Population Structure Analysis Revealed Water or Ancient Hub-induced Genetic Diversity of Cultivated-type Tea Plants in Guizhou Plateau

Zhifei Zhao
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Qinfei Song
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Dingchen Bai
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Suzhen Niu (✉ niusuzhen@163.com)
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Yingqin He
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Dahe Qiao
Institute of Tea Science, Guizhou Academy of Agricultural Sciences, Guiyang 550006, Guizhou Province, RP China

Zhengwu Chen
Institute of Tea Science, Guizhou Academy of Agricultural Sciences, Guiyang 550006, Guizhou Province, RP China

Caiyun Li
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Jing Luo
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Fang Li
College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

Research Article

Keywords: Genotyping-by-sequencing, Population structure, Genetic diversity, Cultivated-type tea plant, Guizhou Plateau

Posted Date: October 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-964671/v1

Abstract

Background

Tea plants originated from the southwest of China. Guizhou is one of the origin center of tea plants, which is rich in tea plant germplasm resources. However, the distribution characteristics and transmission model of tea plant were still unclear.

Results

We collected 253 cultivated-type tea plant accessions from Guizhou plateau and analyzed the genetic diversity, PCA, phylogenetic, population structure, LD, and development of core collection using the genotyping-by-sequencing (GBS) approach. A total of 112,072 high-quality SNPs were identified, which was further used to analyze the genetic diversity and population structure. In this study, we found that the genetic diversity in cultivated-type tea accessions of PR Basin were significantly higher than that in cultivated-type tea accessions of YR Basin. Moreover, four groups, including three pure groups (CG-1, CG-2 and CG-3) and one admixture group (CG-4), were identified based on population structure analysis, which was verified by PAC and phylogenetic analysis. Our results showed that the highest GD and *Fst* values were found in CG-2 vs CG-3, followed by CG-1 vs CG-2 and CG-1 vs CG-3. The lowest GD and *Fst* values were detected in CG-4 vs CG-1, CG-4 vs CG-2, and CG-4 vs CG-3.

Conclusions

This study provided the evidence to confirm the contribution of PR and YR Basins and ancient hub road section to the transmission of cultivated-type tea accessions in Guizhou plateau. The genetic diversity, population structure and core collection revealed by our study will benefit further genetic studies, germplasm protection, and breeding.

Background

Tea (*Camellia sinensis*) is one of the three drinks with the largest consumption, embodied numerous cultural, health and economic values in the world [1]. Tea extracts are rich in secondary metabolites [2], including polyphenol, theanine, caffeine, polysaccharides and volatile oils, have many health benefits such as antioxidant, improving attention, diuretic effect, hypoglycemic effect and immunomodulatory [2–6]. Tea plants originated from the southwest of China, have been widely cultivated over 60 countries and spread to over 160 countries around the world, which has a significant impact on the agricultural economy [1, 7]. In stark contrast to flourishing development of the spread and cultivation of tea, low breeding efficiency and lack excellent varieties, which was the future challenges in global tea industry [8]. The germplasms were the invaluable fundamental resources of crop genetic improvement and determine the success in breeding program [9]. The research of genetic diversity of tea germplasms has increased our knowledge of the origins and population structures of tea plants, that will benefit the improved variety breeding and development of tea industry [8].

Previous studies showed that many ancient tea plants are distributed in the Yangtze River Basin and southern reaches of it, especially in Guizhou, Yunnan, and Guangxi provinces. And higher levels of tea plant genetic diversity were detected in these regions [1, 8, 10]. Guizhou is one of the origin center of tea plants, which is located the upstream of two basins in China (Yangtze River Basin, YR and Pearl River Basin,PR) [11, 12], and contain rich in tea germplasms with high levels genetic diversity and various morphological characteristics, such as modern landraces, ancient landraces and wild germplasm, due to the slow land use and economic development, self-incompatibility and allogamy of tea plants, and long history of cultivation [9, 13–15].

Previous studies have reported the genetic diversity of tea plants through various molecular markers, such as RAPD [16], SSR [17], EST-SSR [18], and AFLP [19]. With advent of next generation technologies, genotyping-by-sequencing (GBS) is a rapid and low-cost tool to genotype breeding populations [20], allow plant breeders to implement genome-wide association studies (GWAS), genomic diversity study, genetic linkage analysis, molecular marker discovery, and genomic selection (GS) under a large scale of plant breeding programs [21]. So far, GBS was widely used in many plants, such as wheat [22], maize [23], pepper [24], and pine [25, 26]. Previous studies reported that a total of 79,016 high-quality SNPs were identified, and further analysis revealed that the genetic diversity in ancient landraces and admixed wild type was higher than that in the pure wild type and modern landraces [14, 27].

Previous studies revealed that wind [28], water [29], animals [30, 31] and human activity [32] contributed to the distribution and genetic diversity of plants. To further explore the distribution characteristics and transmission model effect of the basins and ancient hubs on the distribution and genetic diversity of tea plants, we sampled 253 cultivated-type tea accessions from 32 regions distributed in seven water systems in PR and YR basins of Guizhou Plateau. Base on GBS analysis, the genetic diversity, population structure, and linkage disequilibrium were investigated using the SNPs data of 253 cultivated-type tea accessions. Moreover, we further explore the contribution of basins and ancient hubs to the genetic diversity of cultivated-type tea plant population distributed in PR and YR basins of Guizhou Plateau. In addition, the core collection of these tea plant accessions was constructed.

Results

Sequencing and variant discovery

A total of 253 cultivated-type tea plant accessions, including 172 ancient landraces and 81 modern landraces distributed in PR and YR Basins of Guizhou Plateau [14], were used in this study. Of these, the geographic distribution of 249 accessions in the PR and YR basins of Guizhou Plateau was shown in Fig. 1A, and the other four accessions introduced from other provinces and cultivated in many tea gardens in Guizhou Plateau were also included in this study. GBS analysis of 253 cultivated-type tea plant accessions was performed using Illumina HiSeq Xten platform. A total of 255.2 GB clean data with an average of 1.00 Gb clean data for each accession was obtained (Additional file 1: Table S1). We mapped the clean reads in to released tea reference genome sequence (http://tpia.teaplant.org/). Our results showed that 29,393,327 single nucleotide polymorphisms (SNPs) were identified using GATK (v3.7.0). After filtering, 112,072 high-quality SNPs were identified and calculated heterozygosity values, fund that the average of heterozygosity of each accession is 7.89% (Additional file 1: Table S2). Based on the nucleotide substitution, 112,072 SNPs were classified into transitions and transversions. Transitions were observed in 1,585,601 (77.87%) and transversions in 450,567 SNPs (22.13%). The frequency of substitutions was 137,380 (6.75%) A/T, 112,861 (5.54%) A/C, 117,244 (5.76%) G/T, 83,082 (4.08%) C/G, 805,072 (39.54%) C/T, and 780,529 (38.33%) A/G, with the transitions to transversions ratio of 3.51 (Table 1).

Table 1
Genetic diversity parameters of 253 cultivated-type tea accessions in Guizhou plateau.

	Transitions		**Transversions**			
	CT	AG	AT	AC	CG	GT
Numbers of allelic sites	805,072	780,529	137,380	112,861	83,082	117,244
Percentage of allelic sites	39.54%	38.33%	6.75%	5.54%	4.08%	5.76%
Total (Percentage)	1,585,601(77.87%)		450,567(22.13%)			

Estimation of genetic diversity

The average heterozygosity (*Pi*), observed heterozygosity (*Ho*), minor allele frequency (MAF) and inbreeding coefficient (*Fis*) were used as indicators of genetic diversity. In this study, the *Pi*, *Ho*, MAF, and *Fis* of 253 cultivated-type tea plant accessions were 0.230, 0.082, 0.149 and 0.657, respectively (Table 2). We first compare the genetic diversity of two tea plant populations distributed in PR Basin and YR Basin of Guizhou Plateau, and the result showed that *Pi, Ho*, and MAF values in tea population of PR Basin were significantly higher than that in tea population of YR Basin except the *Fis* value (Table 2, Fig. 1A). In the PR Basin, the *Pi*, *Ho* and MAF values in tea population of WS02 were significantly higher than those in other water systems. The *Fis* value in WS04 was higher than those in other water systems. In the YR Basin, the *Pi* and MAF in WS06 were significantly higher than those in WS05 and WS07 except that *Ho* value. The *Fis* value in WS07 was higher than other water systems. We estimated the genetic diversity of two cultivated statuses (ancient landrace and modern landrace) of the cultivated-type tea population, the *Pi* and MAF was significantly higher in ancient landrace than in modern landrace. The *Ho* was significantly higher in modern landrace than in ancient landrace (Table 2).

The positive Tajima’s D value of all test tea populations suggested that the evolution of these tea populations displayed the population bottlenecks, and/or balancing selection (Table 2). The analysis of the genetic differentiation coefficient (*Fst*) and genetic distance (*GD*) of seven water systems showed that the pairwise *Fst* value were less than 0.05 among seven water systems. The highest pairwise genetic distance (*GD*) was in WS04 vs WS02, WS05. The lowest pairwise genetic distance (*GD*) was in WS01 vs WS06, WS07. (Table 3).

Table 2
Genetic diversity parameters of 253 cultivated-type tea accessions in Guizhou plateau.

Type		Number	Tajima D	Pi	Ho	MAF	Fis
Basins	PR	75	0.752	0.234a	0.091c	0.151a	0.625ab
	YR	174	1.048	0.228d	0.079g	0.148bc	0.664ab
Water systems	WS01	11	0.387	0.212g	0.103b	0.142d	0.460c
	WS02	26	0.380	0.232ab	0.111a	0.151a	0.563b
	WS03	27	0.321	0.225e	0.081ef	0.147c	0.653ab
	WS04	11	0.337	0.218f	0.066h	0.142d	0.695a
	WS05	9	0.408	0.216f	0.083de	0.142d	0.612ab
	WS06	151	0.999	0.229cd	0.081f	0.149bc	0.658ab
	WS07	14	0.216	0.204h	0.058i	0.133e	0.704a
Cultivation status	ML	81	0.593	0.218f	0.084d	0.142d	0.627ab
	AL	172	1.098	0.232ab	0.082ef	0.151a	0.662ab
All	all	253	1.236	0.230bc	0.082ef	0.149b	0.657ab

Note: *Pi* heterozygosity, *Ho* observed heterozygosity, *MAF* minor allele frequency, *Fis* inbreeding coefficient; The different letters indicate a significant difference in p= 0.05 levels by the T-test; *PR*, Pearl River Basin contains *WS01* Liujiang River System, *WS02* Hongshui River System, *WS03* Beipanjiang River System and *WS04* Nanpanjiang River System. *YR* Yangtze River Basin contains *WS05* Yuanjiang River System, *WS06* Wujiang River System, *WS07* Chishui River System.

Table 3
pairwise Fst and genetic distance among seven water systems of 253 accessions in Guizhou plateau

	WS01	WS02	WS03	WS04	WS05	WS06	WS07
WS01		0.211	0.211	0.215	0.211	0.209	0.209
WS02	0.037b		0.220	0.225	0.224	0.221	0.219
WS03	0.034d	0.020h		0.220	0.219	0.218	0.219
WS04	0.033d	0.019i	0.004m		0.225	0.221	0.222
WS05	0.040a	0.027e	0.011l	0.013k		0.221	0.223
WS06	0.027e	0.021g	0.011l	0.002o	0.011l		0.217
WS07	0.035c	0.016j	0.013k	0.011l	0.023f	0.003n	

Note: The bottom left is the value of pairwise genetic differentiation coefficient (Fst); The upper right is the value of pairwise genetic distance; The different letters indicate a significant difference in p= 0.05 levels by the T-test; WS01 Liujiang River System, WS02 Hongshui River System, WS03 Beipanjiang River System, WS04 Nanpanjiang River System, WS05 Yuanjiang River System, WS06 Wujiang River System, WS07 Chishui River System.

Population structure, PCA, and phylogenetic analysis

A total of 112,072 high-quality SNPs were used to analyze the population structure, PCA and phylogenetic analysis of 253 cultivated-type tea plant accessions. Population structure analysis revealed that the minimum cross-validation (CV) error occurred when *K* equals 3, indicating that three ancestral groups and one admixture group were identified (Fig. 2A, Additional file 2: Fig. S1). The accessions with the membership coefficients greater than 0.80 were assigned to corresponding pure group, while those with the membership coefficients less than 0.80 were assigned to the admixture group.

The first pure group contained 35 accessions, including 32 (91%) modern landraces and 3 (9%) ancient landraces, and the four introduce varieties were classed into this group (Additional file 3: Table S2) (referred to as ‘the modern landraces group’ or ‘CG-1’ from now). The second pure group contained 25 accessions (2 modern landraces and 23 ancient landraces), and further analysis showed that these ancient landraces mainly derived from PR Basin (Additional file 3: Table S3) (referred to as ‘the PR ancient landraces group’ or ‘CG-2’ from now). The third pure group was composed of 60 ancient landraces, including 53 (88%) from YR Basin and 7 (12%) from PR Basin (Additional file 3: Table S3) (referred to as ‘the YR ancient landraces group’ or ‘CG-3’ from now). In addition, 133 (52.56%) accessions were assigned to admixed group (referred to as ‘the ancient hubs group’ or ‘CG-4’ from now), which contained 47 modern landraces and 86 ancient landraces. Among them, 38 accessions were collected from PR Basin, which contained 11, 6, 13, and 8 accessions in WS01, WS02, WS03, and WS04, respectively. Ninety-five accessions were sampled from YR Basin, including 8 accessions from WS05, 84 accessions from WS06, and 3 accessions from WS07 (Additional file 3: Table S3).

To further verify the stability of potential population structure, the principal component analysis (PCA) and NJ phylogenetic tree were used to explore the cluster relationships among the 253 cultivated-type tea accessions using 112,072 SNPs. PCA and NJ tree result reveals four major clusters corresponding to the clusters CG-1, CG-2, CG-3 and CG-4, which further verifies the accuracy of population structure. (Fig. 2B, Fig. 2C).

Linkage disequilibrium analysis

Linkage disequilibrium (LD) was commonly used to reveal much about domestication and breeding history. We report LD estimation in a population of 253 accessions using 29,393,327 non-LD pruned SNPs. Our results showed that the LD decayed rapidly with the increasing physical distance. LD dropped to half of its the maximum value at 10 kb (Fig. 2D). Moreover, the lowest LD decay was observed in CG-2, while the most rapidly LD decay was found in CG-4 among 4 inferred populations (Additional file 4: Fig. S1).

Genetic differentiation analysis of inferred populations

Based on the population structure analysis, the genetic diversity among the inferred populations (CG-1, CG-2, CG-3, and CG-4), including Tajima's D, *Pi*, *Ho*, and MAF were calculated (Fig. 3). Our results showed that the *Pi, Ho*, and MAF values in CG-4 population were significantly higher than those in pure populations (CG-1, CG-2, and CG-3). Except MAF value was no significant difference detected in CG-4 and CG-3. Moreover, the *Pi, Ho*, and MAF values in CG-3 population were higher than that in CG-1 and CG-2 populations. The *Pi, Ho*, and MAF values in CG-2 population were higher than that in CG-1 population. The positive Tajima's D values were observed in four inferred populations suggested that the evolution of these four inferred populations displayed the population bottlenecks, and/or balancing selection (Fig. 3).

We analyzed the pairwise genetic differentiation coefficient (*Fst*) across four inferred populations. The mean *Fst* between CG-4 and CG-1 was 0.060, suggesting that there is moderate divergence between CG-4 and CG-1 populations. In addition, the mean *Fst* in CG-1 vs CG-2, CG-1 vs CG-3 and CG-2 vs CG-3 were 0.090, 0.089, and 0.091, respectively, suggesting that there is not only moderate divergence between modern landraces population (CG-1) and ancient landraces populations (CG-2 and CG-3), but also exists in ancient landraces of CG-2 and CG-3 populations. The highest genetic distance (GD) was observed in CG-2 vs CG-3, and the lowest GD was found in CG-4 vs CG-1 (Fig. 3).

Development core collection

The core and mini-core sets were developed to choose a minimum number of accessions representing the maximum diversity of the original collection, which was used in molecular marker assisted breeding, GWAS and other purpose [33–35]. The maximum length subtree method implicated in DARwin v.6.0.17 was repeatedly used to remove the most redundant accessions until the pruned edge and sphericity index percentage tend to level corresponding to 195 accessions (Additional file 3: Table S1). The 195 accessions were selected to represent 253 cultivated-type tea accessions (referred 'core set' from now). On x-axis where the number of accessions decreased from 195 to 85, the pruned edge and sphericity index percentage were increasing stably and slowly, indicating that the information of 111 accessions had no significant difference, suggesting the sphericity index and pruned edge had no significant impact after removing these accessions (Additional file 5: Fig. S1). The mini-core set was constituted by the remaining 85 accessions (Additional file 3: Table S1).

To further estimate the quality of mini-core and core collections, we construction the NJ tree to verify whether the backbone of the NJ tree has changed. The 253 cultivated-type tea plant accessions could be further divided into seven clusters based on the topology of the NJ tree: cluster I~VII (Fig. 4A). The genetic distance matrix between individuals using a SNP database was used to construct the NJ tree. The cluster I contained one ancient landrace from WS02 of PR Basin, and the cluster II was consisted of 29 accessions, including 15 modern landraces and 14 ancient landraces, mainly distributed in YR Basin. The cluster III contained 69 accessions, including 68 ancient landraces and 1 modern landraces, mainly distributed in WS06. The cluster IV consisted of 12 accessions, including 11 ancient landraces and 1 modern landrace. ten ancient landraces from WS06, 1 modern landrace from WS02 and 1 ancient landraces from WS03. The cluster V contained 44 accessions, 37 accessions of which was ancient landraces and 7 accessions of which was modern landraces, 35 accessions from YR Basin and 9 accessions from PR Basin. The cluster VI was consisted of 42 accessions, including 36 ancient landraces and 6 modern landraces, thirty accessions from PR Basin and 12 from YR Basin. The cluster VII contained 56 accessions, including 51 modern landraces and 5 ancient landraces, 17 accessions from PR, 35 accessions from YR and 4 accessions from OT (Additional file 3: Table S1).

We further analyzed the MAF, *Pi*, and genetic distance among whole set (253 cultivated-type tea accessions), core set, and mini-core set. There is no significant difference in MAF and *Pi* between core set and whole set, while the mini-core set captured 97% *Pi* and MAF of the whole set. The genetic distance of the mini-core and core sets decreased slightly, while the minimum value (lower limit) of the genetic distance range of whole set, core set and mini-core set were 0.036, 0.076 and 0.076 (Fig. 4A, Table 4). The proportion of accession with pairwise genetic distance value in 0.200-0.250 among core set and mini-core set increased obviously (Fig. 4B). In general, these results suggested that the min-core and core sets contained accessions from seven clusters of NJ tree, two basins, seven water systems, and two cultivated statuses, and can represent the genetic diversity of whole set. (Fig. 4, Additional file 3: Table S1).

Table 4
genetic diversity of mini-core and core set of cultivated-type tea plant of Guizhou plateau.

group	simple size	Pi	MAF	AGD	GDR
mini-core	85	0.223b	0.145b	0.211c	0.076-0.265
core set	195	0.230a	0.149a	0.217b	0.076-0.273
whole set	253	0.230a	0.149a	0.265a	0.036-0.347

Note: *Pi* heterozygosity, *MAF* minor allele frequency, *AGD* average genetic distance, *GDR* Genetic distance range;

Discussion

Previous studies revealed that wind [28], water [29], animals [30, 31] and human activity [32] contributed to population distribution, genetic exchange, species population expansion and so on. However, the distribution characteristics and transmission model of cultivated-type tea plants was still unclear. In this study, 253 cultivated-type tea accessions, distributed in PR and YR Basin, were collected from Guizhou Plateau for the first time. The population structure analysis, genetic diversity analysis, core collection construction and genetic exchange mechanism of these cultivated-type tea accessions were performed, and further analysis revealed that the ancient hubs and basins played important roles in the distribution characteristics and transmission model of cultivated-type tea plant in Guizhou Plateau.

Genetic diversity of cultivated-type tea plants

GBS was widely used for analysis the population structure in many plants, such as maize [36], common bean (*Phaseolus vulgaris L.*) [37], wheat [38], and tea plants [14, 27]. Previous study showed that 390.3 Gb clean data was obtained from 415 tea accessions with an average of 0.94 Gb clean data per accession. A total of 79,016 high-quality SNPs were identified [14]. We generated 255.2 Gb clean data from 253 tea accessions with an average of 1.00 Gb clean data per accession, and identified 112,072 high-quality SNPs, which was higher than the data reported in the previous study. Moreover, the transition/transversion rate was 3.51, which was higher than those in common bean (*Phaseolus vulgaris L.*) (1.27) [37], apricot (1.78~1.79) [39] and lettuce (2.10) [40] and lower than the previous report of tea plant (4.02) [14]. This result suggested that transitions were better tolerated the natural resistance, which could be that they were synonymous mutations in protein-coding sequences [41].

The *Pi*, *Ho* and MAF values of cultivated-type tea population in PR Basin were significantly higher than those in YR Basin, and the *Fis* of cultivated-type tea population in PR Basin was lower than that in YR Basin (Table 2, Fig. 1B). These results further indicated that the genetic diversity of cultivated-type tea population in PR Basin was significantly higher than that in YR Basin. In addition, the higher and lower genetic diversity were detected in WS02 and WS04 of PR Basin, respectively. And the higher and lower genetic diversity were detected in WS06 and WS07 of YR Basin, respectively (Table 2). A plausible explanation of these results could be WS02 admixed a great of modern landraces on the basis of the initial ancient landraces, and frequent genetic exchange occurred among the individuals of these two cultivated statuses. As an important river traffic, Wujiang River run through the whole Wujiang water system (WS06) and promoted the frequent genetic exchange of cultivated-type tea plant population. And WS04 and WS07 located at the edge of two basins, and few corridors were provided to promote the genetic exchange. The genetic diversity level was higher in ancient landraces than in modern landraces, except *Ho* value. That could be the cultivation of the ancient landraces was not for breeding purposes [14], while in order to serve production, the modern landraces had suffered a certain degree selects [14, 42].

Previous studies showed that there exist population bottlenecks and/or balancing selection when the positive Tajima's D values was detected in a population [43, 44]. The positive Tajima's D values were observed in all these populations, suggesting that these populations existed population bottlenecks, and/or balancing selection (Table 2). The *Fst* is widely used as a measure of population structure and *Fst* between 0.00-0.05 indicate little divergence and 0.05–0.15 moderate divergence [45–47]. All the pairwise *Fst* value of the seven water systems were in the range of 0.00-0.05, indicating there were little divergence in these water systems.

Population structure, PCA and phylogenetic tree analysis of cultivated-type tea plants

Population structure analysis showed that 253 cultivated-type tea accessions from Guizhou Plateau were grouped into four groups: three pure groups (CG-1, CG-2, and CG-3) and one admixture group (CG-4) (Fig. 1A). Our results showed that the CG-1 contain the four varieties introduce from other Province, which further revels that there were similar domestication and artificial selection direction between modern breeding and modern landraces. The highest GD and *Fst* values were detected in CG-2 vs CG-3, suggesting that the genetic background of the wild ancestors of CG-2 and CG-3 was very different owning to the geographical separation caused by YR and PR basins. The GD and *Fst* values in CG-1 vs CG-2 and CG-1 vs CG-3 were higher than those in CG-4 vs CG-1, CG-4 vs CG-2, and CG-4 vs CG-3, and lower genetic diversity was detected in CG-1, suggesting that the moderate divergence and higher genetic distant between modern landraces (CG-1) and ancient landraces (CG-2 and CG-3), narrow genetic diversity in CG-1 were mainly caused by artificial selection and domestication. The highest level of genetic diversity was detected in CG-4, and minimum GD and *Fst* values were discovered in the admixture group (CG-4) vs other groups, suggesting that the members of CG-4 group were derived from the cross-breeding among CG-1, CG-2, and CG-3. Moreover, the members of CG-4 group were distributed in the nearby river, ancient hub road section, and the junction of two basins (Fig. 1B, Additional file 6: Fig. S1), which caused the frequently cross-breeding among CG-1, CG-2, and CG-3.

Development core collection

Low breeding efficiency and lack excellent varieties were the major challenges in global tea industry [8]. Tea germplasms, as the invaluable fundamental resources of biotechnology studies and variety improvement, enhanced the extraordinarily rapid development of tea plant genomics, genetics, and breeding [48–52]. The core collection was used for detecting the novel variation, selecting excellent varieties, and providing the excellent germplasms for breeders by using smaller populations and greater genetic diversity [53]. So far, development of core collection was widely used in many plants, such as cowpea [54], alpine plum [55], walnut [56], tea plant [27, 48] and so on. However, development of the core collection of cultivated-type tea accessions from Guizhou Plateau has not been reported. In this study, the core set and mini-core set we developed contain 77.0% and 33.6% of number of individuals of initial set, separately. These accessions of mini-core and core sets accounting for two cultivated statuses (modern landraces and ancient landraces), two basins (YR Basin and PR Basin) and seven water systems (WS01~WS07). In addition, the contain number proportion of accession was consisted with the genetic diversity of modern landraces and ancient landraces. The core set were almost as the initial set in *Pi*, MAF and *AGD*. These results indicated that the core set can well represented 253 cultivated-type tea accessions for further researches. In order to save management costs, the mini-core set with smaller number represent most of genetic of initial set, which will be used in field experiment, association analysis and providing parent lines for variety improvement [57].

Conclusions

Base on the GBS analysis of cultivated-type tea plant accessions, a total of 112,072 high-quality SNPs were identified, which was further used to analyze the genetic diversity and population structure. In this study, we found that the genetic diversity in cultivated-type tea accessions of PR Basin were significantly higher than that in cultivated-type tea accessions of YR Basin. Moreover, four groups, including three pure groups (CG-1, CG-2, and CG-3) and one admixture group (CG-4), were identified based on population structure analysis, which was verified by PAC and phylogenetic analysis. Our results showed that the highest GD and *Fst* values were found in CG-2 vs CG-3 owning to the geographical separation caused by YR and PR basins. The moderate divergence between modern landraces (CG-1) and ancient landraces (CG-2 and CG-3), and lowest genetic diversity in CG-1 were mainly only caused by artificial selection and domestication. The members of CG-4 group were derived from the cross-breeding among CG-1, CG-2, and CG-3. Moreover, the members of CG-4 group were distributed in the nearby river, ancient hub road section, and the junction of two basins which caused the frequently cross-breeding among CG-1, CG-2, and CG-3. Finally, we developed the mini-core and core collections of cultivated-type tea plant, this information will benefit the germplasm protection and management, genome-wide association studies and breeding.

Methods

Plant materials

A total of 253 cultivated-type tea plant accessions [14], including 172 ancient landraces and 81 modern landraces, were collected and used in this study. Among 253 tea accessions, 249 cultivated-type tea accessions were collected from 32 regions of Guizhou Plateau, the other four accessions were introduced from Fujian, Zhejiang and Hunan provinces, and as the main varieties cultivated in most tea garden of Guizhou Plateau (Additional file 3: Table S1). Among them, 249 accessions were distributed in two basins (PY and YR) (Additional file 7: Fig. S1). The PR Basin contains four water systems (Liujiang WS01, Hongshui WS02, Beipanjiang WS03, and Nanpanjiang WS04), and the YR Basin contains four water systems (Yuanjiang WS05, Wujiang WS06, Chishui WS07) (Additional file 7: Fig S2) [12]. Eleven individuals from WS01, 26 individuals from WS02, 27 individuals from WS03, 11 individuals from WS04, 9 individuals from WS05, 151 individuals from WS06 and 14 individuals from WS07 (Additional file 3: Table S1).

DNA extraction, Library construction and Sequencing

The Plant Genomic DNA Rapid Extraction kit (Beijing Biomed Gene Technology Co., Ltd., Beijing, China) was used to extract genomic DNA from the samples according to the instructions provided by the manufacture. The DNA isolated from each sample was digested by restriction endonuclease SacI and MseI (5 U, NEB) and the adaptors “SacAD and MseAD” with unique barcodes were ligated with the DNA fragments, which were separated on 2% agarose gel and the length range from 500 to 550 bp was chosen for amplification before sequencing on an Illumina Hi-seq platform. The original paired-end sequence length was 150 bp [14, 58].

Sequence alignment and SNP identification

The barcodes were used to de-multiplex the raw DNA reads with trimming the adaptors using a custom perl script. Only reads with quality values >5 were retained and mapped to the reference genome (http://tpia.teaplant.org/) using the BWA-MEM (v0.7.10) software with default parameters [1]. The SNPs were filtered according to the methods used by Niu et al. [14] based on the following criteria: (1) variants must be 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” was used in variant filtration in GATK (v3.7.0) to filter the SNPs; (3) SNPs with minor allele frequency (MAF) higher than 0.05 or missing data rate lower than 20% were reserved by VCFtools (v0.1.160). Finally, 112,072 SNPs from 253 tea accessions were selected and performed the following analysis (Additional file 8, Additional file 9).

Linkage disequilibrium (LD) and Population structure

Linkage disequilibrium (LD) was calculated based on the correlation coefficient (r^2) statistics for genome-wide unpruned pairwise SNPs using PopLDdecay (v3.29) with the default parameters [59].

The VCFtools (v0.1.160) was used to convert VCF files to pedigree files. The ADMIXTURE (v 1.30) was used to estimate admixture proportions among the cultivated-type tea populations through assuming the number of ancestries (K) from 1 to 9. The most optimal K value was confirmed base on the minimum CV error estimated by ADMIXITURE software [60]. The threshold of membership coefficient was set at 0.8 to distinguish between the pure and admixture group. Principal component analysis (PCA) was performed using TASSEL (v5.2.72) [61]. The Neighbor-Joining (NJ) tree was constructed in MEGA (v10.2.4) with default parameter [62].

Genetic diversity

The observed heterozygosity (*Ho*), minor allele frequency (MAF) and inbreeding coefficient (*Fis*) of each inferred population were calculated using Plink (version 1.90). The heterozygosity (*Pi*) and Tajima’s D of each inferred population and genetic differentiation coefficient (*Fst*) of pairwise inferred populations were computed using VCFtools. The MEGA (v10.2.4) was used to compute genetic distance (GD) of pairwise inferred populations. The significant differences of these indexes were examined in SPSS (v 25. Ink).

Development of core collection

The NJ tree was generated based on the 112,072 SNPs. Then, the ‘maximum length subtree function’ was used to develop the core collection as described previously for tea. The threshold and steps of development the core collections were completely referred to our previous report [27].

Abbreviations

LD: linkage disequilibrium; GBS: genotyping-by-sequencing; GD: Genetic distant; PCA: principal component analyses; NJ tree: neighbor Joining tree; GWAS: genome-wide association studies; Pi: heterozygosity; Ho: observed heterozygosity; MAF: minor allele frequency; Fis: inbreeding coefficient;

Fst: differentiation coefficient; AGD: average genetic distance; GDR; Genetic distance range; CV error: Cross-validation error.

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

Availability of data and material

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 [64] 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.

Competing interests

The authors declare that they have no competing interests and consent for publication.

Funding

This work was funded by Science and Technology Plan Project of Guizhou province, in RP China ([2019]1404) for design of the study, Project of the National Science Foundation, in RP China (32060700) for design of the study, Science and Technology Plan Project of Guizhou province, in RP China ([2017]2558) for design of the study, Project of the key field project of Natural Science Foundation of Guizhou Provincial Department of education (KY [2021] 042) for data analysis, Project of the National key R & D plan (2021YFD1100307) and Science and Technology Plan Project of Guizhou province, in RP China ([2020]2103).

Authors' contributions

Z.F.Z, S.Z.N and Z.W.C conceived and supervised the study. Q.F.S analyzed and interpreted the genetic diversity, linkage disequilibrium and population structure. D.C.B and Y.Q.H processed and analyzed the sequencing data. C.Y.L, and J.L wrote and reviewed the manuscript. D.H.Q and F.L performed the filtered the genotyping data. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank College of tea science of Guizhou University for providing research facilities and computing facilities. We thank B.X for his guidance and software suggestions in data processing, W.H.A, Z.H, L.M.H and Y.J.C for their management of the tea germplasm gardens, and S.Y for drawing a map of Guizhou Plateau.

Authors' information

[1] College of Tea Science / Tea Engineering Technology Research Center, Guizhou University, Guiyang 550025, Guizhou Province, RP China

[2] Institute of Tea Science, Guizhou Academy of Agricultural Sciences, Guiyang 550006, Guizhou Province, RP China

References

1. 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 genome evolution and adaptation of tea plants. Mol Plant. 2020;13(7):1013–1026.
2. Fang K, Xia Z, Li H, Jiang X, Qin D, Wang Q et al. Genome-wide association analysis identified molecular markers associated with important tea flavor-related metabolites. Hortic Res-England. 2021;8(1):42.
3. Park J, Park R, Jang M, Park Y-I. Therapeutic potential of EGCG, a green tea polyphenol, for treatment of coronavirus diseases. Life (Basel). 2021;11(3):197.
4. Baba Y, Inagaki S, Nakagawa S, Kaneko T, Kobayashi M, Takihara T. Effects of l-theanine on cognitive function in middle-aged and older subjects: a randomized placebo-controlled study. J Med Food. 2021;24(4):333–341.
5. Hu T, Wu P, Zhan J, Wang W, Shen J, Ho C-T et al. Influencing factors on the physicochemical characteristics of tea polysaccharides. Molecules. 2021;26(11):3457.
6. Barghouthy Y, Corrales M, Doizi S, Somani BK, Traxer O. Tea and coffee consumption and pathophysiology related to kidney stone formation: a systematic review. World Journal of Urology. 2021;39(7):2417–2426.

7. Hazra A, Dasgupta N, Sengupta C, Bera B, Das S. Tea: A worthwhile, popular beverage crop since time immemorial: Agronomic Crops; 2019.
8. Xia EH, Tong W, Wu Q, Wei S, Zhao J, Zhang ZZ et al. Tea plant genomics: achievements, challenges and perspectives. Hortic Res. 2020;7(1):7.
9. Kottawa-Arachchi JD, Gunasekare MTK, Ranatunga MAB. Biochemical diversity of global tea [Camellia sinensis (L.) O. Kuntze] germplasm and its exploitation: a review. Genet Resour Crop Ev. 2019;66(1):259–273.
10. Zhang W, Rong J, Wei C, Gao L-M, Chen J. Domestication origin and spread of cultivated tea plants. Biodiversity Science. 2018;26:357–372.
11. Han G, Liu C-Q: Water geochemistry of two large rivers in Guizhou Province, China: implications for crustal weathering and its controlling factors in karst region. In: Eleventh Annual VM Goldschmidt Conference: 2001. 3301.
12. Luo Z, Wu M, Yin Z. Analysis on the general situation and basic characteristics of river system in Guizhou Province. Jilin Water Resources. 2017;(12):29–32.
13. Weiwei L. The spatiotemporal development and reasons of tea planting in Guizhou during the Ming and Qing Dynasties. Journal of Guangxi Vocational and Technical College. 2018;v.11;No.66(06):41-44+69.
14. 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.
15. Chen L, Gao Q-k, Chen D-m, Xu C-jJB, Conservation. The use of RAPD markers for detecting genetic diversity, relationship and molecular identification of Chinese elite tea genetic resources [Camellia sinensis (L.) O. Kuntze] preserved in a tea germplasm repository. 2005;14(6):1433–1444.
16. Yan D, Liu S, Luo X, Jie W, Lu J, Fan F. Analysis of genetic diversity with RAPD markers for local tea populations in Guizhou. Chinese Agricultural Science Bulletin. 2015;31(19):30.
17. Tan LQ, Liu QL, Zhou B, Yang CJ, Zou X, Yu YY et al. Paternity analysis using SSR markers reveals that the anthocyanin-rich tea cultivar 'Ziyan' is self-compatible. Sci Hortic-Amsterdam. 2019;245:258–262.
18. Huang S, Wen L, Peng J, Zhang F, Tan Y, Long L et al. Genetic relationship analysis of wild tea tree germplasm resources in part of Guangxi based on EST-SSR markers. Guihaia. 2019.
19. Zhou Q, Sun W, Lai Z. Differential expression of genes in purple-shoot tea tender leaves and mature leaves during leaf growth. J Sci Food Agr. 2016;96(6):1982–1989.
20. 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.
21. Akhatar J, Goyal A, Kaur N, Atri C, Mittal M, Singh MP et al. Genome wide association analyses to understand genetic basis of flowering and plant height under three levels of nitrogen application in Brassica juncea (L.) Czern & Coss. 2021;11(1):1–14.

22. Pang Y, Liu C, Wang D, St Amand P, Bernardo A, Li W et al. High-resolution genome-wide association study identifies genomic regions and candidate genes for important agronomic traits in wheat. Mol Plant. 2020;13(9):1311–1327.

23. Kolkman JM, Strable J, Harline K, Kroon DE, Wiesner-Hanks T, Bradbury PJ et al. Maize introgression library provides evidence for the involvement of liguleless1 in resistance to northern leaf blight. G3-Genes Genomes Genet. 2020;10(10):3611–3622.

24. Feng S, Liu Z, Hu Y, Tian J, Yang T, Wei A. Genomic analysis reveals the genetic diversity, population structure, evolutionary history and relationships of Chinese pepper. Hortic Res. 2020;7(1):158.

25. Caballero M, Lauer E, Bennett J, Zaman S, McEvoy S, Acosta J et al. Toward genomic selection in pinus taeda: Integrating resources to support array design in a complex conifer genome. Appl Plant Sci. 2021;9(6):e11439-e11439.

26. Calleja-Rodriguez A, Pan J, Funda T, Chen Z, Baison J, Isik F et al. Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. Bmc Genomics. 2020;21(1):796.

27. 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;8:e8572.

28. Liu Y, Yi F, Yang G, Wang Y, Pubu C, He R et al. Geographic population genetic structure and diversity of Sophora moorcroftiana based on genotyping-by-sequencing (GBS). Peerj. 2020;8:e9609.

29. Dehgan B, Yuen CJBG. Seed morphology in relation to dispersal, evolution, and propagation of Cycas L. 1983;144(3):412–418.

30. Rubalcava-Castillo FA, Sosa-Ramirez J, Luna-Ruiz JD, Valdivia-Flores AG, Iniguez-Davalos LI. Seed dispersal by carnivores in temperate and tropical dry forests. Ecol Evol. 2021;11(9):3794–3807.

31. Mulder AJE, Aalderen R, Leeuwen CHA. Tracking temperate fish reveals their relevance for plant seed dispersal. Funct Ecol. 2021;35(5):1134–1144.

32. Maebara Y, Tamaoki M, Iguchi Y, Nakahama N, Hanai T, Nishino A et al. Genetic diversity of invasive Spartina alterniflora Loisel.(Poaceae) introduced unintentionally into Japan and its invasion pathway. Front Plant Sci. 2020;11:556039.

33. 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. Theoretical and Applied Genetics. 2021;134(1):191–212.

34. Sokolkova A, Burlyaeva M, Valiannikova T, Vishnyakova M, Schafleitner R, Lee CR et al. Genome-wide association study in accessions of the mini-core collection of mungbean (Vigna radiata) from the World Vegetable Gene Bank (Taiwan). Bmc Plant Biol. 2020;20.

35. 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.

36. Wang N, Yuan Y, Wang H, Yu D, Liu Y, Zhang A et al. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. Sci Rep. 2020;10(1):16308.

37. Delfini J, Moda-Cirino V, Neto JD, Ruas PM, Sant'Ana GC, Gepts P et al. Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. Sci Rep-Uk. 2021;11(1):1–12.

38. Yang X, Tan B, Liu H, Zhu W, Xu L, Wang Y et al. Genetic Diversity and Population Structure of Asian and European Common Wheat Accessions Based on Genotyping-By-Sequencing. Front Genet. 2020;11:580782.

39. Li WW, Liu LQ, Wang YA, Zhang QP, Fan GQ, Zhang SK et al. Genetic diversity, population structure, and relationships of apricot (Prunus) based on restriction site-associated DNA sequencing. Hortic Res-England. 2020;7(1).

40. Park S, Kumar P, Shi A, Mou B. Population genetics and genome-wide association studies provide insights into the influence of selective breeding on genetic variation in lettuce. The plant genome. 2021;14(2):e20086.

41. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL et al. Transversions have larger regulatory effects than transitions. Bmc Genomics. 2017;18(1):394.

42. Chen L, Zhou Z-X, Yang Y-JJE. Genetic improvement and breeding of tea plant (Camellia sinensis) in China: from individual selection to hybridization and molecular breeding. 2007;154(1):239–248.

43. 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.

44. 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. Evol Bioinform. 2021;17.

45. Slatkin MJE. Rare alleles as indicators of gene flow. Evolution. 1985;39(1):53–65.

46. Yang TY, Gao TX, Meng W, Jiang YL. Genome-wide population structure and genetic diversity of Japanese whiting (Sillago japonica) inferred from genotyping-by-sequencing (GBS): Implications for fisheries management. Fish Res. 2020;225.

47. 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.

48. Taniguchi F, Kimura K, Saba T, Ogino A, Yamaguchi S, Tanaka J. Worldwide core collections of tea (Camellia sinensis) based on SSR markers. Tree Genet Genomes. 2014;10(6):1555–1565.

49. Yang H, Wei CL, Liu HW, Wu JL, Li ZG, Zhang L et al. Genetic divergence between camellia sinensis and its wild relatives revealed via genome-wide SNPs from RAD sequencing. Plos One. 2016;11(3):e0151424.

50. 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(10):1938–1953.

51. Meegahakumbura MK, Wambulwa MC, Li MM, Thapa KK, Sun YS, Moller M et al. Domestication origin and breeding history of the tea plant (camellia sinensis) in China and India based on nuclear microsatellites and cpDNA sequence data. Front Plant Sci. 2017;8:2270.
52. Koech RK, Malebe PM, Nyarukowa C, Mose R, Kamunya SM, Joubert F et al. Functional annotation of putative QTL associated with black tea quality and drought tolerance traits. Sci Rep. 2019;9(1):1465.
53. Pascual L, Fernandez M, Aparicio N, Lopez-Fernandez M, Fite R, Giraldo P et al. Development of a multipurpose core collection of bread wheat based on high-throughput genotyping data. Agronomy-Basel. 2020;10(4).
54. Egbadzor KF, Ofori K, Yeboah M, Aboagye LM, Opoku-Agyeman MO, Danquah EY et al. Diversity in 113 cowpea [Vigna unguiculata (L.) Walp] accessions assessed with 458 SNP markers. Springerplus. 2014;3:541.
55. Liu S, Decroocq S, Harte E, Tricon D, Chague A, Balakishiyeva G et al. Genetic diversity and population structure analyses in the Alpine plum (Prunus brigantina Vill.) confirm its affiliation to the Armeniaca section. Tree Genet Genomes. 2021;17(1).
56. Bernard A, Barreneche T, Donkpegan A, Lheureux F, Dirlewanger E. Comparison of structure analyses and core collections for the management of walnut genetic resources. Tree Genet Genomes. 2020;16(5).
57. Sokolkova A, Burlyaeva M, Valiannikova T, Vishnyakova M, Schafleitner R, Lee CR et al. Genome-wide association study in accessions of the mini-core collection of mungbean (Vigna radiata) from the World Vegetable Gene Bank (Taiwan). Bmc Plant Biol. 2020;20(Suppl 1):363.
58. Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. Trends in genetics: TIG. 2001;17(9):481–485.
59. 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):1786–1788.
60. Liu C, Shringarpure S, Lange K, Novembre J. Exploring population structure with admixture models and principal component analysis. In: Statistical Population Genomics. Humana, New York, NY; 2020. 67–86.
61. 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(19):2633–2635.
62. Stecher G, Tamura K, Kumar S. Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol. 2020;37(4):1237–1239.
63. Xia R, Chen C, Ding p, Yao L, Lin H, Zhong C et al. The ancient Chinese road traffic, Guizhou Road history Beijing: Beijing: People's Communications Press; 1989.
64. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T et al. GSA: Genome Sequence Archive*. Proteom & Bioinf. 2017;15(1):14–18.

Figures

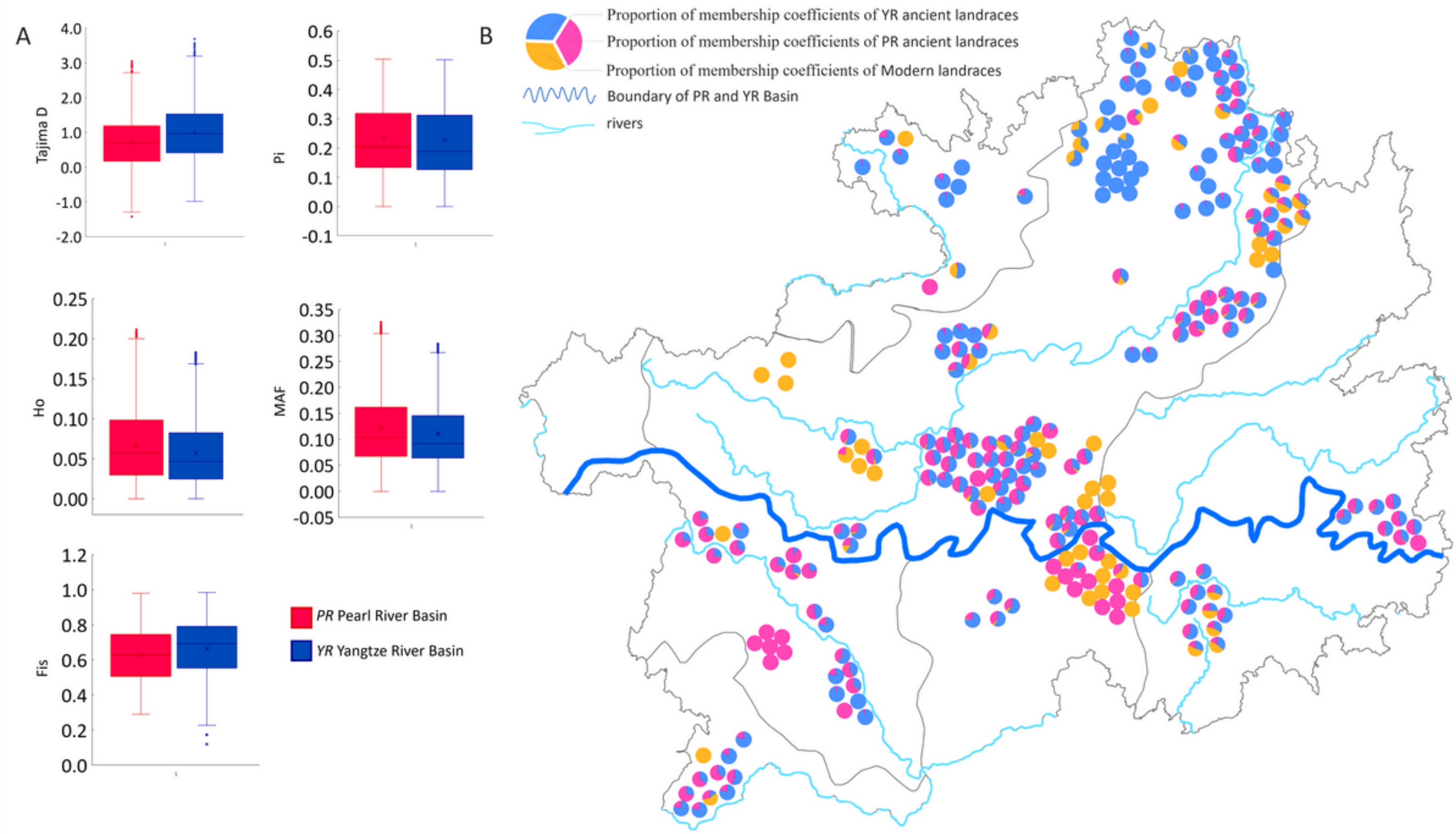

Figure 1

Geographic distribution and genetic diversity of 253 accessions. (A) Comparison of the genomic characteristics of two basins, PR the Pearl River Basin were showed in red and YR Yangtze River Basin were in blue. Pi heterozygosity, Ho observed heterozygosity, MAF minor allele frequency, Fis inbreeding coefficient. (B) Geographic distribution of each accession was represented by a pie chart of membership coefficient in ADMIXTURE on the Guizhou map [12]. For the three membership coefficients, CG-1 (the modern landraces group) was in yellow, CG-2 (the PR ancient landraces group) was in red and CG-3 (the YR ancient landraces group) was in blue in the pie chart.

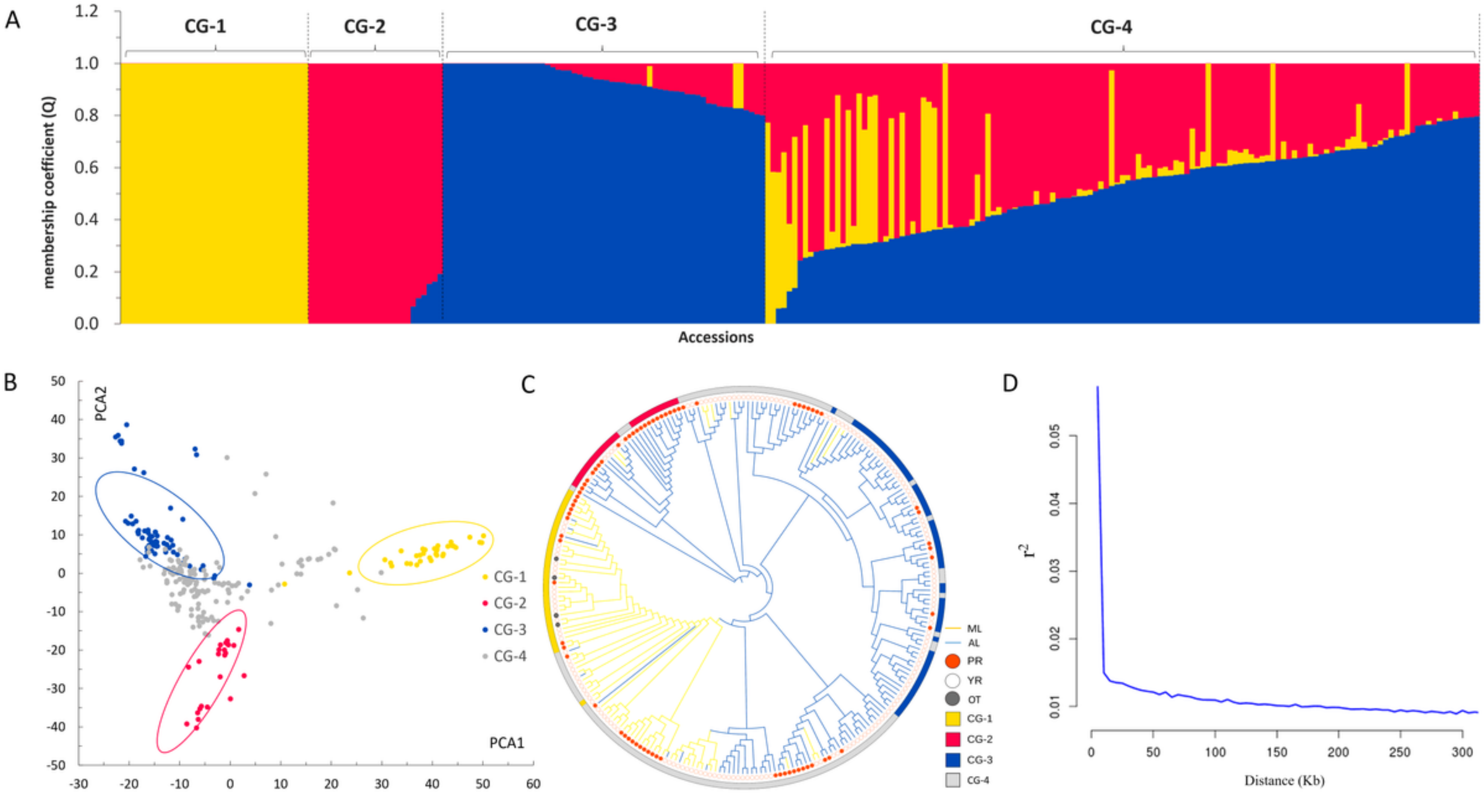

Figure 2

Population structure of 253 accessions (A) Inferred population structure of the collection using ADMIXTURE (v1.30). Bar plot of individual membership coefficients for the genetic clusters inferred using ADMIXITURE (K = 3) and the reduced dataset (112,070 SNPs data). Individual membership coefficients (Q) were sorted within each cluster. CG-1, CG-2 and CG-3 are shown in yellow, red and blue, respectively. (B) Principal component analysis (PCA). The four PCA scatter diagram was made by the first and second principal components. Four inferred populations were identified in ADMIXITURE, CG-1 was showed in yellow, CG-2 in red, CG-3 in blue and CG-4 in grey. (C)NJ tree compared the classification of inferred populations, cultivated statuses (ML modern landraces, AL ancient landraces and OT other) and two basins (PR the Pearl River Basin, YR Yangtze River Basin). (D) The LD decay plot of 253 accessions.

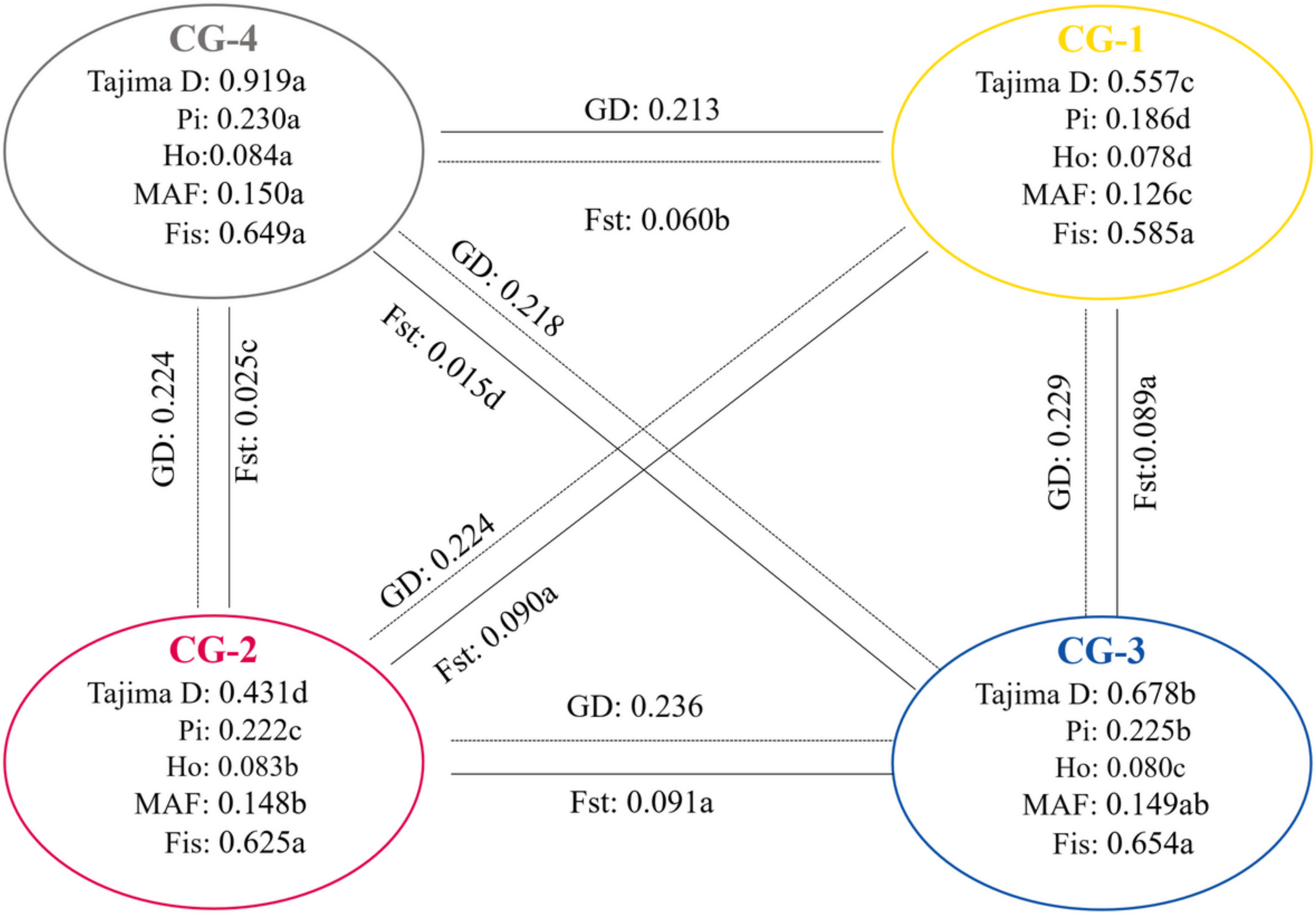

Figure 3

Genetic diversity of four inferred populations of 253 accessions. Pi heterozygosity, Ho observed heterozygosity, MAF minor allele frequency, Fis inbreeding coefficient, GD genetic distance Fst differentiation coefficient. The different letters indicate a significant difference in p= 0.05 levels by the T-test. CG-1, CG-2 and CG-3 are pure groups and the CG-4 is the admixture group base on ADMIXTURE software at K=3.

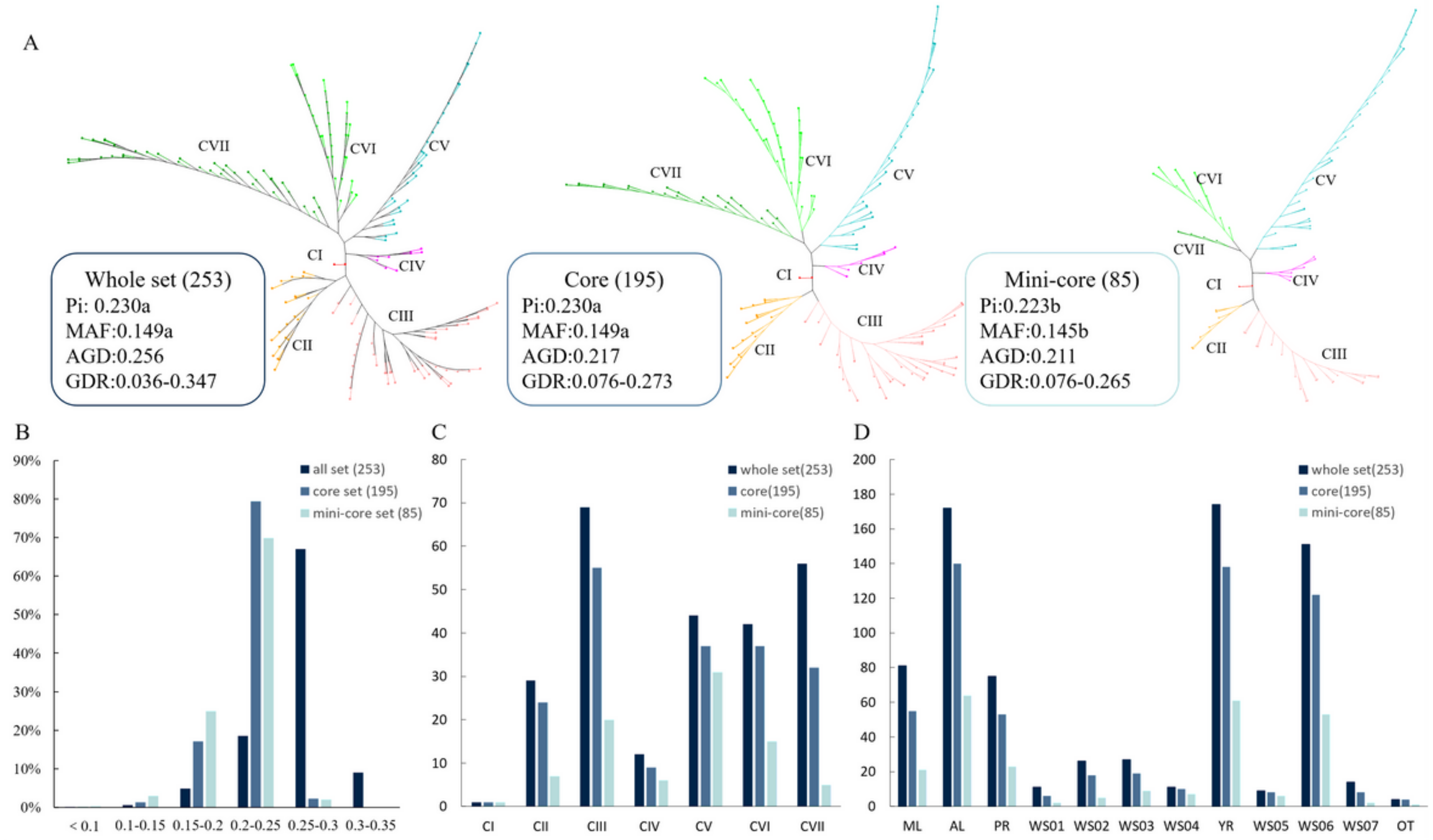

Figure 4

Summary of comparison information among core, mini-core and whole sets. (A) The NJ tree and genetic diversity of whole set, core set and mini-core set. Pi heterozygosity, MAF minor allele frequency, AGD average genetic distance, GDR genetic distance range. CI cluster I, CII cluster II, CIII cluster III, CIV cluster IV, CV cluster V, CVI cluster VI, CVII cluster VII. The different letters indicate a significant difference in p= 0.05 levels by the T-test; (B) Frequency distribution categories of pairwise genetic distance of 253 accessions. (C) The histogram of the numbers of accessions of whole set, core set and mini-core set in seven groups according to NJ tree. (D) The histogram of the numbers of accessions of whole set, core set and mini-core set in cultivated statuses (ML modern landraces, AL ancient landraces), two basins (PR the Pearl River Basin, YR Yangtze River Basin) and seven water systems (WS01 Liujiang River System, WS02 Hongshui River System, WS03 Beipanjiang River System, WS04 Nanpanjiang River System, WS05 Yuanjiang River System, WS06 Wujiang River System, WS07 Chishui River System; PR Pearl River Basin, YR Yangtze River Basin).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile2.pdf

- Additionalfile3.xlsx
- Additionalfile4.pdf
- Additionalfile5.pdf
- Additionalfile6.tif
- Additionalfile7.pdf
- Additionalfile8.txt
- Additionalfile9.txt