis (Table 8). For MLT, ten SNPs of 18 high-quality SNPs were located in the coding region. Functional annotation of the coding sequence where the ten SNPs were located revealed that three SNPs

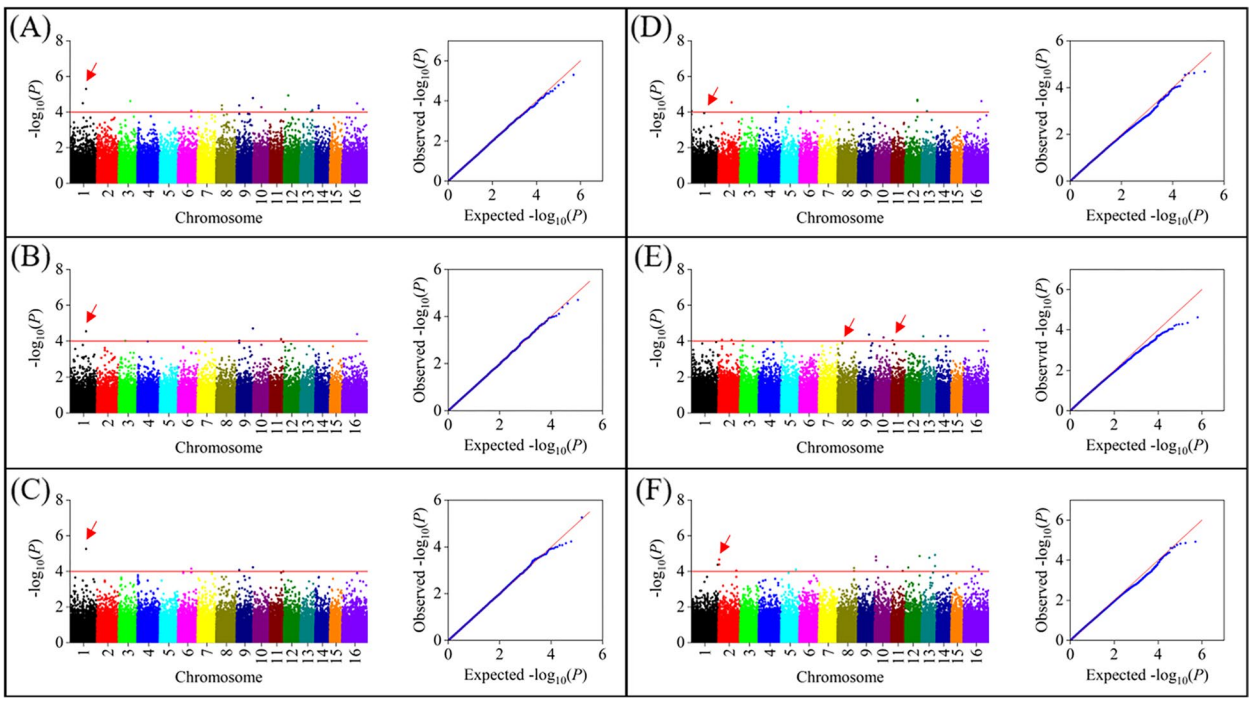

Fig. 5 Manhattan plots and Q-Q plots of the 3-years (2019, 2020, 2021) cMLM-P + K model for MLZ traits of the tea plant. **A** MLZ-cMLM-P + K-2019; **B** MLZ-cMLM-P + K-2020; **C** MLZ-cMLM-P + K-2021; The red dashed horizontal line indicates the significance threshold ($-\log_{10}^{(P)}$ is about equal to 4.0); The red arrows represent SNPS that have been jointly identified for MLZ traits in three years; "16" represents Contig (A continuous long DNA sequence overlapping fragment formed by joining short DNA fragments in the process of genome sequencing.) in Manhattan plots. Manhattan plots and Q-Q plots of the 1-year optimal model for three qualitative traits (MLC, MLS, MLT). **D** MLC-GLM-Q; **E** MLS-cMLM-P + K; (F) MLT-GLM-P; The red dashed horizontal line indicates the significance threshold ($-\log_{10}^{(P)}$ is about equal to 4.0); The red arrows represent the ultimately significant SNPS associated with the three traits; "16" represents Contig (A continuous long DNA sequence overlapping fragment formed by joining short DNA fragments in the process of genome sequencing.) in Manhattan plots

Table 6 The optimal model of three qualitative traits of 338 tea accessions was used for SNPs analysis

Traits	Model	SNPs	Chromosome	Position	$-Log_{10}^{(p)}$	Allele	R^2(%)
MLC	GLM(Q)	S1-P1076408	1	1,076,408	4.063	T/G	6.62
		S2-P128618813	2	128,618,813	4.543	G/A	7.43
		S5-P87525173	5	87,525,173	4.302	C/T	7.49
		S6-P112505345	6	112,505,345	4.018	A/G	7.65
		S6-P18460797	6	18,460,797	4.026	A/G	6.59
		S12-P108970717	12	108,970,717	4.621	A/G	8.56
		S12-P108970772	12	108,970,772	4.682	T/C	8.74
		S13-P38603004	13	38,603,004	4.056	G/A	7.31
		SCONTIG614-P349959	CONTIG614	349,959	4.610	G/A	7.71
MLT	GLM(P)	S1-P211517178	1	211,517,178	4.365	G/A	7.90
		S2-P11825464	2	11,825,464	4.655	G/A	8.36
		S2-P11825496	2	11,825,496	4.381	G/T	7.84
		S2-P171922400	2	171,922,400	4.046	G/A	7.66
		S5-P162371923	5	162,371,923	4.103	C/T	5.66
		S8-P129108645	8	129,108,645	4.020	A/G	6.53
		S8-P129108677	8	129,108,677	4.185	G/A	6.94
		S10-P119785042	10	119,785,042	4.252	G/A	7.53
		S10-P13082621	10	13,082,621	4.825	G/A	8.86
		S10-P13082689	10	13,082,689	4.609	C/A	8.36
		S11-P101072100	11	101,072,100	4.014	G/A	7.29
		S12-P32519734	12	32,519,734	4.214	A/G	7.29
		S12-P136348822	12	136,348,822	4.861	A/G	8.93
		S13-P55838677	13	55,838,677	4.758	T/C	7.92
		S13-P122895304	13	122,895,304	4.921	A/C	8.46
		S13-P122895559	13	122,895,559	4.302	A/G	7.30
		SCONTIG18-P388182	CONTIG18	388,182	4.263	A/G	6.98
		SCONTIG446-P62289	CONTIG446	62,289	4.108	A/T	7.38
MLS	cMLM(P + K)	S2-P40432827	2	40,432,827	4.062	G/T	7.21
		S2-P130091909	2	130,091,909	4.050	A/C	7.36
		S3-P30907968	3	30,907,968	4.028	T/C	7.50
		S9-P110799549	9	110,799,549	4.354	A/G	7.39
		S10-P87179990	10	87,179,990	4.201	C/T	7.61
		S11-P12773211	11	12,773,211	4.033	T/A	6.80
		S13-P6815607	13	6,815,607	4.256	G/A	7.77
		S14-P34150418	14	34,150,418	4.272	A/T	7.50
		S14-P107071402	14	107,071,402	4.282	G/A	7.42
		SCONTIG736-P863256	CONTIG736	863,256	4.616	G/A	8.55

(P-129108645, P-129108677, and P-101072100) associated with MLT were distributed in coding sequences of TEA021901.1 gene contained two SNPs (P-129108645 and P-129108677) and the TEA002469.1 gene (P-101072100), respectively. And they encoded Beta-glucosidase 12 based on gene function annotation of the TPIA and NCBI databases (Table 8). For MLZ, 22 high-quality SNPs were identified in the three years. Nine SNPs (P-122201532, P-116917717, P-124718592, P-163812618, P-100806529, P-11244003, P-92754559, P-110363923, and P-110363973) of 22 high-quality SNPs were distributed in the coding region. Functional annotation of the coding sequence where the nine SNPs were located revealed that two SNPs (P-122201532, P-163812618) were present in the coding sequences of the TEA005350.1 and TEA029641.1 genes, respectively. These two genes encode the TONSOKU protein and Pyrophosphate fructose 6-phosphate 1-phosphate transferase subunit α based on gene function annotation of TPIA and NCBI database (Table 8). For MLS, five SNPs (P-3090796, P-130091909, P-110799549, P-12773211, and

Table 7 SNP loci were significantly associated with mature leaf size (MLZ) using cMLM-P + K in the three years

SNPs	Chromosome	Position	$-Log_{10}{}^{P}$	R^2(%)	Allele	Year		
						2019	2020	2021
S1-P96765712	1	96,765,712	4.49	7.85	A/G	√		
S1-P122201532	1	122,201,532	5.30	9.59	T/G	√	√	√
			4.54	8.30				
			5.25	9.74				
S3-P116917717	3	116,917,717	4.62	8.75	C/T	√		
S3-P69973921	3	69,973,921	4.00	9.00	C/T		√	
S6-P124718592	6	124,718,592	4.08	7.27	G/A	√		√
			4.16	7.74				
S7-P15417610	7	15,417,610	4.02	7.86	G/A	√		
S8_P52030704	8	52,030,704	4.37	8.67	G/A	√		
S8_P52030778	8	52,030,778	4.16	8.47	G/A	√		
S9-P27970985	9	27,970,985	4.38	8.62	C/T	√		√
			4.75	8.36				
S9-P163812618	9	163,812,618	4.78	8.82	A/G	√	√	
			4.70	8.77				
			4.23	7.97				
S9-P27970980	9	27,970,980	4.02	7.79	G/C		√	√
			4.07	7.89				
S10-P78788354	10	78,788,354	4.28	8.61	G/A	√		
S11-P100806529	11	100,806,529	4.10	8.64	G/A		√	
S12-P11244003	12	11,244,003	4.16	7.72	C/T	√		
S12-P38652091	12	38,652,091	4.94	9.50	T/C	√		
S13-P92754559	13	92,754,559	4.02	6.85	A/C	√		
S13-P110363923	13	110,363,923	4.11	7.58	T/C	√		
S13-P110363973	13	110,363,973	4.04	7.46	G/A	√		
S14_P31141612	14	31,141,612	4.36	8.82	G/A	√		
S14_P31141847	14	31,141,847	4.21	8.75	G/C	√		
SCONTIG419-P527748	CONTIG419	527,748	4.48	9.33	C/T	√	√	
			4.38	9.79				
SCONTIG758-P387428	CONTIG758	387,428	4.12	7.68	A/T	√		

Table 8 Functional annotation of candidate genes significantly associated with four leaf traits

Traits	SNP locus	Chromosome	Candidate gene	Function annotation
MLZ	S1_P122201532	1	TEA005350.1	Tetratripeptide repeat
	S9_P163812618	9	TEA029641.1	PFP-α: pyrophosphate fructose 6-phosphate 1-phosphate transferase subunit α
MLC	S1_P1076408	1	TEA027527.1	RAB geranyl geranyl transferase type 2 subunit β1
MLT	S8_P129108645	8	TEA021901.1	1,3-beta-glucan synthase
	S11_P101072100	11	TEA002469.1	Beta-glucosidase 12
MLS	S2_P130091909	2	TEA026128.1	Protein containing lob (lateral organ boundary) domain 22

P-107071402) of ten high-quality SNPs were discovered in the coding region. Functional annotation of the coding sequence where the five SNPs were located revealed that only one SNP (P-130091909) was distributed in the coding sequence of the TEA026128.1 gene, which was associated with MLS (Table 8).

RT-qPCR analysis

To determine whether potential candidate genes were involved in the accumulation of leaf size and leaf color, we determined the expression level of the TEA005350.1 gene in three leaf sizes (leaflet, middle leaf, large leaf) and TEA027527.1 gene in four leaf colors (yellow-green, light green, green, dark green) by RT-qPCR. The results showed that the TEA005350.1 gene was differentially expressed in tea plant leaflets, middle leaves, and large leaves. The expression level was highest in leaflets (Fig. 6A, Table S7). Therefore, the TEA005350.1 gene negatively regulates the formation of tea plant leaf size. The expression level of the TEA027527.1 gene in yellow-green leaves was the highest (Fig. 6B, Table S8).

Developing and verifying the dCAPS marker

Since an SNP mutation Chr 1 122,201,532 (T/G) in the coding region of TEA005350.1 was detected as significantly associated with MLZ, we developed a PCR-based dCAPS marker that can be used in molecular marker-assisted breeding. In this study, we introduced a "one base" mismatch in the forward primer to design dCAPS markers and developed a pair of dCAPS marker primers using restriction endonuclease NlaIII, whose digestion site was CATG (Fig. 7A). To verify the reliability of dCAPS markers, we applied the primers to six tea accessions with known MLZ genotypes, and all obtained 216 bp of PCR products (Fig. 7B, Figure S3), which indicates that the dCAPS markers developed can be used for PCR amplification.

Discussion

The plant materials used for the association study should have a wide range of genetic diversity to capture historical recombination events in maximum quantity, so it was extremely important to select plant materials. A core collection is considered a small set representing the maximum diversity of the raw accession collection [33]. Although commercial varieties of tea, including "Fuding Dabaicha" and "Longjing", have been widely planted in southwestern China due to their high yield and absolute economic values, many wild tea germplasm resources in Guizhou province have not been exploited. Guizhou, one of the original centers of the tea plants, is rich in wild tea

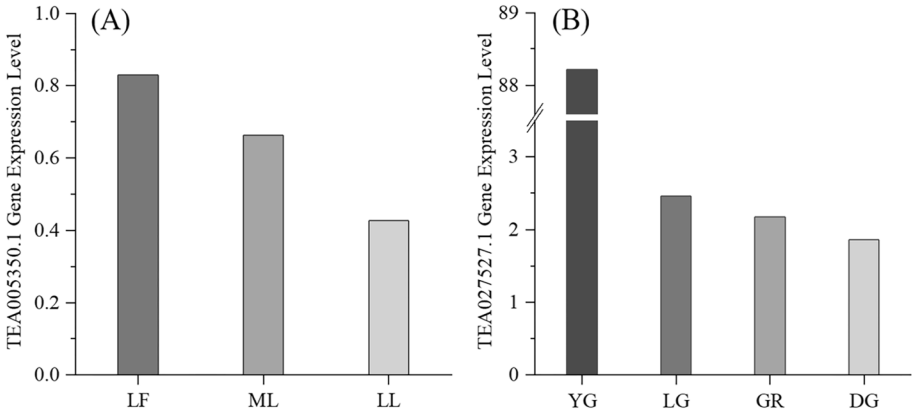

Fig. 6 **A** The expression level of TEA005350.1 gene, a candidate gene associated with mature leaf size (MLZ), in leaflet, middle, and large leaves. L.F.: leaflet; M.L.: middle leaf; L.L.: large leaf. **B** Expression level of TEA027527.1 gene associated with mature leaf color (MLC) in yellow-green, light green, green, and dark green leaves. Y.G.: yellow-green; L.G.: light green; G.R.: green; D.G.: dark green

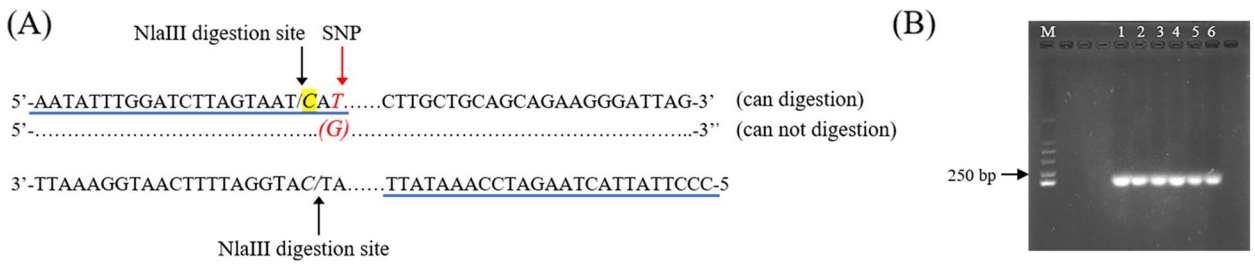

Fig. 7 Design and verification of the dCAPS marker. **A** Establishment of a dCAPS marker with the NlaIII restriction enzyme (the underlined part is the primer sequence, italics are digestion sites, the yellow bases are the introduced one mismatch into the forward primer, the red base is the SNP site); **B** PCR products of six tea accessions using the dCAPS primer

tree germplasm resources. Tea is the most drank beverage in the world, processed through the young leaves. Leaf-related traits affect plant architecture, yield, and quality potential in tea plants. In this study, the investigation of four leaf traits in 338 individuals indicated that four leaf traits of 338 tea plants displayed broad variation. Moreover, a total of 100,829 high-quality SNPs were obtained from 338 tea accessions, and the number of SNPs was higher than that the previous study reported of 415 Guizhou tea accessions with 79,016 high-quality SNPs [4], suggesting that 338 tea accessions of Guizhou province are rich in genetic diversity and can be used in GWAS [4].

Based on the population structural analysis, the 338 tea plants were classified into five groups (four ancestral groups (G-I, G-II, G-III and G-IV) and one admixture group (G-V)), consistent with the phylogenetic analysis and PCA results. G–IV represented ancient landraces, including *C. sinensis* (95.57%) and *C. remotiserrata* (4.43%). A reasonable explanation about this result may be that the ancient landraces grew on the edge of terraced fields or served as fences to separate fields owned by different farmers [4]. Due to human activities, the geographic isolation between the ancient landraces and *C. remotiserrata* (wild tree) was narrowed, thus promoting cross-pollination between different germplasm, and the *C. remotiserrata* gradually acquired the genetic background of the landraces [4]. However, previous studies had revealed that 253 tea accessions from Guizhou and 100 tea accessions from Hunan were classified into three groups, which were significantly different from our 338 individuals divided into five groups due to the rich genetic diversity of C. *remotiserrata* and C. *tachangensis* tea population. Moreover, our results further showed that tea plants from the Zunyi, Bijie, and Southwest Guizhou Autonomous Prefectures were a sister branch to the cultivars, which is consistent with the results of a previous study [4].

GWASs are considered a powerful strategy for exploring the genetic basis of complex traits at the genome-wide level [34–37]. GBS-based GWAS analysis has been widely used in many horticultural plants and crops, such as tea plants, peaches, octoploid strawberry plants, melons, and radishes. For example, the GBS method was used to detect the 18,373 SNPs of 220 Brazilian peach germplasms, and the result of GWAS revealed that 13 SNPs were associated with five peach fruit traits [38]. Kishor et al. [39] revealed that 18 significant SNPs from 48 commercial *Oriental melon* varieties were associated with various morphological traits, and four potential candidate genes were annotated. Lee et al. [40] investigated the genetic diversity of 225 radish accessions and identified 44 SNPs and 20 potential candidate genes significantly associated with Fusarium wilt resistance. Compared with other crops, there are few studies based on GBS to perform GWAS of leaf traits in tea plants. Lu et al. [7] reported that eight high-quality SNPs corresponding to six candidate genes were significantly associated with three leaf traits such as leaf color, depth of leaf serration, and density of leaf serration by using a total of 8,082,370 high-quality SNPs identified from 120 ancient tea plants. In our previous report [6], nine SNPs were found to be significantly associated with four leaf traits (MLA, MLL, MLW, and MLS) using GBS-based GWAS, but no corresponding candidate gene was found.

Although we used the same method for phenotype analysis of MLZ, MLC, MLS, and MLT for three consecutive years, the phenotypes of MLC, MLS, and MLT were consistent over 3 years, and only the phenotype of MLZ still showed some changes, resulting in some differences in the identified SNPs associated with MLZ between the three years. This is mainly related to the fact that the environment greatly affects the quantitative trait MLZ of leaves [6]. To reduce the environmental impacts, we used three years of the data of MLZ for comprehensive analysis in this study.

GLM (GLM-Q, GLM-P), MLM (MLM-P + K, MLM-Q + K), and cMLM (cMLM-P + K, cMLM-Q + K) were the six most common statistical models in GWAS. An advantage of GLM is that it has a wide detection range and can detect many SNPs associated with the target traits, but its detection accuracy is not as accurate as MLM. Although MLM can improve the accuracy of GWAS by taking the kinship matrix as a random effect, it may also filter out some SNPs markers that are truly related to the target traits due to its strict control standard. The cMLM model aims to redetect those false negative SNPs filtered out by the MLM model [2]. Based on the characteristics of these six models, we selected the most appropriate model for each trait in this study for GWAS. In this study, a total of 59 high-quality SNPs were significantly associated with four leaf traits, such as MLC (nine SNPs) by using the GLM-P model, MLT (18 SNPs) by using GLM-Q model, MLZ (22 SNPs) by using cMLM-P + k model, and MLS (10 SNPs) by using the cMLM-P + k model.

LD decay distance was used to determine the region of potential candidate genes in GWAS. In Xu's study, an SNP located on chromosome 19 associated with the starch content of leaf in Nicotiana tabacum was selected for linkage disequilibrium analysis to obtain the target candidate region [41]. Compared with self-pollinated species, the LD of cross-pollinated species, such as tea plants, declines faster because the recombination efficiency of the latter was lower [42, 43]. Our results showed that the LD levelness was the highest at

$r^2 = 0.24$, gradually beginning to decrease at $r^2 = 0.12$ when the physical distance increased to 7 kb, and low ($r^2 = 0.035$) when the physical distance increased to 50 kb. The result was lower than self-pollination plants *Prunus avium* [44] and rice [20]. This may be due to the self-incompatibility of tea plants [45]. Based on LD decay distance, genes located in the 7 kb region around SNPs associated with the four leaf traits were identified as likely candidate genes. However, based on the current reference genome, no functionally annotated genes were found in this region. We selected 50 kb as the reasonable distance that caused LD between genes and traits associated with SNPs because we found that LD became too low when the physical distance increased by more than 50 kb. Therefore, we broadened this region to 50 kb [2] and found 26 candidate genes, which methods had been adopted in many other GWAS cases [46, 47].

Further analysis revealed that one SNP (P-1076408) associated with MLC traits was distributed in coding sequences of the TEA027527.1 gene. Three SNPs (P-129108645, P-129108677, and P-101072100) were associated with MLT. The corresponding two functional genes were TEA021901.1 and TEA002469.1, respectively, and only one SNP (P-130091909) distributed in the coding sequence of the TEA026128.1 gene was associated with MLS. Moreover, two SNPs (P-122201532, P-163812618) distributed in the coding sequences of the TEA005350.1 gene and TEA029641.1 gene, respectively, were associated with MLZ within the three years. In addition, the functions of these candidate genes were further investigated based on the homologous genes in *Arabidopsis* and KEGG analysis [48–50], which were verified by RT-qPCR.

The shape of leaves is determined by development processes [51]. At first, the leaf is a small protuberance on the periphery of the shoot tip meristem (SAM), then undergoes asymmetric growth, expansion and maturation, and finally forms a shape [52]. The TEA026128.1 gene belongs to the plant-specific LBD (Lateral Organ Boundaries Domain) gene family. It is expressed in the proximal base of the initial lateral organ and is crucial in regulating plant lateral organ development. We searched the TEA026128.1 gene using the TPIA database and found it was annotated as LOB domain-containing protein 22 (*CsLBD22*). By constructing a phylogenetic tree, Teng et al. [53] have proved that the *CsLBD22* gene of the tea plant and the *AtLBD15* gene of *Arabidopsis thaliana* are orthologous genes. There is evidence that the *AtLBD15* gene is involved in the development of shoot tip meristem and regulates the expression of WUSCHEL [54]. In addition, Shuai et al. [55] showed that the lateral organ boundary (LOB) family genes were expressed in the adaxial side of the lateral organ base, and their ectopic expression led to changes in leaf shape. Our study found that the TEA026128.1 gene (*CsLBD22*) was located within 9.40 kb downstream of SNP significantly associated with MLS traits, proving that the TEA026128.1 gene is a potential candidate gene involved in the formation of MLS in tea plants.

Cell proliferation and expansion can change the number and size of cells during leaf growth, thus affecting the final leaf size [56]. However, cell proliferation and expansion are closely related to cell cycle regulation [56]. The TEA005350.1 gene belongs to the tetratricopeptide-like helix domain superfamily; it encodes a protein that plays an important role in cell division control and plant morphogenesis, mediates protein interactions, and participates in cell cycle regulation. Our results show that the *AtTSK* gene of Arabidopsis had high homology with tea plant TEA005350.1 by Blasting in the TPIA database. Therefore, we annotated the TEA005350.1 gene regarding the function of the *AtTSK* gene. We determined the expression level of the TEA005350.1 gene in three leaf sizes (leaflet, middle leaf, large leaf) by RT-qPCR. The results showed that the TEA005350.1 gene was differentially expressed in tea plant leaflet, middle leaf, and large leaves. The expression level was highest in leaflets. There is evidence that the *AtTSK* gene is expressed in the S phase of the cell cycle, and its defect delays the G2/M transformation process of the Arabidopsis cell cycle, resulting in the accumulation of many cells in the G2 phase [57]. The *AtTSK* gene was also involved in regulating cell division of the shoot tip meristem of *A. thaliana*, and its mutant meristem cells were larger than the wild type [58]. Protein TONSOKU has been studied in *A. thaliana*, *Nicotiana attenuata*, *Helianthus annuus*, and *Dendrobium catenatum*, but there is no relevant report on tea plants at present [59–61], which further confirmed our findings that the TEA005350.1 gene is located 36.84 kb downstream of the SNP (on chromosome 1) significantly associated with MLZ traits is a potential candidate gene associated with MLZ traits of tea plants. Therefore, verifying the accuracy of this gene associated with the mature leaf size of tea plants will be part of our subsequent work.

The TEA029641.1 gene belongs to the Phosphofructokinase superfamily and can catalyze the reversible mutual transformation of fructose-6-phosphate and fructose-1,6-diphosphate, which is a key regulatory step in the glycolysis pathway [62]. Fructose pyrophosphate 6 phosphate 1 phosphotransferase subunit α also participates in the pentose phosphate pathway, fructose and mannose metabolism, and other biological pathways. In RNAi transgenic lines, the homologous gene *AtPFP-α* of TEA029641.1 in *A. thaliana*, showed significant growth retardation and smaller rosette leaves [63], indicating that

the *AtPFP-α* gene affects the size of rosette leaf of Arabidopsis. In plants, glycolysis is the main respiration route and provides ATP, reductant, and precursor materials required for plant growth and development [64]. Therefore, we speculate that the TEA029641.1 gene, which is located within 1 kb downstream of SNP, was significantly associated with MLZ and was a candidate gene of MLZ.

The color of leaves was affected by the kind, content, and proportion of pigments in leaves, among which chlorophyll, carotenoids, and anthocyanins were the most important pigments [65]. The TEA027527.1 gene was annotated as geranylgeranyl transfer type-2 subunit beta-like, referring to TPIA and NCBI databases. It belonged to the protein isoprene transferases family and could catalyze geranylgeranyl pyrophosphate synthase, playing an important role in protein isoprene modification [66]. There is evidence that the rate and yield of carotenoid synthesis from geranylgeranyl pyrophosphate (*GGPP*) were guided by geranylgeranyl pyrophosphate synthase (*GGPPS*) [67]. Based on the TPIA database, the *RGTB1* gene in *A. thaliana* was a homologous gene of the TEA027527.1 gene in the tea plant, and its mutant has defects in etiolation. It has been reported that specific RAB GTPase is up-regulated in etiolated wild-type plants [68]. This found that the TEA027527.1 gene was located at 25.57 kb upstream of SNP on chromosome 1 and was significantly associated with MLC traits. Therefore, the TEA027527.1 gene may be a candidate gene for MLC.

Texture characteristics are closely related to the composition (pectin, cellulose, hemicellulose) of cell walls [69], the destruction of which will make cells lose their support and become soft [70]. Based on the annotation information of the TPIA database, we found that the TEA021901.1 gene and TEA002469.1 gene were annotated as callose synthase 11-like and belonging to the Glycosyl transfer family, and beta-glucosidase 12-like belonged to the Glycoside hydrolase family (GH1). It has been reported that callose is produced by callose synthase, which is one of the basic structures of the cell wall [71]. The GH1 family has glycoside hydrolase activity and can hydrolyze β-D-galactoside and L-arabinoside. Studies by Guo et al. [72] and Yang et al. [69] showed that $(1\rightarrow4)$-β-Galactose and $(1\rightarrow5)$-α-Arabinose are the main components of the cell wall pectin side chain, and the content of pectin was positively correlated with the hardness. Therefore, the TEA002469.1 gene located at 0.03 kb upstream of SNP on chromosome 8 and the TEA002469.1 gene located at 26.9 kb upstream of SNP on chromosome 11 may be candidate genes associated with MLT traits.

Genome-wide association analysis has identified many SNP associated with leaf morphological characteristics [73, 74]. However, marker-assisted selection (MAS) breeding is still relatively less applied. One of the main reasons for this is the lack of effective markers. For breeders, these markers have a high reference value for identifying traits of interest. Wang et al. [2] introduced two base mismatching restriction endonuclease EcoR1 and developed a TBF dCAPS marker for MAS breeding of *C.sinensis*. Hu et al. [56] developed a dCAPS polymorphic marker related to melon seed coat color to construct a genetic map after genotyping the population with insertion-deletion (InDels). In this study, we introduced a base mismatch to develop a dCAPS marker associated with MLZ and verified its feasibility in materials. The development of this dCAPS marker will provide a reference for screening molecular markers closely related to leaf size phenotypic traits.

Conclusions

In this study, we applied GBS to the GWAS of 338 tea accessions genotypes and MLZ, MLC, MLT, and MLS. We found 22, 9, 18, and ten SNPs associated with MLZ, MLC, MLT, and MLS, respectively, in three years. Using common recognition in the same model (cMLM-P + K) of MLZ within the three years, a final SNP (P1-122,201,532) was obtained, proving that it has a relatively strong correlation with MLZ. Then, we searched for candidate genes in the 50 kb region upstream and downstream of the final important SNP site and found two (TEA005350.1, TEA029641.1), one (TEA027527.1), two (TEA021901.1, TEA002469.1), and one (TEA026128.1) potential candidate genes associated with MLZ, MLC, MLT, and MLS, respectively. Based on the strong correlation with MLZ, the expression level of the TEA005350.1 gene in three different types of mature leaves was detected by RT-qPCR. At present, there are few reports on the TEA027527.1 gene associated with MLC. We used RT-qPCR to detect its expression level in different types of mature leaf colors. The TEA005350.1 and TEA027527.1 genes showed differential expression levels. In addition, a dCAPS marker associated with MLZ was designed, which may be helpful for future MAS breeding. Our research demonstrates that GBS-based GWAS is an effective method for analyzing complex traits and finding candidate genes in tea plants.

Methods

Plant materials and phenotyping

A total of 338 tea plant accessions, including 202 cultivated tea plants and 136 wild-type tea plants, were used for this study (Table S5). Among them, 336 accessions were collected from 30 counties in Guizhou Province (Fig. 8), and the other two varieties were collected from Fujian and Zhejiang Province, respectively (Table S5). The materials were planted in the tea genetic resource

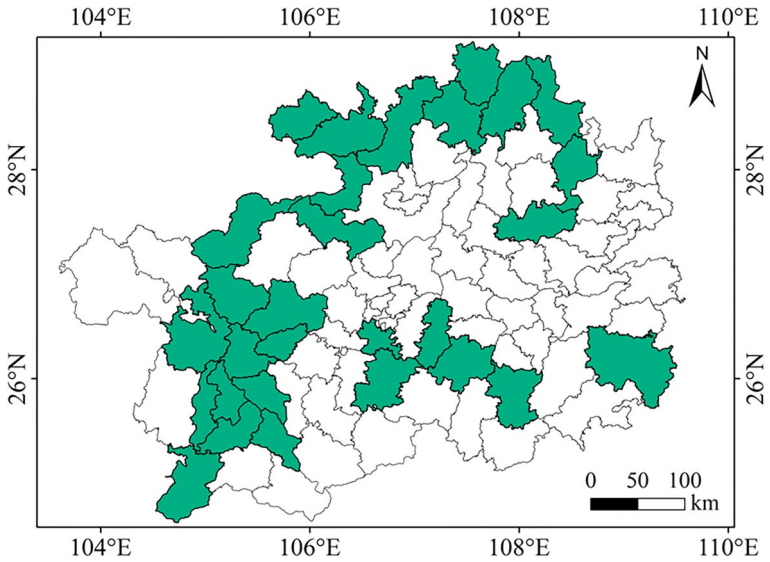

Fig. 8 Distribution of 336 materials from 30 counties in Guizhou Province

nursery (106° 40′ E and 26° 20′ N) at Guizhou University and were collected for GBS sequencing. We further investigated four leaf traits of tea plants: mature leaf size (MLZ: Leaflet area < 20 cm^2, 20 cm^2 < Middle leaf area < 40 cm^2, 40 cm^2 < Large leaf area < 60 cm^2) (Fig. 9A); mature leaf color (MLC: Yellow-green, Light green, Green, and Dark green) (Fig. 9B); mature leaf shape (MLS: Oval, Long oval, Lanceolate, Round) (Fig. 9C); and mature leaf texture (MLT: Soft, Medium, Hard). In 2019, 2020, and 2021 (these three consecutive years were regarded as three independent environments), a field experiment was conducted to investigate the above four traits on five mature leaves of 338 tea accessions according to descriptors for tea germplasm resources (NY/T 2943–2016) and assign values to qualitative traits by the scoring method [30] (Table S9). For quantitative trait MLZ, the average phenotypic value of each year in three years was considered for association analysis. Qualitative traits (MLC, MLT, MLS) were verified in the second and third-year surveys, and no significant difference was observed between the three years' phenotypic values; we used one year's data for association analysis (Table S9). The descriptive statistical analysis of four phenotypic traits (e.g., mean, minimum and maximum, standard deviation (SD), and coefficient of variation (CV)) was used to evaluate the population diversity of 338 tea accessions. SPSS estimated analysis of variance (ANOVA) of MLZ based on the average value of each character added in the three years. The component of variation assessed generalized heritability (h_B^2).

DNA extraction, library construction, and sequencing

A Plant Genomic DNA Rapid Extraction Kit (Beijing Biomed Gene T echnology Co. Ltd., Beijing, China) was used to extract genomic DNA. Restriction

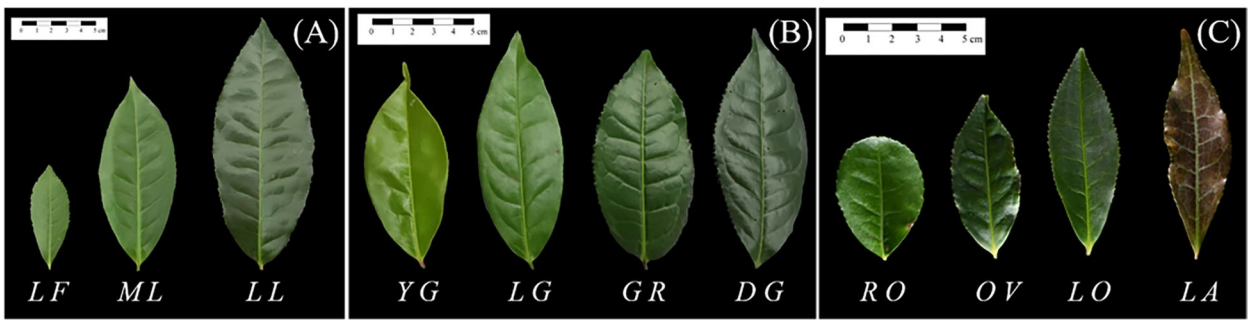

Fig. 9 **A** Three morphological characteristics of "Mature leaf size (MLZ)" phenotypic trait. L.F.: leaflet; M.L.: middle leaf; L.L.: large leaf. **B** Morphological characteristics of four "Mature leaf colors (MLC)". Y.G.: yellow-green; L.G.: light green; G.R.: green; D.G.: dark green. **C** Morphological characteristics of four "Mature leaf shapes (MLS)". R.O.: round; O.V.: oval; L.O.: long oval; L.A.: lanceolate

endonuclease SacI and MseI (5 U; New England Biolabs (NEB), Ipswich, USA) were used for digesting the DNA isolated from each sample. After digestion, adaptors "SacAD and MseAD" with unique barcodes were ligated to the digested DNA fragments. Then, the PCR Primer Cocktail and PCR Master Mix were used to amplify the purified DNA fragments. The 500–550 bp amplified fragment was retrieved through electrophoresis using 2% agarose gel, and purified using QIAquick Gel Extraction Kit (Qiagen) [75]. The average length of DNA fragments was determined using the Agilent DNA 12,000 kit and 2100 Bioanalyzer system (Agilent). Quantitative real-time PCR with a TaqMan probe was used to quantify the final DNA library and the Illumina HiSeq X ten platform with the paired-end 150 (PE150) sequencing strategy was used to sequence [75].

Sequence alignment and SNP identification

We used barcodes to de-multiplex the raw DNA reads and trim the adaptors with a custom perl script. Ultimately, the reads with quality values > 5 were retained as the clean data, then the BWA-MEM v.0.7.10 with parameters '-T 20 -k 30' was used to align with the reference genome (Accession Number. SRA536878, http://tpia.teaplant.org/) [76, 77]. GATK v.3.7.0 (https://github.com/broadinstitute/gatk/releases) was used call for SNPs [78]. Based on the method of Chen et al. [79], Hussain et al. [80], and Eltaher et al. [81], we summarized and then used the following filtering criteria: (1) The variants were bi-allelic SNPs; (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; (3) Minor allele frequency (MAF) lower than 0.05 or missing data rate higher than 20% were filtered out by VCFtools v.0.1.16 (https://github.com/vcftoools/vcftools) [82]. Three-hundred-thirty-eight accessions and 100,829 SNPs were retained for further analysis after the filter. Genetic diversity analysis.

Plink v.1.90 (https://www.cog-genomics.org/plink2/) [83] was used to calculate *Ho*, *He*, *MAF*, and *Fis* for each inferred population. VCFtools was used to calculate *Pi* and *Tajima's D* for each inferred population [82]. SPSS v.25 (IBM Corp., Armonk, NY, USA) was used to determine significant differences between these indices [84].

Population structure, PCA, and phylogenetic analysis

Admixture v.1.3.0 [85] was used to estimate the proportions of admixtures among the 338 tea accession populations by running k = 2–9. Then, the best value of K was determined by cross-validation (CV) and log-likelihood estimates. We set the threshold to 0.8 to distinguish the pure group from the mixed group. PCA was analyzed by Tassel v.5.2.43 [86]. The neighbor-joining tree was constructed using MEGA-X [87].

Linkage disequilibrium analysis

Based on the allele frequency correlation (r^2), the PopLDdecay V.3.30 was used to calculate the paired linkage disequilibrium of 29,393,327 genome-wide unfiltered SNPs from more than 500 kb, and the LD attenuation map was generated [88].

Genome-wide association study

To ensure the accuracy of the results, SNPs with a minor allele frequency less than 0.05 or a maximum deletion genotype frequency greater than 20% were screened for GWAS. The six linear regression models were the general linear model (GLM) (GLM-Q and GLM-P), mixed linear model (MLM) (MLM-Q + K and MLM-P + K), and compressed mixed linear model (cMLM) (cMLM-Q + K and cMLM-P + K). The Q-matrix (Q) or PCA-matrix (P) was taken as the fixed effect in GLM models to control possible false positives resulting from the confounding of population structure. The kinship matrix (K) was considered a covariate factor in MLMs and cMLMs to reduce unequal relatedness among genotypes [89]. K was the kinship matrix constructed by the TASSEL software, the best admission results representing population membership were denoted as Q, and P was the result of principal component analysis.. The Quantile–Quantile plots (Q-Q plots) showed the fit between the model and the data, which was used to judge the deviation between the observed and expected values. Manhattan plots showed the correlation between SNPs sites on each chromosome and this trait. Comparing the Q-Q plots of the output of the six GWAS models, the best model fitting the curve of the expected value to the observed value was entered into the later analysis as the optimal model for each trait [87]. $-\mathrm{Log}_{10}{}^{(P)} \geq 4.0$ [90] was selected as the threshold to identify SNPs sites closely associated with four leaf traits.

Candidate genes prediction

We used the best model of four traits to screen significant SNPs to reduce false positives, and candidate genes were searched within 50 kb upstream and downstream of the linked SNP loci. The functional annotation information of candidate genes was obtained from National Center for Biotechnology Information Database (NCBI, https://www.ncbi.nlm.nih.gov/) and the candidate genes that may be related to traits were predicted combined with

Tea Plant Information Archive Database (TPIA, http://tpia.teaplant.org/).

Extraction of RNA and RT-qPCR analysis

We selected 15 tea materials (including 5 large leaf materials, 5 middle leaf materials and 5 leaflet materials) with different leaf sizes and 20 tea materials with different leaf colors (including 5 yellow green materials, 5 light green materials, 5 green samples and 5 dark green materials) for RT-qPCR. For reverse transcription, the total RNA of 35 tea accessions was extracted using the UNIQ-10 column Trizol total RNA Extraction Kit (Sangon Biotech Co. Ltd., Shanghai). RT-qPCR was performed using ChamQ Universal SYBR qPCR master Mix Kit (novozan Biology), using cDNA as a template to detect the expression level of the TEA005350.1 and TEA027527.1 genes in tea varieties with different leaf sizes and colors. The results were analyzed using the $2^{(-\triangle\triangle Ct)}$ [91] method, and GADPH (Forward primer: AGCTGCACAACCAACTGTTTG, Reverse primer: AGCTGCACAACCAACTGTTTG) was used as an internal reference gene for relative quantification analysis with three replicates for each sample.

Developing and verifying a dCAPS marker associated with MLZ

SnpEff software was used to annotate the.vcf sequence files for obtaining SNP mutation site information, and we used the TPIA database to call the genes' coding sequences (CDSs). In addition, the online software dCAPS Finder 2.0 (http://helix.wustl.edu/dcaps/dcaps.html) was used to design markers at about 200 kb upstream and downstream of the SNP site and found the corresponding specific endonuclease digestion sites. Primer Premier 5.0 software was used to design specific reverse primers. Finally, genomic DNA was extracted from six tea accessions with known MLZ, and the designed markers were used for PCR amplification to verify the feasibility of the dCAPS markers obtained.

Abbreviations

MLZ	Mature leaf size
MLC	Mature leaf color
MLS	Mature leaf shape
MLT	Mature leaf texture
SNPs	Single nucleotide polymorphisms
CVerror	Cross-validation error
GBS	Genotyping-by-sequencing
GWAS	Genome-wide association study
LD	Linkage disequilibrium
NJ tree	Neighbor-Joining tree
PCA	Principal component analysis
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
dCAPS	Derived cleaved amplified polymorphism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-023-04192-0.

Additional file 1: Figure S1. Dynamics of population structure under different K (K = 2-9) values of 338 tea accessions. **Figure S2.** The Q-Q plots and Manhattan plots of other models of 4 traits except the optimal model. (A1) MLZ-2019-cMLM-Q+K; (A2) MLZ-2019-GLM-P; (A3) MLZ-2019-GLM-Q; (A4) MLZ-2019-MLM-P+K; (A5) MLZ-2019-MLM-Q+K; (B1) MLZ-2020-cMLM-Q+K; (B2) MLZ- 2020-GLM-P; (B3) MLZ-2020-GLM-Q; (B4) MLZ-2020-MLM-P+K; (B5) MLZ-2020-MLM-Q+K; (C1) MLZ-2021-cMLM-Q+K; (C2) MLZ-2021-GLM-P; (C3) MLZ-2021-GLM-Q; (C4) MLZ-2021-MLM-P+K; (C5) MLZ-2021-MLM-Q+K; (D1) MLC-cMLM-P+K; (D2) MLC-cMLM-Q+K; (D3) MLC-GLM-P; (D4) MLC-MLM-P+K; (D5) MLC-MLM-Q+K; (E1) MLS-cMLM-Q+K; (E2) MLS-GLM-P; (E3) MLS-GLM-Q; (E4) MLS-MLM-P+K; (E5) MLS-MLM-Q+K; (F1) MLT-cMLM-P+K; (F2) MLT-cMLM-Q+K; (F3) MLT-GLM-Q; (F4) MLT-MLM-P+K; (F5) MLT-MLM-Q+K. **Figure S3.** The labeled complete gels of PCR products of six tea accessions using the dCAPS primer. **Table S1.** The quality control (QC) data of each sample. **Table S2.** Genotyping of 100,829 SNPs based on GBS in 168 tea accessions. **Table S3.** Genotyping of 100,829 SNPs based on GBS in 170 tea accessions. **Table S4.** Distribution information of 100,829 SNPs on 15 chromosomes of tea plant.**Table S5.** Information of 338 tea accessions used in the present study. **Table S6.** Statistics of the number and ratio of the accessions of species, and both cultivation status in five inferred populations.**Table S7.** Analysis of TEA005350.1 gene expression of 15 tea plant accessions. **Table S8.** Analysis of TEA027527.1 gene expression of 20 tea plant accessions. **Table S9.** Four leaf phenotypic traits (mature leaf size, mature leaf color, mature leaf shape and mature leaf texture) data and their assignment of 338 tea accessions in 2019, 2020 and 2021, respectively.

Acknowledgements

We thank the College of Tea Science of Guizhou University for providing research facilities and computing facilities. We thank BX and JXW for their guidance and software suggestions in data processing, JDH, LXD, and HYW for their management of the tea germplasm gardens.

Authors' contributions

Y.J.C and S.Z.N conceived and supervised the study. Q.F.S and L.M.H analyzed and interpreted the linkage disequilibrium, population structure, and phylogenetic tree. D.C.B and Y.Q.H processed and analyzed the sequencing data. X.Y.D analyzed and screened the SNP sites. All authors read and approved the final version of the manuscript.

Funding

This work was funded by Project of the National key R & D plan (2021YFD1200203-1), Project of the National Science Foundation, in RP China (32060700) for design of the study, Guiyang Science and Technology Plan Project (Construction Technology Contract [2023] 48-21), Project of the key field project of Natural Science Foundation of Guizhou Provincial Department of education (KY [2021] 042) for data analysis, Science and Technology Plan Project of Guizhou province, in RP China ([2021] General 126) for design of the study.

Availability of data and materials

The plant materials were growing in our resource nursery which 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 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 plant material is derived from our tea germplasm nursery. 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. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, People's Republic of China. [2]Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region, Ministry of Education, Institute of Agro-Bioengineering, Guizhou University, Guiyang 550025, Guizhou Province, People's Republic of China. [3]School of Architecture, Guizhou University, Guiyang 550025, Guizhou Province, People's Republic of China.

Received: 25 December 2022 Accepted: 24 March 2023
Published online: 12 April 2023

References

1. Daglia M, Antiochia R, Sobolev AP, Mannina L. Untargeted and targeted methodologies in the study of tea (Camellia sinensis L.). Food Res Int. 2014;63:275–89.
2. 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;67:10380–91.
3. Dai Y. The overlap of suitable tea plant habitat with Asian elephant (Elephus maximus) distribution in southwestern China and its potential impact on species conservation and local economy. Environ Sci Pollut Res. 2022;29:5960–70.
4. Niu S, Song Q, Koiwa H, 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):1–12.
5. Cheng L, Dong X, Liu Q, et al. SLAF-Seq Technology-Based Genome-Wide Association and Population Structure Analyses of Ancient Camellia sinensis (L.) Kuntze in Sandu County, China. Forests. 2022;13(11):1885.
6. Niu S, Koiwa H, Song Q, Qiao D, Chen J, Zhao D, et al. Development of core-collections for Guizhou tea genetic resources and GWAS of leaf size using SNP developed by genotyping-by-sequencing. PeerJ. 2020;2020:1–22.
7. Lu L, Chen H, Wang X, Zhao Y, Yao X, Xiong B, et al. Genome-level diversification of eight ancient tea populations in the Guizhou and Yunnan regions identifies candidate genes for core agronomic traits. Horticulture Research. 2021;8. https://doi.org/10.1038/s41438-021-00617-9.
8. Cao Q, Yang G, Wang F, Chen L, Xu B, Zhao C, et al. Discrimination of tea plant variety using in-situ multispectral imaging system and multi-feature analysis. Computers and Electronics in Agriculture. 2022;202:107360. https://doi.org/10.1016/j.compag.2022.107360.
9. Zaman F, Zhang E, Xia L, Deng X, Ilyas M, Ali A, et al. Natural variation of main biochemical components, morphological and yield traits among a panel of 87 tea [Camellia sinensis (L.) O. Kuntze] cultivars. Horticultural Plant J. 2022; https://doi.org/10.1016/j.hpj.2022.08.007.
10. Zeng W, Zeng Z, Teng J, Rothenberg DO, Zhou M, Lai R, et al. Comparative analysis of purine alkaloids and main quality components of the three Camellia species in China. Foods. 2022;11:1–16.
11. Karamat U, Sun X, Li N, Zhao J. Genetic regulators of leaf size in Brassica crops. Horticulture Research. 2021;8. https://doi.org/10.1038/s41438-021-00526-x.
12. Zhang Y, Wang B, Qi S, Dong M, Wang Z, Li Y, et al. Ploidy and hybridity effects on leaf size, cell size and related genes expression in triploids, diploids and their parents in Populus. Planta. 2019;249:635–46.
13. Babu BK, Mathur RK, Ravichandran G, Anitha P, Venu MVB. Genome-wide association study for leaf area, rachis length and total dry weight in oil palm (Eleaeisguineensis) using genotyping by sequencing. PLoS ONE. 2019;14:1–10.
14. Tan LQ, Wang LY, Xu LY, Wu LY, Peng M, Zhang CC, et al. SSR-based genetic mapping and QTL analysis for timing of spring bud flush, young shoot color, and mature leaf size in tea plant (Camellia sinensis). Tree Genetics and Genomes. 2016;12. https://doi.org/10.1007/s11295-016-1008-9.
15. Li DM, Zhu GF. High-density genetic linkage map construction and QTLs Identification Associated with four leaf-related traits in lady's slipper orchids (Paphiopedilum concolor × Paphiopedilum hirsutissimum). Horticulturae. 2022;8(9):842.
16. Liu Z, She H, Xu Z, et al. Quantitative trait loci (QTL) analysis of leaf related traits in spinach (Spinacia oleracea L.). BMC Plant Biol. 2021;21(1):290.
17. Zhang J, Zhang D, Fan Y, et al. The identification of grain size genes by RapMap reveals directional selection during rice domestication. Nat Commun. 2021;12(1):5673.
18. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465:627–31.
19. Elassbli H, Abdelraheem A, Zhu Y, Teng Z, Wheeler TA, Kuraparthy V, et al. Evaluation and genome-wide association study of resistance to bacterial blight race 18 in U.S. Upland cotton germplasm. Mol Genet Genomics. 2021;296:719–29.
20. Korinsak S, Darwell CT, Wanchana S, Praphaisal L, Korinsak S, Thunnom B, et al. Identification of Bacterial Blight Resistance Loci in Rice (Oryza sativa L.) against Diverse Xoo Thai Strains by Genome-Wide Association Study. Plants. 2021;10:518.
21. Ibba MI, Juliana P, Hernández-Espinosa N, Posadas-Romano G, Dreisigacker S, Sehgal D, et al. Genome-wide association analysis for arabinoxylan content in common wheat (T. Aestivum L.) flour. J Cereal Sci. 2021;98.
22. Kumar S, Deng CH, Molloy C, Kirk C, Plunkett B, Lin-wang K, et al. Extreme-phenotype GWAS unravels a complex nexus between apple (Malus domestica) red-flesh colour and internal flesh browning. 2022. p. 1–14.
23. Zhou C, Guo Y, Chen Y, Zhang H, El-Kassaby YA, Li W. Genome wide association study identifies candidate genes related to the earlywood tracheid properties in Picea crassifolia Kom. Forests. 2022;13:1–16.
24. Yamashita H, Uchida T, Tanaka Y, Katai H, Nagano AJ, Morita A, et al. Genomic predictions and genome-wide association studies based on RAD-seq of quality-related metabolites for the genomics-assisted breeding of tea plants. Sci Rep. 2020;10:1–10.
25. Nair RJ, Pandey MK. Role of molecular markers in crop breeding: a review. Agric Rev. 2021; Of. https://doi.org/10.18805/ag.r-2322.
26. Bangarwa SK, Solanki KL. An introduction to DNA-markers and their role in crop improvement. 2021;10:638–43.
27. Hasan N, Choudhary S, Naaz N, Sharma N, Laskar RA. Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. Journal of Genetic Engineering and Biotechnology. 2021;19:1–26. https://doi.org/10.1186/s43141-021-00231-1.
28. Taranto F, D'Agostino N, Greco B, Cardi T, Tripodi P. Genome-wide SNP discovery and population structure analysis in pepper (Capsicum annuum) using genotyping by sequencing. BMC Genomics. 2016;17:1–13.
29. Hill CB, Li C. Genetic improvement of heat stress tolerance in cereal crops. Agronomy. 2022;12:1–31.
30. Dong H, Chen H. Genetic difference analysis of suspected "Bee Sugar Plum" germplasm based on leaf phenotypic traits. Northern Hortic. 2022;15:25–33.
31. Pandey J, Scheuring DC, Koym JW, Coombs J, Novy RG, Thompson AL, et al. Genetic diversity and population structure of advanced clones selected over forty years by a potato breeding program in the USA. Sci Rep. 2021;11(1):8344.
32. Tandoh KZ, Amenga-Etego L, Quashie NB, Awandare G, Wilson M, Duah-Quashie NO. Plasmodium falciparum malaria parasites in Ghana show signatures of balancing selection at artemisinin resistance predisposing background genes. EvolBioinforma. 2021;17. https://doi.org/10.1177/1176934321999640.
33. Egan LM, Conaty WC, Stiller WN. Core Collections : Is There Any Value for Cotton Breeding ? Frontiers in Plant Science. 2022;13:895155. https://doi.org/10.3389/fpls.2022.895155.
34. Tibbs Cortes L, Zhang Z, Yu J. Status and prospects of genome-wide association studies in plants. Plant Genome. 2021;14:1–17.

35. Saini DK, Chopra Y, Singh J, Sandhu KS, Kumar A, Bazzer S, et al. Comprehensive evaluation of mapping complex traits in wheat using genome-wide association studies. Netherlands: Springer; 2022.
36. Zia MAB, Demirel U, Nadeem MA, Ali F, Dawood A, Ijaz M, et al. Genome-wide association studies (GWAS) revealed a genetic basis associated with floratraits in potato germplasm. Turk J Agric For. 2022;46:90–103.
37. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1:59. https://doi.org/10.1038/s43586-021-00056-9.
38. Thurow LB, Gasic K, Bassols Raseira M do C, Bonow S, Marques Castro C. Genome-wide SNP discovery through genotyping by sequencing, population structure, and linkage disequilibrium in Brazilian peach breeding germplasm. Tree Genetics and Genomes. 2020;16:1–14. https://doi.org/10.1007/s11295-019-1406-x.
39. Kishor DS, Noh Y, Song WH, Lee GP, Park Y, Jung JK, et al. SNP marker assay and candidate gene identification for sex expression via genotyping-by-sequencing-based genome-wide associations (GWAS) analyses in Oriental melon (Cucumis melo L.var.makuwa). Scientia Horticulturae. 2021;276:109711.
40. Lee ON, Koo H, Yu JW, Park HY. Genotyping-by-sequencing-based genome-wide association studies of fusarium wilt resistance in radishes (Raphanus sativus l.). Genes. 2021;12:1–15.
41. Xu X, Wang Z, Xu S, Xu M, He L, Zhang J, et al. Identifying loci controlling total starch content of leaf in Nicotiana tabacum through genome-wide association study. Funct Integr Genomics. 2022;22:537–52.
42. Maruki T, Lynch M. Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. Genetics. 2014;197:1303–13.
43. Zhu X, Dong L, Jiang L, Li H, Sun L, Zhang H, et al. Constructing a linkage-linkage disequilibrium map using dominant-segregating markers. DNA Res. 2015;23:1–10.
44. Campoy JA, Lerigoleur-Balsemin E, Christmann H, Beauvieux R, Girollet N, Quero-García J, et al. Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of Prunus avium L. landraces and bred cultivars. BMC Plant Biol. 2016;16:1–15.
45. Tan L, Cui D, Wang L, Liu Q, Zhang D, Hu X, et al. Genetic analysis of the early bud flush trait of tea plants (Camellia sinensis) in the cultivar "Emei Wenchun" and its open-pollinated offspring. Horticulture Research. 2022;9. https://doi.org/10.1093/hr/uhac086.
46. Zhao Y, Wang R, Liu Q, Dong X, Zhao DG. Genetic diversity of ancient camellia sinensis (L.) o.kuntze in sandu county of Guizhou Province in China. Diversity. 2021;13:276. https://doi.org/10.3390/d13060276.
47. Li S, Liu SL, Pei SY, Ning MM, Tang SQ. Genetic diversity and population structure of Camellia huana (Theaceae), a limestone species with narrow geographic range, based on chloroplast DNA sequence and microsatellite markers. Plant Divers. 2020;42:343–50.
48. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30 ([PMID:10592173]).
49. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28:1947–51 ([PMID:31441146]).
50. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51:D587–92 ([PMID:36300620]).
51. Du L, Adkins S, Xu M. Leaf Development in Medicago truncatula. Genes. 2022;13:1203. https://doi.org/10.3390/genes13071203.
52. Dkhar J, Pareek A. What determines a leaf's shape? EvoDevo. 2014;5:1–19. https://doi.org/10.1186/2041-9139-5-47.
53. Teng RM, Wang YX, Wang WL, Li H, Shen W, Zhuang J. Genome-wide identification, classification and expression pattern of LBD gene family in Camellia sinensis. Biotechnol Biotechnol Equip. 2018;32:1387–97.
54. Sun XD, Feng ZH, Meng LS, Zhu J, Geitmann A. Arabidopsis ASL11/LBD15 is involved in shoot apical meristem development and regulates WUS expression. Planta. 2013;237:1367–78.
55. Shuai B, Reynaga-Peña CG, Springer PS. The Lateral Organ Boundaries gene defines a novel, plant-specific gene family. Plant Physiology. 2002;129:747–61. https://doi.org/10.1104/pp.010926.
56. Hu Y, Xiong J, Shalby N, Zhuo C, Jia Y, Yang QY, et al. Comparison of dynamic 3D chromatin architecture uncovers heterosis for leaf size in Brassica napus. J Adv Res. 2022; xxxx. https://doi.org/10.1016/j.jare.2022.01.001.
57. Suzuki T, Nakajima S, Inagaki S, Hirano-Nakakita M, Matsuoka K, Demura T, et al. TONSOKU is expressed in S phase of the cell cycle and its defect delays cell cycle progression in Arabidopsis. Plant Cell Physiol. 2005;46:736–42.
58. Suzuki T, Inagaki S, Nakajima S, Akashi T, Ohto MA, Kobayashi M, et al. A novel Arabidopsis gene Tonsoku is required for proper cell arrangement in root and shoot apical meristems. Plant J. 2004;38:673–84.
59. Schäfer M, Meza-Canales ID, Navarro-Quezada A, Brütting C, Vanková R, Baldwin IT, et al. Cytokinin levels and signaling respond to wounding and the perception of herbivore elicitors in Nicotiana attenuata. J Integr Plant Biol. 2015;57:198–212.
60. Aoyagi Blue Y, Satake A. Analyses of gene copy number variation in diverse epigenetic regulatory gene families across plants: Increased copy numbers of BRUSHY1/TONSOKU/MGOUN3 (BRU1/TSK/MGO3) and SILENCING DEFECTIVE 3 (SDE3) in long-lived trees. Plant Gene. 2022;32:100384.
61. Brzezinka K, Altmann S, Bäurle I. BRUSHY1/TONSOKU/MGOUN3 is required for heat stress memory. Plant Cell Environ. 2019;42:771–81.
62. Mehari TG, Xu Y, Umer MJ, Hui F, Cai X, Zhou Z, et al. Genome-Wide Identification and Expression Analysis Elucidates the Potential Role of PFK Gene Family in Drought Stress Tolerance and Sugar Metabolism in Cotton. Frontiers in Genetics. 2022;13. https://doi.org/10.3389/fgene.2022.922024.
63. Lim H, Cho MH, Jeon JS, Bhoo SH, Kwon YK, Hahn TR. Altered expression of pyrophosphate: Fructose-6-phosphate 1-phosphotransferase affects the growth of transgenic Arabidopsis plants. Mol Cells. 2009;27:641–9.
64. Lim H, Hwang H, Kim T, Kim S, Chung H, Lee D, et al. Transcriptomic analysis of rice plants overexpressing psgapdh in response to salinity stress. Genes. 2021;12:641. https://doi.org/10.3390/genes12050641.
65. Tian Y, Wang H, Zhang Z, Zhao X, Wang Y, Zhang L. An RNA-seq Analysis Reveals Differential Transcriptional Responses to Different Light Qualities in Leaf Color of Camellia sinensis cv. Huangjinya J Plant Growth Regul. 2022;41:612–27.
66. Andrews M, Huizinga DH, Crowell DN. The CaaX specificities of Arabidopsis protein prenyltransferases explain era1 and ggb phenotypes. BMC Plant Biology. 2010;10:1–11. https://doi.org/10.1186/1471-2229-10-118.
67. Dong C, Qu G, Guo J, Wei F, Gao S, Sun Z, et al. Rational design of geranylgeranyl diphosphate synthase enhances carotenoid production and improves photosynthetic efficiency in Nicotiana tabacum. Sci Bull. 2022;67:315–27.
68. Hála M, Soukupová H, Synek L, Žárský V. Arabidopsis RAB geranylgeranyl transferase β-subunit mutant is constitutively photomorphogenic, and has shoot growth and gravitropic defects. Plant J. 2010;62:615–27.
69. Yang L, Cong P, He J, Bu H, Qin S, Lyu D. Differential pulp cell wall structures lead to diverse fruit textures in apple (Malus domestica). Protoplasma. 2022;259:1205–17.
70. Zhang W, Guo M, Yang W, Liu Y, Wang Y, Chen G. The Role of Cell Wall Polysaccharides Disassembly and Enzyme Activity Changes in the Softening Process of Hami Melon (Cucumis melo L.). Foods. 2022;11:841. https://doi.org/10.3390/foods11060841.
71. Wang B, Andargie M, Fang R. The function and biosynthesis of callose in high plants. Heliyon. 2022;8: e09248.
72. Guo Y, Wu B, Guo X, Liu D, Qiu C, Ma H. Effect of thermosonication on texture degradation of carrot tissue in relation to alterations in cell membrane and cell wall structure. Food Chem. 2022;393:133335.
73. Muhammad A, Li J, Hu W, Yu J, Khan SU, Khan MHU, et al. Uncovering genomic regions controlling plant architectural traits in hexaploid wheat using different GWAS models. Sci Rep. 2021;11:1–14.
74. Yang W, Yao D, Wu H, Zhao W, Chen Y, Tong C. Multivariate genome-wide association study of leaf shape in a Populus deltoides and P. Simonii F1 pedigree. PLoS ONE. 2021;16:1–20.
75. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.53.
76. Xia EH, Li FD, Tong W, Li PH, Wu Q, Zhao HJ, et al. Tea Plant Information Archive: a comprehensive genomics and bioinformatics platform for tea plant. Plant Biotechnol J. 2019;17:1938–53.
77. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Preprint at https://arxiv.org/abs/1303.3997. 2013.

78. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
79. Chen W, Hou L, Zhang Z, Pang X, Li Y. Genetic diversity, population structure, and linkage disequilibrium of a core collection of ziziphus jujuba assessed with genome-wide snps developed by genotyping-by-sequencing and SSR Markers. Front Plant Sci. 2017;8:1–14.
80. Hussain W, Stephen Baenziger P, Belamkar V, Guttieri MJ, Venegas JP, Easterly A, et al. Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. Sci Rep. 2017;7:1–15.
81. Eltaher S, Sallam A, Belamkar V, Emara HA, Nower AA, Salem KFM, et al. Genetic diversity and population structure of F3:6 Nebraska Winter wheat genotypes using genotyping-by-sequencing. Frontiers in Genetics. 2018;9:76. https://doi.org/10.3389/fgene.2018.00076.
82. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011;27(15):2156–8.
83. Slifer S. PLINK: key functions for data analysis. Curr Protoc Hum Genet. 2018;97: e59.
84. Evans BA, Rozen DE. Significant variation in transformation frequency in Streptococcus pneumoniae. ISME J. 2013;7(4):791–9.
85. Wang J, Zhang Z, Gong Z, Liang Y, Ai X, Sang Z, et al. Analysis of the genetic structure and diversity of upland cotton groups in different planting areas based on SNP markers. Gene. 2021;2022(809): 146042.
86. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
87. Saitou N, Nei M. ESCALA CIWA-AR Escala CIWA-Ar(Clinical Institute Withdrawal Assesment for Alcohol) Evaluación del Síndrome de Abstinencia Alcohólica. Mol Biol Evol. 1987;4:406–25.
88. 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;35:1786–8.
89. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8.
90. Huang L, Min Y, Schiessl S, Xiong X, Jan HU, He X, et al. Integrative analysis of GWAS and transcriptome to reveal novel loci regulation flowering time in semi-winter rapeseed. Plant Sci. 2021;310:110980.
91. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods. 2001;25:402–8.

Publisher's Note