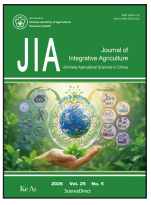
Journal of Integrative Agriculture 2026, 25(5): 1927–1938

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

Core germplasm construction of tea plant populations based on genome-wide SNP and catechins in Shaanxi Province, China

Xinyu Wang[1*], Xiufeng Li[2*], Dan Chen[1*], Jingwen Gao[1], Shuangqian Hao[1], He Zhang[1], Ziyan Zhao[1], Mengwei Shen[1], Huirui Chen[1], Fuqiang Qi[1], Keyi Zhang[1], Haozhe Zhou[1], Yanjun Xi[2], Jie Zhou[1], Youben Yu[1#], Qingshan Xu[1#]

[1] *College of Horticulture, Northwest A&F University, Yangling 712100, China*
[2] *Hanzhong Agricultural Technology Extension and Training Center, Hanzhong 723000, China*

Highlights

- A core germplasm representing the full genetic and metabolic diversity of tea plants was constructed by integrating genome-wide SNP and targeted metabolite data.
- The effectiveness of this core collection for genetic discovery was validated, as it successfully replicated all marker-trait associations identified in the original population.

Abstract

Genetic diversity is crucial to genetic research and crop breeding, and core collections are important resources for capturing this diversity. Recently, the core germplasm of tea plants was constructed mainly based on phenotypic data or molecular markers; however, the effective construction of core germplasm resources for plant breeding programs requires consideration of multiple aspects. In this study, we collected 320 tea germplasm resources and analyzed their single-nucleotide polymorphisms (SNPs) and metabolite data. Abundant genetic diversity in tea plants was inferred from the mean values of observed heterozygosity (Ho=0.340), expected heterozygosity (He=0.327), minor allele frequency (MAF=0.229), and polymorphic information content (PIC=0.268), based on the data from 2,118,060 high-quality SNP markers. A mean genetic diversity index (H′) value of 1.902 suggested significant metabolic variation. The 320 tea samples were categorized into six groups based on phylogenetic analysis, reflecting the influence of geographical origins on genetic diversity. Based on the genetic and metabolic data, a preliminary core collection of 106 accessions was developed to effectively represent most of the original panel's molecular, metabolic, population, and regional diversity. Genome-wide association studies of the core panel successfully replicated the marker-trait associations found in the original panel. This study contributes to the conservation and management of tea plant germplasm.

Keywords: tea plant, core collection, genetic diversity, SNPs, catechin index

1. Introduction

Tea [*Camellia sinensis* (L.) O. Kuntze] is a beverage crop highly regarded for its rich secondary metabolites, which are beneficial for human health (Liao *et al.* 2021; Zhang *et al.* 2021, 2025). The genetic diversity of tea plants is a crucial component of biodiversity and is essential for the success of tea breeding programs (Koorevaar *et al.* 2023). Self-incompatibility in tea plants is a key survival mechanism that enhances hybrid vigor and contributes to the maintenance of genetic diversity within the tea population (Lin *et al.* 2015; Zhao *et al.* 2022; Zhang *et al.* 2023; Jing *et al.* 2025). The Shaanxi tea region, one of the northernmost tea-producing areas in China, has a long history of tea cultivation. Historically, the complex terrain and limited transportation infrastructure in Shaanxi have hindered the cross-regional introduction and artificial improvement of tea varieties, resulting in the development of distinct regional populations and high-quality varieties. Based on geography, cultivation history, and chemical composition, these varieties can be broadly categorized into seven populations: Ziyang (ZY),

Received 30 September 2024; Received in revised form 1 January 2025; Accepted 23 March 2025; Available online 27 March 2025
[#]Correspondence Qingshan Xu, E-mail: xuqingshan@nwsuaf.edu.cn; Youben Yu, E-mail: yyben@163.com
[*]These authors contributed equally to this study.

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2025.03.024

Xixiang Daheba (XXDHB), Nanzheng Beiba (NZBB), Baihe Gefeng (BHGF), Kucha (KC), Ningqiang Guangping (NQGP), and Shanyang Manchuan (SYMC) (Hu 2012). The diverse tea germplasm resources exemplify distinct genetic diversity and represent a valuable natural gene pool for tea cultivation in northern China.

A core collection, representing a subset of germplasm resources that captures the maximum genetic, phenotypic, and geographic diversity with minimal kinship, is typically constructed using phenotypic data, molecular markers, or a combination of both (Korte *et al.* 2012; Kaur *et al.* 2023; Chen *et al.* 2024). Among these methods, using phenotypic data alone is the most straightforward approach for constructing core germplasm resources (Chen *et al.* 2024). Based on the phenotypic data of 18 traits from 104 individuals of *Juglans regia* L., 27 core germplasms were constructed (Mahmoodi *et al.* 2019). In another study, core germplasm was constructed from the phenotypic data of 780 *Pinus yunnanensis* individuals using a 30% sampling ratio (Wang *et al.* 2021). Molecular markers, such as random amplified polymorphisms (RAPDs), restriction fragment length polymorphisms (RFLPs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs), are commonly used to establish core collections because of their high stability and reliability, as demonstrated in sorghum (Upadhyaya *et al.* 2014), pomegranate (Wang *et al.* 2023), Chinese chestnut (Nie *et al.* 2021), and cucumber (Wang *et al.* 2018). It has also been reported that integrating phenotypic and genotypic data provides greater advantages over using either type of data alone, as it enables a more comprehensive construction of a core collection (Li *et al.* 2010). Based on genotypic data (SNPs) from 3,000 rice accessions and phenotypic data on 18 agronomic traits from 2,266 rice accessions, 520 core germplasm lines were constructed (Kumar *et al.* 2020). Analysis of SSR data and six primary coumarins from 208 *Angelica biserrata* accessions resulted in a core collection consisting of 42 accessions (Liu *et al.* 2020).

The germplasm of tea plants from various regions was collected and analyzed, resulting in multiple core germplasm collections being established. For example, a collection of 573 tea plant germplasms from Anhui Province, China yielded a core collection of 115 samples using SSR markers (Tao *et al.* 2023). A total of 788 tea plant accessions from 14 countries worldwide were collected, and a core collection of 192 tea plant accessions was developed based on SSR markers (Taniguchi *et al.* 2014). A total of 229 germplasms of *Camellia tunghinensis* from Guangxi Province were analyzed, creating a core collection consisting of 77 samples using SNP markers (Zhu *et al.* 2023). A core collection of 50 tea accessions was established from 468 accessions to represent genetic diversity using SNPs (Sun *et al.* 2024). Similarly, a core collection of 415 tea plants from Guizhou constituted 37% of the total collection obtained through SNP analysis (Niu *et al.* 2020). Despite these advances, there have been few reports that use both phenotypic data and molecular markers.

Effectively constructing a core germplasm resource for a plant breeding program requires consideration of various factors. The catechin index (CI) [(EC+ECG)/(EGC+EGCG)] has been used to assess genetic diversity by measuring the proportion of catechins in tea leaves (Saravanan *et al.* 2005; Yao *et al.* 2005; Magoma *et al.* 2020). This index provides a quantitative measure of the relative abundance of key catechins, which are essential for evaluating tea germplasm's overall quality and potential. Therefore, incorporating catechin analysis into the core collection development may offer additional insights. This study analyzed 320 locally cultivated tea germplasm resources from Shaanxi Province to assess their diversity and population structure. CI and SNP analyses were used to assess genetic diversity and generate data for constructing the core germplasm. Based on these analyses, a core germplasm collection of 106 individuals, representing all populations from each sampling region, was established. This core collection aids in better protecting and managing tea plant resources in Shaanxi Province and provides a reference for the effective utilization of tea plant resources.

2. Materials and methods

2.1. Plant materials

A total of 320 local germplasm varieties were collected from six tea populations in southern Shaanxi Province, China between 8:00 and 10:00 am on April 11, 2023 (Table 1). These samples were preserved at the Tea Sci-tech Demonstration Base of Northwest A&F University in Xixiang County (32°50′N, 107°40′E) and the Hanzhong Agricultural Technology Extension and Training Center (33°08′N, 107°04′E). Detailed information on these samples, including population type, accession name, geographical location, and sample type, is provided in Appendix A. Irrigation and fertilization practices in the two tea gardens were identical. Healthy young leaves (one bud and two leaves) were harvested, immediately frozen in liquid nitrogen, and stored at –80°C until use.

2.2. DNA extraction, amplification, and sequencing

Genomic DNA was extracted from fresh young leaf tissues (one bud and two leaves) using a modified cetyl trimethyl ammonium bromide (CTAB) method (Clarke 2009). DNA concentration and quality were measured using a NanoDrop 2000c (Thermo Scientific, MA, USA), and DNA integrity was

Table 1 Source and sample analysis of tea germplasm resources

Population type[1)]	Origin[2)]	Sample type	Sample size
NZBB	MX, Shaanxi	Landraces	2
	NZ, Shaanxi	Landraces	108
XXDHB	XX, Shaanxi	Landraces	92
	CG, Shaanxi	Landraces	7
ZY	ZB, Shaanxi	Landraces	31
	ZY, Shaanxi	Landraces	27
SYMC	SN, Shaanxi	Landraces	5
	SY, Shaanxi	Landraces	4
NQGP	NQ, Shaanxi	Landraces	39
BHGF	BH, Shaanxi	Landraces	5

[1)] NZBB, Nanzheng Beiba; XXDHB, Xixiang Daheba; ZY, Ziyang; SYMC, Shanyang Manchuan; NQGP, Ningqiang Guangping; BHGF, Baihe Gefeng.
[2)] MX, Mianxian; NZ, Nanzheng; XX, Xixiang; CG, Chenggu; ZB, Zhenba; ZY, Ziyang; SN, Shangnan; SY, Shanyang; NQ, Ningqiang; BH, Baihe.

assessed by 1.0% agarose gel electrophoresis. Following genomic DNA evaluation, the DNA was fragmented and subjected to end repair, A-base addition at the 3′ ends, and adapter ligation to prepare for PCR amplification and library construction, which were subsequently assessed for quality. The ligation products were denatured into single strands and circularized to obtain single-stranded circular DNA molecules, with non-circularized linear DNA being removed. Single-stranded circular DNA molecules were amplified using rolling circle replication to generate DNA nanoballs (DNBs) containing multiple DNA copies. The DNBs were loaded onto a high-density DNA nanochip with a nanowell array, and sequencing was performed using combinatorial probe-anchor synthesis (cPAS) technology.

2.3. Extraction and HPLC analysis of catechins

Tea catechins were isolated and detected using high-performance liquid chromatography (HPLC; LC-2030CD, Shimadzu, Japan) according to the Chinese National Standard GB/T 8313-2018 2018, with minor modifications. A ZORBAX SB-Aq C18 column (4.6 mm×250 mm, 5 μm, Agilent, USA) was used for chromatographic separation. The binary gradient elution system comprised 1% formic acid (solvent A) and acetonitrile (solvent B), with the following linear gradient elution procedure: 0–8 min, 5–15% B; 8–20 min, 15–20% B; 20–23 min, 20–27% B; 23–40 min, 27% B; 40–45 min, 27–85% B; 45–48 min, 85–5% B. The column temperature was set to 30°C, with an injection volume of 5 μL and a flow rate of 1 mL min^{-1}. Linear regression analysis was performed using the concentrations of the four catechin standard solutions as the independent variable (x) and the peak area as the dependent variable (y).

Standard substances EGCG (≥98% purity), ECG (≥98% purity), EGC (≥98% purity), and EC (≥98% purity) were obtained from Shanghai Yuanye Bio-Technology Co., Ltd. (Shanghai, China). LC-MS-grade methanol, acetonitrile, and formic acid were purchased from Sigma-Aldrich Corporation (Shanghai, China). Water was purified using a Milli-Q purification system (Millipore, Bedford, MA, USA).

2.4. SNP calling and filtering

Adapter sequences and low-quality reads in the raw sequences were filtered using Trimmomatic-0.36 (Bolger *et al.* 2014). The clean reads were mapped to the tea reference genome “Shuchazao v2.0” using BWA-MEM v0.7.17 with default parameters (Li and Durbin 2009). PCR duplicates were identified and removed using Picard (http://broadinstitute.github.io/picard/). SNPs were identified using GATK v4.2.3 (McKenna *et al.* 2010). High-quality variants were obtained following the method described by Leydon *et al.* (2017) and Zhang *et al.* (2020): (1) SNPs must be biallelic, (2) raw SNPs were filtered using GATK VariantFiltration with thresholds (QD<2.0||FS>60.0||MQ<40.0||MQRankSum<−12.5||ReadPosRankSum<−8.0), (3) SNPs within a 10 bp range of InDels were excluded, (4) SNPs occurring within a 10-base window with three or more instances were filtered out, (5) SNPs with a minor allele frequency (MAF)<0.1 or >10% missing data were filtered out using Plink v1.90 (Purcell *et al.* 2007). In total, 11,654,404 high-quality SNPs were retained from the initial set of 320 varieties for further analysis.

2.5. Analysis of population structure and genetic diversity parameters

Linkage disequilibrium (LD) analysis was conducted using the Plink v1.90 software tool with the parameters “indep-pairwise 50 10 0.2”. The resulting LD-filtered SNPs were then analyzed using Admixture v1.3.0 to determine the optimal K value based on minimum cross-validation (CV) error and to evaluate the population structure across K=2–12. A phylogenetic tree of the 320 individuals was constructed using the neighbor-joining (NJ) method based on an identity-by-state (IBS) kinship matrix and visualized using iTOL (https://itol.embl.de/). PCA was performed using Plink v1.90 (Purcell *et al.* 2007). K-means clustering of the metabolic data was performed using the R packages “stats” and “cluster”.

SNP filtering with Plink v1.90 using the parameters ‘--maf 0.1 --geno 0’ retained 2,118,060 SNPs. These SNPs were then used to calculate indices of genetic diversity and variation, including observed heterozygosity (Ho), expected heterozygosity (He), MAF, inbreeding coefficient (Fis), nucleotide diversity (π), and polymorphic information content (PIC), using Plink v1.90, VCFtools v0.1.16, and custom shell scripts (Purcell *et al.* 2007; Danecek *et al.* 2011). The genetic differentiation coefficient (Fst) was calculated for all loci within 500,000-bp non-overlapping windows across the genome of each population, using VCFtools v0.1.16. The resulting Fst values were then used to estimate the gene flow (Nm) between populations by applying the following equation: Nm=(1–Fst)/4Fst. Descriptive statistics for catechin compounds, including maximum, minimum, mean, standard deviation, range, and coefficient of variation, were computed using the R package “pastecs”. The genetic diversity index (H′) was calculated using Shannon’s index.

2.6. Construction of core collections

CoreHunter 3 has been widely used to extract representative samples from both phenotypic and genotypic data, effectively preserving the diversity of germplasm resources (Beukelaer *et al.* 2018). Based on SNP markers from 320 germplasm samples, gradient extraction of the core germplasm was performed with proportions of 0.35, 0.30, 0.25, 0.20, 0.15, and 0.10. The best representative and most genetically diverse subset was selected as Core1 by comparing genetic diversity parameters. Additionally, a candidate subset, Core2, was constructed from the phenotypic data using the CI values from 295 germplasm samples. The construction strategy aims to best represent the diversity of the original germplasm by minimizing the mean difference (MD) and variance difference percentage (VD) and maximizing the range compliance rate (CR) and coefficient of variation percentage (VR). The final core germplasm set was determined to exhibit high genetic diversity and accurately reflect the phenotypic diversity by integrating the results from both genotypic and phenotypic analyses.

To investigate significant differences between the core

collections and the original samples, the R package 'agricolae' was used to perform an LSD test with Bonferroni correction at a significance level of 0.05 on seven main genetic parameters (Ho, He, MAF, Fis, Pi, and PIC). Principal Coordinate Analysis (PCoA) of the original and core germplasms was performed to assess sample distributions. The significance of the differences in CI values among the different sets was evaluated using one-way ANOVA with SPSS 24.0 (IBM Corp 2016).

2.7. Genome-wide association study of core collections

All genome-wide association studies (GWAS) analyses were performed using TASSEL v5.2.93 (Bradbury *et al.* 2007) based on a mixed linear model (MLM) for the five tea catechin traits. The kinship matrix (*K*) was constructed using Tassel software and incorporated as a covariate in MLM to account for unequal relatedness among genotypes. The Q matrix, derived from the population structure (*K*=2), was used to control for population structure and reduce false positives. As previously described, SNP data and various phenotypic data were combined with their kinship (*K*) and Q matrices using MLM. Quantile-Quantile (Q-Q) plots were used to assess the fit between the model and the data, helping to evaluate the deviation between observed and expected values. Manhattan plots illustrate the correlation between SNP sites on each chromosome and the trait of interest. A threshold of $-\log_{10}(P)\geq 4.0$ was applied to identify SNP sites associated with the four catechins (EC, ECG, EGC, EGCG), while a more stringent threshold of $-\log_{10}(P)\geq 5.0$ was applied to pinpoint SNP sites closely linked to the CI value.

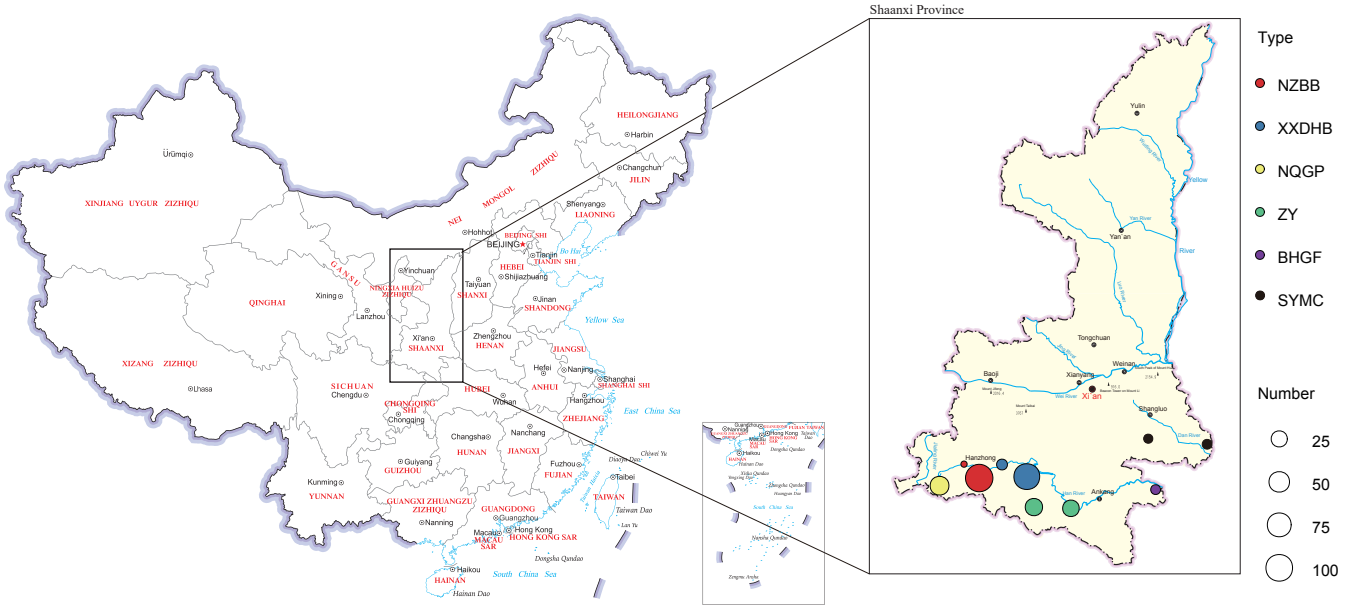
3. Results

3.1. Sequencing and SNP variant discovery

To obtain the SNP data, we conducted whole-genome resequencing (WGS) for 320 tea germplasm accessions representing six distinct populations from Shaanxi Province. (Fig. 1; Appendix A). After filtering the low-quality reads, 11 terabases (Tb) of clean data were generated, with an average of 35.2 gigabases (Gb) per accession and an average coverage depth of approximately 15×. All clean reads were mapped to the reference genome "ShuchazaoV2", resulting in the identification of 11,654,404 high-quality SNPs with MAF>0.05 and a missing data rate<10%, ensuring the reliability and accuracy of subsequent analyses on genetic diversity and population structure. These SNPs were unevenly distributed across the 15 tea chromosomes. Most SNPs were located on chromosomes 1, 2, and 4, with chromosome 15 containing the fewest (Appendix B). At the genomic level, 72.98% of SNPs were transitions, while 27.02% were transversions, resulting in a transition-to-transversion (ts/tv) ratio of 2.70. The frequencies of the two transition types were similar (36.48% for C/T and 36.50% for A/G), and the frequencies of the G/C, G/T, A/T, and A/C transversions were 4.65, 7.02, 8.35, and 7.01%, respectively (Appendix C).

3.2. Genetic diversity analysis of tea plants in Shaanxi Province

To characterize the genetic diversity, we calculated several metrics, including Ho, He, π, MAF, Fis, and PIC for SNPs using Plink, VCFtools, and shell scripts. For the entire population, the mean values of Ho, He, MAF, Fis, π, and PIC were 0.340, 0.327, 0.229, –0.040, 3.272×10^{-3}, and 0.268, respectively (Table 2). Further analysis of the six populations revealed Ho values ranging from 0.274 to 0.351 (mean=0.308), He values from 0.222 to 0.328 (mean=0.274),



审图号：GS 京 (2026) 0655 号

Fig. 1 Geographical distribution of 320 tea plant resources. The tea germplasm populations include the Nanzheng Beiba (NZBB), Xixiang Daheba (XXDHB), Ningqiang Guangping (NQGP), Ziyang (ZY), Baihe Gefeng (BHGF), and Shanyang Manchuan (SYMC). Circle colors represent different populations and circle sizes represent the number of collected samples.

MAF values from 0.167 to 0.231 (mean=0.197), and π values from 2.480×10^{-3} to 3.296×10^{-3} (mean=2.88×10^{-3}). The data suggest that the XXDHB, NZBB, ZY, and NQGP populations exhibit high genetic diversity, as indicated by the higher values across multiple metrics such as Ho, He, and π. In contrast, BHGF and SYMC populations show moderate genetic diversity (Table 2).

3.3. Population structure and genetic divergence of tea accessions

We analyzed the population structure of cultivars by evaluating the dynamic changes in the population structure across various *K* values (*K*=2–12). The admixture analysis classified the 320 accessions into two distinct ancestries using the optimal *K* model (*K*=2), as determined by the lowest cross-validation (CV) error score (Appendix D). The first ancestral group (Pop1) included 51 accessions, with BHGF and SYMC predominantly associated with this group. The second ancestral group (Pop2) comprised 269 accessions, including ZY (52 accessions), NQGP (31), XXDHB (89), NZBB (95), and SY (2). By increasing the Admixture *K* value to 5, Pop1 maintained a consistent structure with all individuals clustered within a single branch, whereas Pop2 was further subdivided into four distinct subgroups: Pop2-a, Pop2-b, Pop2-c, and Pop2-d (Fig. 2; Appendix E). Subcluster Pop2-a contained 52 accessions and was predominantly composed of ZY (36), XXDHB (18), and NQGP (16) populations. Pop2-b was the smallest Pop2 sub-cluster, containing only six accessions from the NZBB population. The largest subcluster, Pop2-c, contained 97 accessions and was primarily dominated by the XXDHB (65) population, followed by the subcluster Pop2-d, which contained 87 accessions and was mainly composed of the NZBB (77) population (Appendix F). This finding suggests that Pop1 and Pop2 exhibit distinct genetic differences, with Pop1 likely representing a group with a relatively uniform or homogeneous genetic background, whereas Pop2 comprises multiple subpopulations with greater genetic diversity.

The Fst analysis and Nm assessed the genetic differentiation among the six population groups. The results

Table 2 Genetic diversity analysis of different local population species[1)]

Type[2)]	Number	Ho	He	MAF	Fis	π	PIC
BHGF	5	0.274 g	0.222 g	0.167 f	–0.233 a	2.480×10^{-3} c	0.179 g
SYMC	9	0.299 f	0.270 f	0.194 e	–0.106 a	2.859×10^{-3} b	0.220 f
XXDHB	99	0.351 a	0.328 a	0.231 a	–0.069 a	3.296×10^{-3} a	0.269 a
NQGP	39	0.331 e	0.320 e	0.225 d	–0.035 a	3.235×10^{-3} a	0.262 e
NZBB	110	0.343 b	0.326 c	0.228 c	–0.052 a	3.265×10^{-3} a	0.267 c
ZY	58	0.337 d	0.325 d	0.228 bc	–0.037 a	3.272×10^{-3} a	0.266 d
ALL	320	0.340 c	0.327 b	0.229 b	–0.040 a	3.272×10^{-3} a	0.268 b

[1)] Ho, observed heterozygosity; He, expected heterozygosity; MAF, minor allele frequency allele frequency; Fis, inbreeding coefficient; π, nucleotide diversity; PIC, polymorphic information content.
[2)] BHGF, Baihe Gefeng; SYMC, Shanyang Manchuan; XXDHB, Xixiang Daheba; NQGP, Ningqiang Guangping; NZBB, Nanzheng Beiba; ZY, Ziyang.
Significant differences in genetic diversity parameters are represented by different letters, indicating the level of significance of the LSD test at a threshold of 0.05.

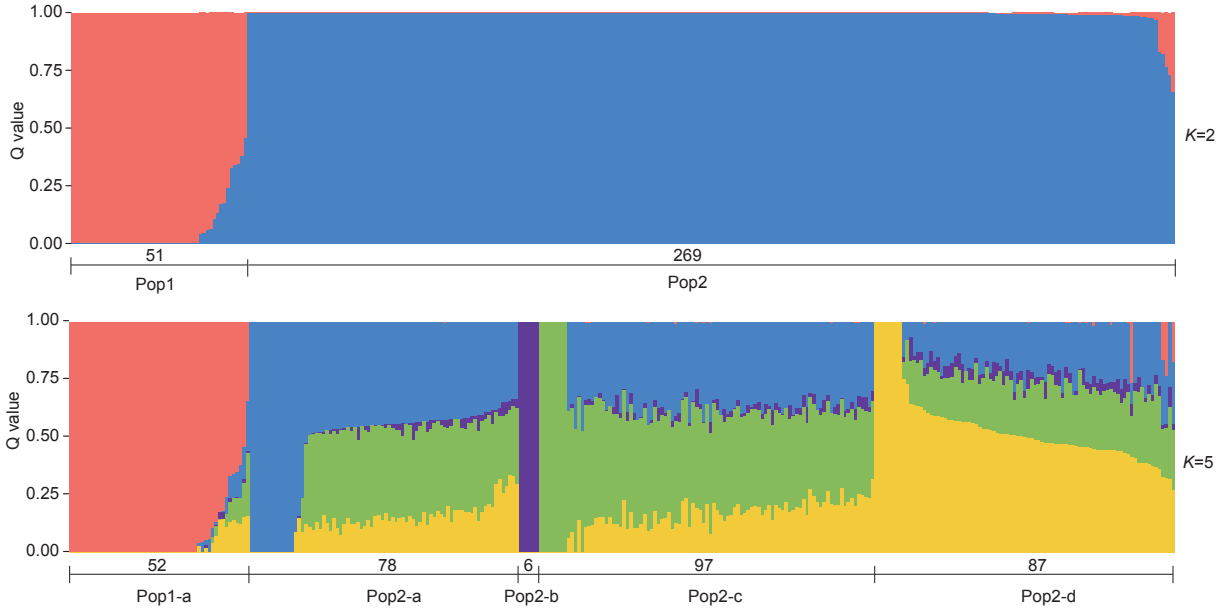


Fig. 2 Population structure analysis for each individual was inferred using Admixture (*K*=2, 5). The colors distinguish between different group structures.

revealed that the Fst value ranged from 0.0027 to 0.0746 (mean=0.028), while Nm values ranged from 3.10 to 91.97 (mean=27.34). The highest Fst value was observed in ZY *vs*. BHGF (0.0746), followed by XXDHB *vs*. BHGF (0.0741), NZBB *vs*. BHGF (0.0709), NQGP *vs*. BHGF (0.0634), and the corresponding Nm values were 3.10, 3.12, 3.28, and 3.70, respectively. These findings suggest that gene flow between these populations was limited and with moderate divergence. The minimum Fst value was recorded for ZY *vs*. XXDHB (0.0027), followed by NZBB *vs*. XXDHB (0.0042); the remaining population pairs had Fst values ranging from 0.0266 to 0.0048. These results suggest a slight divergence between these populations, with a higher Nm value reflecting frequent gene flow. These results demonstrate differential gene flow among the six tea populations in Shaanxi Province, highlighting the complex genetic structure of their germplasm resources (Table 3).

3.4. Phylogenetic relationship of tea plants in Shaanxi Province

A total of 320 accessions were analyzed using a neighbor-joining (NJ) phylogenetic tree to elucidate the relationships among the inferred populations, which were categorized into six distinct groups (Fig. 3-A). Group 1 contained 41 accessions including NQGP (22), NZBB (5), ZY (9), and XXDHB (5) populations. In Group 2, the ZY population accounted for 64% (28), and the XXDHB population constituted 18% (8). Group 3 contained 40 accessions; the XXDHB population constituted 82.5% (33), with contributions from NZBB (3), ZY (2), NQGP (1), and SYMC (1) comprising the remaining 17.5%. Group 4 predominantly comprised NZBB (73) and XXDHB (16), making it the branch with the highest concentration of tea plant resources. Group 5 consisted of three populations: XXDHB (15), NZBB (9), and ZY (1). Group 6 represents a diverse assemblage encompassing resources from all six populations. The numbers of accessions that constituted different groups in the neighbor-joining analysis are provided in Appendix G. Notably, all Pop1 individuals from the admixture analysis were classified within the same branch and clustered together, indicating close genetic relatedness (Fig. 3-A). Additionally, we observed that tea germplasms from the same geographic location tended to cluster within the same taxa, suggesting that geographic factors play a significant role in population clustering analysis. We also observed individuals from the same population but different regions clustered within the same branch. For example, most individuals from the ZY population originating from Zhenba and Ziyang Counties were grouped in Group 2. This suggests that genetic factors are also major contributors to population clustering analysis. PCoA was performed to validate the findings of the phylogenetic relationship analysis. The 320 original accessions were distributed along the first two coordinate axes, with PCoA1 explaining 19.99% of the total variation and PCoA2 accounting for 3.49% (Fig. 3-B). Individuals within each group were distinctly clustered, which was consistent with the results of phylogenetic tree analysis.

Table 3 Fst and Nm values between populations[1)]

	BHGF	SYMC	ZY	NQGP	XXDHB	NZBB
BHGF		18.3312	3.1008	3.6927	3.1248	3.2756
SYMC	0.0130 c		9.6034	19.9433	9.1772	9.6999
ZY	0.0746 a	0.0254 a		51.3496	91.9669	48.7312
NQGP	0.0634 b	0.0190 b	0.0048 a		44.3849	34.7199
XXDHB	0.0741 a	0.0266 a	0.0027 b	0.0056 b		59.0198
NZBB	0.0709 a	0.0251 a	0.0051 a	0.0071 a	0.0042 n	

[1)] Fst values (lower diagonal) and Nm values (upper diagonal) between different populations (BHGF, Baihe Gefeng; SYMC, Shanyang Manchuan; ZY, Ziyang; NQGP, Ningqiang Guangping; XXDHB, Xixiang Daheba; NZBB, Nanzheng Beiba).
Significant differences in Fst are indicated by different letters, representing levels of significance determined by the LSD test at a threshold of 0.05. Fst values of 0–0.05, 0.05–0.15, and 0.15–0.25 indicate little, moderate, and large divergence, respectively.

3.5. Cluster analysis of catechin compounds

To understand the variation in catechin compounds among different individuals, we performed statistical analysis on the catechin components (EC, ECG, EGC, and EGCG) and CI values of 295 cultivated tea germplasm samples selected from 320 sequenced samples. Among the four catechin components, EGCG was the most abundant (average 23,638.6 μg g^{-1}), followed by ECG (average 14,468.4 μg g^{-1}), EGC (average 4,297.0 μg g^{-1}), and EC (average 665.1 μg g^{-1}) (Appendix H-a). Continuous variation was observed across all traits, with normal or approximately normal distributions (Appendix H-b–f). The Shannon's genetic diversity index (H′) for the four catechins (EC, ECG, EGC, EGCG) ranged from 1.636 to 2.119, with an average of 1.969 (>1.85), indicating a high level of genetic diversity among these compounds. The coefficient of variation (CV%) ranged from 17.0 to 36.4%, with an average of 26.1%, indicating considerable variability in the catechin content across samples (Table 4). The CI values ranged from 0.301 to 1.192, with an average of 0.550 and a CV of 24.7%, and the H′ for CI was 1.636, reflecting moderate genetic diversity among the CI values in the tea germplasm samples.

Analysis of the metabolic data (four catechins and CI values) using K-means clustering with the Elbow method grouped the 320 accessions into three major clusters comprising 87, 94, and 114 individuals (Fig. 4-A; Appendix I). Each cluster contained individuals from Groups 1–6, as determined by neighbor-joining (NJ) analysis, indicating that the phylogenetic tree and catechin-based clustering did not fully align (Fig. 4-B). However, the distribution was not random. For instance, individuals from Groups 4 and 6 were predominantly clustered in clusters 2 and 3, respectively, which may suggest that these individuals share similar catechin compound metabolic profiles. Individuals from Groups 1, 2, 3, and 5 were evenly distributed across the three clusters, indicating that individuals within these groups exhibited greater genetic and metabolic diversity (Appendix J). These results revealed significant differences in both metabolite and genetic characteristics across the 320 tea accessions, suggesting that diversity within the core set may not be fully captured using only one type of data set.

3.6. Construction of core collection

To balance and capture the maximum possible phenotypic

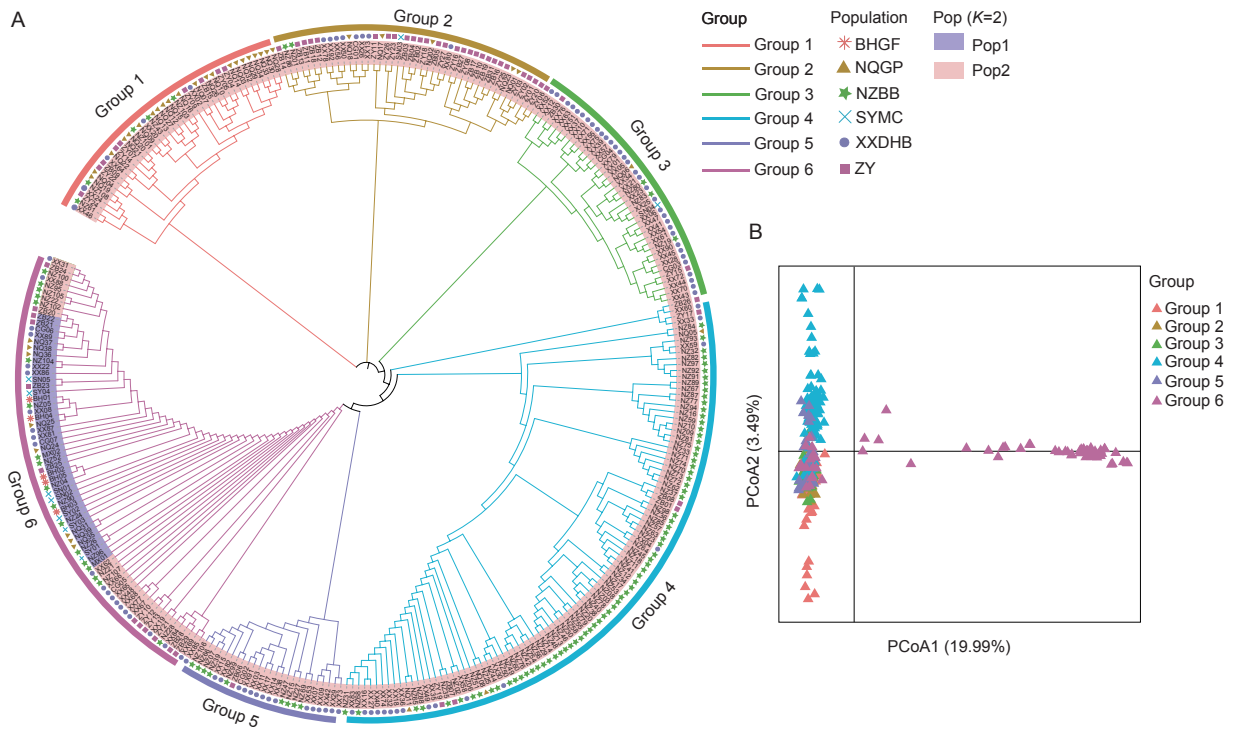


Fig. 3 Phylogeny and principal component analysis of local cultivated germplasm resources in Shaanxi Province, China. A, a phylogenetic analysis was conducted on 320 tea plants. Different colors represented distinct groups; various shapes indicated populations; background colors denoted different Pops. B, principal component analysis of the 320 accessions, showing principal component axes 1 and 2. The color codes represent different groups.

Table 4 Statistical analysis of catechin compounds in cultivated tea germplasm in Shaanxi Province, China

Trait	Mean	Max.	Min.	Range	SD	CV%	H′
EC (μg g^{-1})	665.1	1320.9	273.3	1,047.6	183.8	27.6	2.001
ECG (μg g^{-1})	14,468.4	24,349.8	6,132.7	18,217.1	3,365.5	23.3	1.933
EGC (μg g^{-1})	4,297.0	9,354.8	1,481.8	7,873.0	1,565.0	36.4	1.821
EGCG (μg g^{-1})	23,638.6	34,920.8	11,702.5	23,218.3	4,010.8	17.0	2.119
CI	0.550	1.192	0.301	0.891	0.139	24.7	1.636

and genotypic variabilities present in the original collection, we developed independent core collections based on genotypic data and CI values. Genotypic data (SNPs) from 320 tea plant germplasm accessions and CI values from 295 tea plant germplasm accessions were used to create representative core sets. For the SNPs-based subsets, the genetic diversity metrics, including Ho, He, MAF, PIC, and π, were higher than those of the original germplasm and superior to those of other candidate core sets at the 25% sampling ratio (Table 5). This indicated that a 25% sampling ratio (Core 1), encompassing 80 materials, was the most effective at eliminating redundant germplasm and preserving the genetic diversity of the original collection. For the subgroup based on CI, we selected an 11% construction proportion as Core 2, with an MD of 6.5%, VD of 42.5%, VR of 133.8%, and CR of 100% (Appendix K). Core 2 showed no significant differences in the catechin index compared to the original set, and the density map of Core 2 was closely mirrored with that of the original set, demonstrating that the proposed Core 2 can effectively capture the phenotypic diversity of the original set (Fig. 5).

Core 1 and Core 2 were merged to form the core collection Core-Merge with only six accessions shared between them, demonstrating that distinct accessions were selected based on phenotypic and genotypic variation (Fig. 6-A). The 106 accessions in the Core-Merge represented 33.3% of the original collection (320 accessions). Regarding genetic diversity, Core-Merge shows significant improvements over the original set in key parameters (He, MAF, Pi, PIC, π) (Table 5). This suggests that Core-Merge reduced redundant germplasm and effectively captured the genetic diversity of the original set. Regarding phenotypic diversity, Core-Merge captured the full range of CI traits present in the original collection, with a Shannon diversity index (H) of 1.68, demonstrating a significant improvement compared to Core 1 (H=1.47) and Core 2 (H=1.56) (Appendix L). Similar to Core 2, Core-Merge did not show a significant difference in the catechin index compared to the original set (Fig. 5), indicating that Core-Merge effectively preserved or captured the characteristics of the catechin index present in the original dataset. In contrast to Core 1 and Core 2,

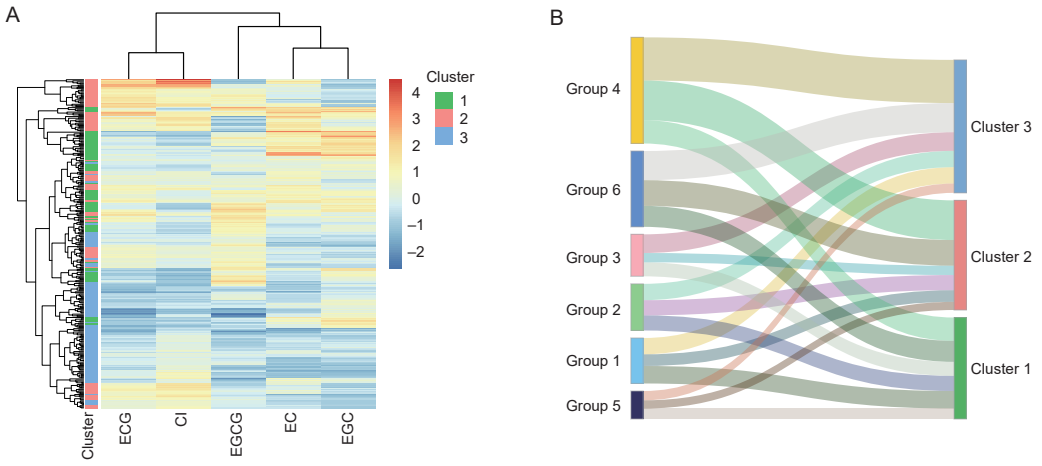


Fig. 4 Clustering and group distribution of tea plant individuals based on metabolic and phylogenetic analyses. A, heatmap of metabolite levels (ECG, CI, EGC, EC, and EGCG) showing clustering of 295 tea plants into three clusters. B, Sankey diagram showing the distribution flow of six groups of tea plants (Groups 1–6) into three clusters.

Table 5 Comparison of genetic diversity among 320 germplasm samples with different proportions[1)]

Candidate set	Ho	He	MAF	Fis	π	PIC
100%	0.3404 c	0.3273 d	0.2287 e	–0.0402 a	3.272×10^{-3} a	0.2683 c
35%	0.3413 b	0.3294 ab	0.2316 ab	–0.0361 a	3.303×10^{-3} a	0.2693 a
30%	0.3412 b	0.3293 ab	0.2317 ab	–0.0359 a	3.306×10^{-3} a	0.2692 a
25%	0.3419 a	0.3296 a	0.2319 a	–0.0373 a	3.311×10^{-3} a	0.2693 a
20%	0.3404 c	0.3292 b	0.2318 ab	–0.0340 a	3.313×10^{-3} a	0.2690 b
15%	0.3393 d	0.3285 c	0.2315 b	–0.0329 a	3.314×10^{-3} a	0.2684 b
10%	0.3342 e	0.3268 e	0.2307 c	–0.0228 a	3.315×10^{-3} a	0.2668 d
Core2	0.3314 f	0.3172 f	0.2227 f	–0.0446 a	3.218×10^{-3} a	0.2600 e
Core-Merge	0.3390 d	0.3283 c	0.2302 d	–0.0324 a	3.293×10^{-3} a	0.2688 b

[1)] Ho, observed heterozygosity; He, expected heterozygosity; MAF, minor allele frequency allele frequency; Fis, inbreeding coefficient; π, nucleotide diversity; PIC, polymorphic information content.
Significant differences in genetic diversity parameters are represented by different letters, indicating the level of significance of the LSD test at a threshold of 0.05.

the Core-Merge included all populations and regions to varying degrees within the 106 retained germplasms (Appendices M–O). In the PCoA, the core germplasm materials exhibited a distribution similar to that of the original germplasm and covered it comprehensively, representing its geometric distribution (Fig. 6-B). These results indicate that we have achieved the initial objective of developing a core collection (Core-Merge) that represents the maximum phenotypic, genotypic, population, and geographic variability present in the original collection of the 320 tea accessions.

3.7. Analysis of kinship in Core-Merge

An ideal core collection should minimize kinship among individuals to ensure that it fully represents the genetic diversity of the entire germplasm. Kinship analysis was conducted to assess genetic diversity within the Core-Merge. The results revealed that 75.09% of the possible pairs of Core-Merge accessions had kinship values of less than zero, while 23.10% had kinship values ranging from 0 to 0.25 (Fig. 7). Approximately 0.80% of the pairs exhibited kinship values between 0.25 and 0.50, and only six pairs had kinship values between 0.50 and 1.0. Therefore, the kinship values for most Core-Merge accessions indicated either a lack or a weak level of genetic relatedness, thereby meeting the criteria for an ideal core collection.

3.8. GWAS analysis of Core-Merge collection and original set

To assess the representativeness of the Core-Merge collection, we compared the number and co-location of significant SNPs associated with the five traits (CI, EC, ECG, EGC, and EGCG) between the original and Core-Merge collections using the MLM method. Q-Q plots were used to evaluate the disparity between the observed and expected *P*-values (Appendices P and Q). For the CI trait, with a significance threshold of $P<1.0\times10^{-5}$, the original collection identified 439 significant SNPs, while the Core-Merge collection contained 1,670 significant SNPs, including 62 co-located SNPs. These SNPs explained 6.8–14.1% and 20.6–34.7% of the phenotypic variation in the original and Core-Merge collections, respectively, indicating an enhanced representativeness of the Core-Merge collection. For ECG, at a significance level of $P<1.0\times10^{-4}$, the original collection had 641 significant SNPs compared to 362 in the Core-Merge collection, with five co-located SNPs. The SNPs accounted for 6.6–8.1% and 21.4–27.1% of the phenotypic variation, respectively, suggesting that

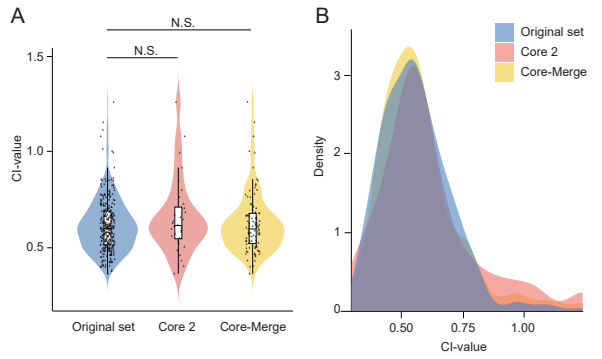


Fig. 5 Comparison of CI-value distribution between original set and core collection. A, distribution of CI-values in Core-Merge (*n*=106), Core 2 (*n*=32), and original set (*n*=295). B, density plot of CI-values in different sets. N.S., not significant.

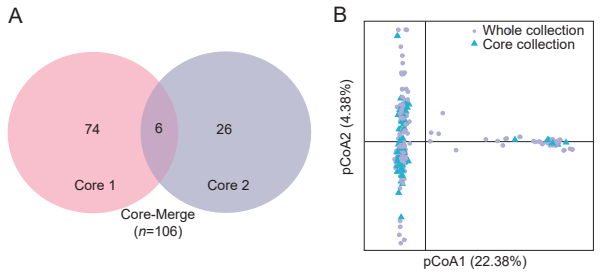


Fig. 6 Analysis of core germplasm distribution. A, the Venn diagram shows the distribution of different core collections developed in this study. B, principal coordinate analysis (PCoA) based on SNP marker genotypic data was conducted for the core germplasm (*n*=106) and the entire germplasm (*n*=320). Varieties in the core collection are represented in purple, while those in the entire collection are represented in blue.

although the Core-Merge collection has fewer SNPs, it retains important co-located SNPs. For the remaining traits (EC, EGC, and EGCG), the number of co-located SNPs was lower, indicating the reduced representativeness of the Core-Merge collection for these traits (Table 6). Overall, the Core-Merge collection effectively captured both genetic diversity and key trait representativeness.

4. Discussion

4.1. Genetic diversity and catechin variation of tea plants in Shaanxi Province

Genetic diversity is fundamental to the success of plant breeding programs, and core collections are critical resources for preserving this diversity (Koorevaar *et al.* 2023). Low genetic diversity in a species leads to diminished genetic variation and adaptability, threatening its long-term survival (Linde and Selmes 2012; Rodríguez-Nevado *et al.* 2018). Genetic diversity plays a crucial role in synthesizing secondary metabolites in plants (Yuan *et al.* 2010). Therefore, assessing the genetic diversity of plant populations is a critical step in constructing a core collection. Despite the long history and extensive cultivation of tea in the Shaanxi Province, the conservation and utilization of its rich genetic resources remain insufficient. This study collected 320 tea plant germplasm resources from six populations. A total of 2,118,060 SNP markers were used to analyze genetic diversity. The average values for Ho, He, MAF, and PIC were 0.340, 0.327, 0.229, and 0.268, respectively (Table 2), which are higher than those reported for tea populations in the Guizhou, Hunan, and Guangxi Provinces (Niu *et al.* 2020; Huang *et al.* 2022; Zhu *et al.* 2023). This indicates that tea plants in the region retain a high level of genetic diversity, which aligns with previous findings (Ban *et al.* 2021). Catechin composition is critical in the biochemical analyses of tea plant germplasm resources. (Gulati *et al.* 2009; Jin *et al.* 2014). The concentrations of the four major catechins (EC, ECG, EGC, and EGCG) were measured. The results revealed substantial variation in catechin composition (H′=1.902>1.85), highlighting the genetic variability within the tea plant germplasm populations. This rich diversity offers a broader genetic basis for constructing core collections and suggests the potential for selecting and breeding tea cultivars with desirable biochemical traits.

4.2. Clustering differences exist between genetic and metabolic data

Understanding the population structure can reveal genetic variations within germplasm resources, thus facilitating the development of more representative core collections (Brown 1989; Van Hintum *et al.* 2000). In this study, molecular markers and metabolic data were used to explore the population structure of tea plant germplasm resources. For molecular markers (SNPs), the NJ tree and PCA divided the 320 tea plants into six distinct groups. NQGP was primarily assigned to Group 1; ZY was predominantly located in Group 2; XXDHB was mainly clustered in Group 3; NZBB was chiefly classified into Group 4; and BHGF and SYMC were distributed across Group 6 (Fig. 3-A). On the other hand, based on the metabolic data, the 320 germplasm resources were grouped into three clusters based on five traits (EC, ECG, EGC, EGCG, and CI), which differed from the clustering obtained from the SNP-based analysis (Fig. 4). Environmental factors, including variations in temperature, precipitation, and light intensity, can significantly affect the composition and concentration of catechins in tea plants (Ahmed *et al.* 2019; Wang *et al.* 2022). Genetic factors also play a crucial role in determining catechin content by influencing the biosynthetic pathways and enzymatic activities (Eungwanichayapant and Popluechai 2009). As a result, the differences observed between clustering based on molecular markers and catechin traits may be because genetic and environmental factors influence the catechin content. The analysis of the differences between molecular markers and metabolite profile structures demonstrated that integrating these two data types facilitated the development of a more comprehensive core germplasm collection.

In addition, the SYMC and BHGF populations and the Mianxian region are limited in number in the present study; however, they represent a crucial component of the tea germplasm in Shaanxi Province. To understand the genetic diversity of tea plants in Shaanxi, we also sampled tea plants

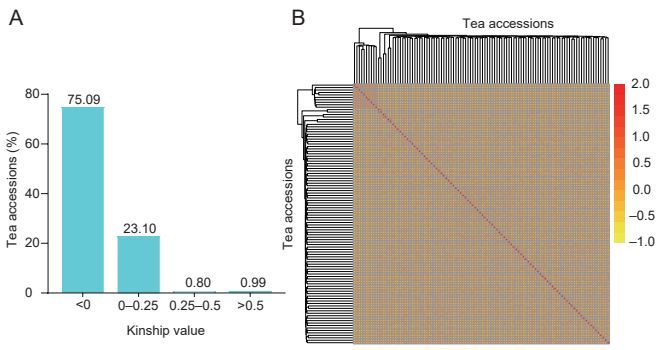


Fig. 7 Analysis of phylogenetic relationships in core sets. A, histogram displaying the phylogenetic relationships of the core collection of tea plants. B, display the kinship matrix of the core collection of tea plants.

Table 6 Comparison of significant SNPs and co-located SNPs between the original and Core-Merge collections for five traits

Trait	*P*-value	No. of SNPs in the original set	No. of SNPs in the Core-Merge	No. of co-located SNPS
CI	1.0×10^{-5}	439	1,670	62
EC	1.0×10^{-4}	956	55	0
ECG	1.0×10^{-4}	641	362	5
EGC	1.0×10^{-4}	1,022	136	3
EGCG	1.0×10^{-4}	459	113	1

from these populations and regions, inevitably introducing unbalanced sampling. However, unbalanced sampling can lead to misassigning individuals to incorrect groups (Fumagalli 2013; McLaughlin and Winker 2020). To address this issue, we employed multiple methods, including Admixture, NJ tree, PCA, and NJ trees with small sample sizes, to cross-validate the reliability and accuracy of our results (Fig. 3; Appendices R and S).

4.3. Integration of genetic and phenotypic data enhancing the core collections

A core collection is a strategically selected subset from a large germplasm bank designed to maintain the maximum genetic diversity (Frankel and Brown 1984). Shaanxi Province possesses rich tea germplasm resources and genetic diversity, which can be effectively conserved and utilized by establishing a core collection. Core collections have been successfully constructed for crops such as *Angelica biserrata* (Liu *et al.* 2020), tomatoes (Roohanitaziani *et al.* 2020), *Perilla* (Sa *et al.* 2021), and apples (Lassois *et al.* 2016). However, most studies focus primarily on molecular markers to construct core collections, emphasizing genetic diversity at the gene level while often overlooking phenotypic differences among varieties. There is a growing need to integrate molecular markers with morpho-physiological phenotypes to develop more precise core collections (Zhang *et al.* 2019; Kumar *et al.* 2020; Li *et al.* 2021; Le *et al.* 2024). For tea plants, intrinsic quality (in terms of the content of active ingredients) is an important phenotypic trait. As a crucial physiological component of tea, the catechin content directly influences its quality and value (Xu *et al.* 2018). It has been reported that the CI value, which is influenced by genetic background, can effectively distinguish genetic differences among tea germplasm resources (Jin *et al.* 2014). In addition, the dihydroxylated:trihydroxylated catechin ratio can be used as a marker for superior quality in tea breeding programmes (Magoma *et al.* 2020). Thus, based on the genetic diversity and population structure, we integrated genotypic data and CI values, utilizing Corehunter3 to construct a core collection that ensures both representation and diversity. The SNP-based candidate collection represented 25% of the original resources, while the catechins-based collection represented 11%. Notably, only 6 resources overlapped between these two datasets. This resulted in a core collection comprising 106 tea plant germplasm resources, representing 33.1% of the original collection (Fig. 6-A). While the core collection ratio is traditionally around 20% (Wang *et al.* 2011), we believe that the higher proportion in our study is justifiable. This is because the expanded core collection ensures that major genetic and phenotypic variants are encompassed and provides a more comprehensive assessment of tea plant germplasm resources. A series of results (Figs. 5, 6-B, and 7, and Table 5) further demonstrated that our integrated approach for constructing a core collection of tea plants based on genetic diversity and metabolic components is both feasible and reliable.

Core collections of plant genomes have been widely utilized in GWAS of agronomic traits (Zhao *et al.* 2018; Tanaka *et al.* 2020; Zhou *et al.* 2021). We compared the number of significant and co-localized SNPs for five traits between the original and Core-Merge collections, further validating the effectiveness of the Core-Merge collection in GWAS. Interestingly, only the CI trait showed a significant increase in the number of significant SNPs and co-localized SNPs within the Core-Merge collection, whereas the performance for the other traits was relatively less pronounced. This phenomenon may be attributed to the fact that the Core-Merge collection was constructed using both SNPs and CI, specifically accounting for the genetic characteristics of the CI trait. Although the CI trait is influenced by the ratios of the EC, ECG, EGC, and EGCG traits, reflecting some degree of association among them, this association may not fully capture the genetic diversity of these traits, resulting in their variation not being adequately represented in the Core-Merge collection. This provides an important insight; it is essential to consider specific traits in detail to achieve diverse breeding goals during the breeding process. Future research should integrate genetic data from multiple traits, optimize sample selection, and consider gene-environment interactions to enhance the representativeness and utility of core collections.

5. Conclusion

In this study, we employed a method combining SNPs and metabolite data to analyze tea germplasm resources' comprehensive genetic diversity and population structure. From an initial set of 320 tea accessions, we created a preliminary core collection of 106 Shaanxi tea accessions. This core collection effectively captured the majority of the

molecular, morphological, population, and regional diversity present in the original population, providing a more accurate representation of the generic variety found in the initial collection. The insights gained from this study are vital for conserving and utilizing tea genetic diversity, significantly contributing to the preservation, genetic research, and breeding of tea germplasm resources.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32472795), the Natural Science Basic Research Program of Shaanxi, China (2024JC-YBMS-145; 2021JQ-162), the China Agriculture Research System of MOF and MARA (CARS-19), the Agricultural Special Fund Project of Shaanxi Province, China (2024NYGG009), and Chinese Universities Scientific Fund (2452023481). The authors are grateful to Miss Jing Zhang, Miss Jing Zhao, and Miss Wenjing Cao (Horticulture Science Research Center, Northwest A&F University, Yangling, China) for providing professional technical assistance with HPLC analysis.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT 4.0 to assist in improving the clarity and readability of the language. All content generated was carefully reviewed and revised by the authors as necessary, and the authors take full responsibility for the final content of the article.

Appendices associated with this paper are available at https://doi.org/10.1016/j.jia.2025.03.024

References

Ahmed S, Griffin T S, Kraner D, Schäffner M K, Sharma D, Hazel M, Leitch A R, Orians C M, Han W Y, Stepp J R, Robbat A, Matyas C, Long C L, Xue D Y, Houser R F, Cash S B. 2019. Environmental factors variably impact tea secondary metabolites in the context of climate change. *Frontiers in Plant Science*, **10**, 939.

Ban Q Y, Pan Y T, Pan C, Hu X, Li Y Y, Jiang C J. 2021. EST-SSR analysis of genetic diversity in tea germplasm resources in Shaanxi Province. *Journal of Northwest A&F University*, **49**, 69–78. (in Chinese)

Beukelaer H D, Davenport G F, Fack V. 2018. Core Hunter 3: Flexible core subset selection. *BMC Bioinformatics*, **19**, 203.

Bolger A, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bradbury P J, Zhang Z W, Kroon D E, Casstevens T M, Ramdoss Y, Buckler E S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Brown A H D. 1989. Core collections: A practical approach to genetic resources management. *Genome*, **31**, 818–824.

Chen J H, Qu K L, Zhang Y C, Sun Y Q, Li B, Kang Y, Dong S J. 2024.. Construction of *Ziziphus jujuba* var. *spinosa* core collection based on phenotypic traits. *Journal of Shenyang Agricultural University*, **55**, 176–186. (in Chinese)

Clarke J D. 2009. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor Protocols*, **3**, 5177.

Danecek P, Auton A, Abecasis G R, Albers C A, Banks E, DePristo M A, Handsaker R E, Lunter G, Marth G, Sherry S T, McVean G, Durbin R. 2011. The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Eungwanichayapant P D, Popluechai S. 2009. Accumulation of catechins in tea in relation to accumulation of mRNA from genes involved in catechin biosynthesis. *Plant Physiology and Biochemistry*, **47**, 94–97.

Frankel O H, Brown A H D. 1984. Current plant genetic resources-a critical appraisal. In: *Genetics*: *New Frontiers*. *vol. 4*. Oxford & IBH Publishing, New Delhi. pp. 1–11.

Fumagalli M. 2013. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, **8**, 79667.

GB/T 8313–2018. 2018. *Method for the Determination of Catechins Content in Tea by HPLC*. Standards Press of China, Beijing. (in Chinese)

Gulati A, Rajkumar S, Karthigeyan S, Sud R K, Vijayan D, Thomas J, Rajkumar R, Das S C, Tamuly P, Hazarika M, Ahuja P S. 2009. Catechin and catechin fractions as biochemical markers to study the diversity of Indian tea (*Camellia sinensis* (L.) O. Kuntze) germplasm. *Chemistry and Biodiversity*, **6**, 1042–1052.

Van Hintum T J L, Brown A H D, Spillane C, Hodgkin T. 2000. *Core Collections of Plant Genetic Resources*. Bioversity International, Rome, Italy. pp. 24–30.

Hu X. 2012. Research progress on tea germplasm resources in Shaanxi. *Journal of Anhui Agricultural Sciences*, **40**, 51–52. (in Chinese)

Huang F Y, Duan J H, Lei Y, Liu Z, Kang Y K, Luo Y, Chen Y Y, Li Y Y, Liu S Q, Li S J, Liu Z H. 2022. Genetic diversity, population structure and core collection analysis of Hunan tea plant germplasm through genotyping-by-sequencing. *Beverage Plant Research*, **2**, 1–7.

IBM Corp. 2016. *IBM SPSS Statistics for Windows*. version 24.0. IBM Corp, Armonk, NY.

Jin J Q, Ma J Q, Ma C L, Yao M Z, Chen L. 2014. Determination of catechin content in representative Chinese tea germplasms. *Journal of Agricultural and Food Chemistry*, **62**, 9436–9441.

Jing X Y, Qian D D, Jiang X H, Wang P, Bao H H, Li D J, Zhu Y H, Zhu G T, Zhang C Z. 2025. Genetic diversity of the self-incompatibility locus in diploid potato. *Journal of Integrative Agriculture*, **24**, 1448–1460.

Kaur V, Gomashe S S, Aravind J, Yadav S K, Sheela, Singh D, Chauhan S S, Kumar V, Jat B, Tayade N R, Saroha A, Kaushik N, Langyan S, Singh M, Wankhede D P, Singh K, Kumar A, Singh G P. 2023. Multi-environment phenotyping of linseed (*Linum usitatissimum* L.) germplasm for morphological and seed quality traits to assemble a core collection. *Industrial Crops and Products*, **206**, 117657.

Koorevaar T, Willemsen J H, Visser R G F, Arens P, Maliepaard C. 2023. Construction of a strawberry breeding core collection to capture and exploit genetic variation. *BMC Genomics*, **24**, 740.

Korte A, Vilhjálmsson B J, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, **44**, 1066–1071.

Kumar A, Kumar S, Singh K B, Prasad M, Thakur J K. 2020. Designing a mini-core collection effectively representing 3004 diverse rice accessions. *Plant Communications*, **1**, 100049.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li S, Ji F F, Hou F F, Shi Q Q, Xing G M, Chen H, Weng Y Q, Kang X P. 2021. Morphological, palynological and molecular assessment of Hemerocallis core collection. *Scientia Horticulturae*, **285**, 110181.

Li X B, Yan W G, Agrama H, Hu B L, Jia L, Jia M, Jackson A, Moldenhauer K, McClung A, Wu D X. 2010. Genotypic and phenotypic characterization of genetic differentiation and diversity in the USDA rice mini-core collection. *Genetica*, **138**, 1221–1230.

Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, Poncet C, Lasserre-Zuber P, Feugey L, Durel C. 2016. Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. *Plant Molecular Biology Reporter*, **34**, 827–844.

Le L L, Yang X M, Xie X Y, Zhang W T, Wang G B, Cao F L. 2024. Construction of the core germplasm of yellowhorn (*Xanthoceras sorbifolium* Bunge) using physiological traits and SSR markers. *Scientia Horticulturae*, **323**, 112556.

Leydon A R, Weinreb C, Venable E, Reinders A, Ward J M, Johnson M A. 2017. The molecular dialog between flowering plant reproductive partners defined by SNP-Informed RNA-Sequencing. *The Plant Cell*, **29**, 984–1006.

Liao Y Y, Zhou X C, Zeng L T. 2021. How does tea (*Camellia sinensis*) produce specialized metabolites which determine its unique quality and function: A review. *Critical Reviews in Food Science and Nutrition*, **62**, 11–17.

Lin Z C, Eaves D J, Sánchez-Morán E, Franklin F H, Franklin-Tong V E. 2015. The *Papaver rhoeas S* determinants confer self-incompatibility to *Arabidopsis thaliana* in planta. *Science*, **350**, 684–687.

Linde C C, Selmes H. 2012. Genetic diversity and mating type distribution of *Tuber melanosporum* and their significance to truffle cultivation in artificially planted truffiéres in Australia. *Applied and Environmental Microbiology*, **78**, 6534–6539.

Liu M, Hu X, Wang X, Zhang J J, Peng X B, Hu Z G, Liu Y F. 2020. Constructing a core collection of the medicinal plant *Angelica biserrata* using genetic and metabolic data. *Frontiers in Plant Science*, **11**, 600249.

Magoma G, Wachira F, Obanda M, Imbuga M, Agong S. 2000. The use of catechins as biochemical markers in diversity studies of tea (*Camellia sinensis*). *Genetic Resources and Crop Evolution*, **47**, 107–114.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo M A. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Mahmoodi R, Dadpour M R, Hassani D, Zeinalabedini M, Vendramin E, Micali S, Nahandi F Z. 2019. Development of a core collection in Iranian walnut (*Juglans regia* L.) germplasm using the phenotypic diversity. *Scientia Horticulturae*, **249**, 439–448.

McLaughlin J F, Winker K. 2020. An empirical examination of sample size effects on population demographic estimates in birds using single nucleotide polymorphism (SNP) data. *PeerJ*, **8**, 9939.

Nie X H, Wang Z H, Liu N W, Song L, Yan B B, Xing Y, Zhang Q, Fang K F, Zhao Y L, Chen X, Wang G P, Qin L, Cao Q Q. 2021. Fingerprinting 146 Chinese chestnut (*Castanea mollissima* Blume) accessions and selecting a core collection using SSR markers. *Journal of Integrative Agriculture*, **20**, 1277–1286.

Niu S Z, Koiwa H, Song Q F, Qiao D H, Chen J, Zhao D G, Chen Z W, Wang Y, Zhang T Y. 2020. Development of core-collections for Guizhou tea genetic resources and GWAS of leaf size using SNP developed by genotyping-by-sequencing. *PeerJ*, **8**, 8572.

Purcell S, Neale B M, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P I W, Daly M J, Sham P C. 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Rodríguez-Nevado C, Lam T T, Holmes E C, Pagán I. 2018. The impact of host genetic diversity on virus evolution and emergence. *Ecology Letters*, **21**, 253–263.

Roohanitaziani R, Maagd R A, Lammers M, Molthoff J, Meijer-Dekens F, Kaauwen M, Richard Finkers R, Tikunov Y, Visser R G F, Bovy A. 2020. Exploration of a resequenced tomato core collection for phenotypic and genotypic variation in plant growth and fruit quality traits. *Genes*, **11**, 1278.

Sa K J, Kim D M, Oh J S, Park H, Hyun D Y, Lee S, Rhee J H, Lee J K. 2021. Construction of a core collection of native *Perilla* germplasm collected from South Korea based on SSR markers and morphological characteristics. *Scientific Reports*, **11**, 23891.

Saravanan M, John K M M, Kumar R R, Pius P K, Sasikumar R. 2005. Genetic diversity of UPASI tea clones (*Camellia sinensis* (L.) O. Kuntze) on the basis of total catechins and their fractions. *Phytochemistry*, **66**, 561–565.

Sun W H, Chen C L, Xu L L, Tao L L, Tong X Y, Tian Y P, Jiang H B, Chen L B, Wen H L, Liu S R, Wei C L, Zhu J Y. 2024. Genetic diversity analysis and core collection construction of tea plant from the Yunnan Province of China using ddRAD sequencing. *BMC Plant Biology*, **24**, 1163.

Tanaka N, Shenton M, Kawahara Y, Kumagai M, Sakai H, Kanamori H, Yonemaru J, Fukuoka S I, Sugimoto K, Ishimoto M, Wu J Z, Ebana K. 2020. Investigation of the genetic diversity of a rice core collection of Japanese landraces using whole-genome sequencing. *Plant and Cell Physiology*, **61**, 2087–2096.

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

Tao L L, Yang T, Chen H R, Wen H , Xie H, Luo L J, Huang K L, Zhu J Y, Liu S R, Wei C L. 2023. Core collection construction of tea plant germplasm in Anhui Province based on genetic diversity analysis using simple sequence repeat markers. *Journal of Integrative Agriculture*, **22**, 2719–2728.

Upadhyaya H D, Wang Y H, Gowda C L L, Sharma S. 2014. Erratum to: Association mapping of maturity and plant height using SNP markers with the sorghum mini core collection. *Theoretical and Applied Genetics*, **127**, 1461.

Wang M, Yang J, Li J L, Zhou X C,Yao Xiao Y, Liao Y Y, Tang J C, Dong F, Zeng L T. 2022. Effects of temperature and light on quality-related metabolites in tea [*Camellia sinensis* (L.) Kuntze] leaves. *Food Research International*, **161**, 111882.

Wang R X, Zhong Y H, Hong W J, Luo H W, Li D L, Zhao L, Zhang H Y, Wang J. 2023. Genetic diversity evaluation and core collection construction of pomegranate (*Punica granatum* L.) using genomic SSR markers. *Scientia Horticulturae*, **319**, 112192.

Wang X, Bao K, Reddy U K, Bai Y, Hammar S A, Iiao C, Wehner T C, Ramírez-Madera A O , Weng Y Q, Grumet R, Fei Z J. 2018. The USDA cucumber (*Cucumis sativus* L.) collection: Genetic diversity, population structure, genome-wide association studies, and core collection development. *Horticulture Research*, **5**, 1–13.

Wang X C, Chen L, Yang Y J. 2011. Establishment of core collection for Chinese tea germplasm based on cultivated region grouping and phenotypic data. *Frontiers of Agriculture in China*, **5**, 344–350.

Wang X L, Cao Z M, Gao C J, Li K. 2021. Strategy for the construction of a core collection for *Pinus yunnanensis* Franch. to optimize timber based on combined phenotype and molecular marker data. *Genetic Resources and Crop Evolution*, **68**, 3219–3240.

Xu Y Q, Zhang Y N, Chen J X, Fang W, Du Q Z, Yin J F. 2018. Quantitative analyses of the bitterness and astringency of catechins from green tea. *Food Chemistry*, **258**, 16–24.

Yao L, Caffin N, D'Arcy B, Jiang Y M, Shi J, Singanusong R, Liu X, Datta N, Kakùda Y, Xu Y. 2005. Seasonal variations of phenolic compounds in Australia-grown tea (*Camellia sinensis*). *Journal of Agricultural and Food Chemistry*, **53**, 6477–6483.

Yuan Q J, Zhang Z Y, Hu J, Guo L, Shao A J, Huang L. 2010. Impacts of recent cultivation on genetic diversity pattern of a medicinal plant, *Scutellaria baicalensis* (Lamiaceae). *BMC Genetics*, **11**, 29.

Zhang D Y, Li Y Y, Zhao X W, Zhang C L, Liu D K, Lan S,Yin W L, Zhong J L. 2023. Molecular insights into self-incompatibility systems: From evolution to breeding. *Plant Communications*, **5**, 100719.

Zhang H X, Bai R, Wu F,Guo W Q, Zhang Y, Qi Y, Zhang Y F, Ma J X, Zhang J Y. 2019. Genetic diversity, phylogenetic structure and development of core collections in *Melilotus* accessions from a Chinese gene bank. *Scientific Reports*, **9**, 13017.

Zhang W Y, Zhang Y J, Qiu H J, Guo Y F, Wan H, Zhang X L, Scossa F, Alseekh S, Zhang Q, Wang P, Xu L, Schmidt M, Jia X X, Li D L, Zhu A T, Guo F, Chen W, Ni D J, Usadel B, Fernie A R, Wen W W. 2020. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications*, **11**, 3719.

Zhang X T, Chen S, Shi L Q, Gong D P, Zhang S C, Zhao Q, Zhan D L, Vasseur L, Wang Y B, Yu J X, Liao Z Y, Xu X D, Qi R, Wang W L, Ma Y R, Wang P J, Ye N X, Ma D N, Shi Y, Wang H F, *et al.* 2021. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nature Genetics*, **53**, 1250–1259.

Zhao M Z, Song J Y, Wu A T, Hu T, Li J Q. 2018. Mining beneficial genes for aluminum tolerance within a core collection of rice tandraces through genome-wide association mapping with high density SNPs from specific-locus amplified fragment sequencing. *Frontiers in Plant Science*, **9**, 1838.

Zhao Y X, Zhao Z, Chen C G, Yu Y, Jeyaraj A, Zhuang J, Arkorful E, Thangaraj K, Periakaruppan R, Kou X B, Chen X, Li X H. 2022. Characterization of self-incompatibility and expression profiles of CsMCU2 related to pollination in different varieties of tea plants. *Scientia Horticulturae*, **293**, 110693.

Zhou X J, Guo J B, Pandey M K, Varshney P K, Huang L, Luo H Y, Liu N, Chen W G, Lei Y, Liao B S, Jiang H F. 2021. Dissection of the genetic basis of yield-related traits in the Chinese peanut mini-Core collection through genome-wide association studies. *Frontiers in Plant Science*, **12**, 637284.

Zhu X L, Zhou R, Qin H Z, Chai S F, Tang J L, Li Y Y, Xiao W. 2023. Genome-wide diversity evaluation and core germplasm extraction in ex situ conservation: A case of golden *Camellia tunghinensis*. *Evolutionary Applications*, **16**, 1519–1530.

Zhang C Y, Li H L, Mei P, Ye Y Y, Liu D D, Gong Y, Liu H R, Yao M Z, Ma C L. 2025. QTL detection and candidate gene analysis of the anthracnose resistance locus in tea plant (*Camellia sinensis*). *Journal of Integrative Agriculture*, **24**, 2240–2250.

Executive Editor-in-Chief Sanwen Huang
Managing Editor Lingyun Weng