frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 September 2025
DOI 10.3389/fpls.2025.1629553

OPEN ACCESS

EDITED BY
Sunchung Park,
Agricultural Research Service (USDA),
United States

REVIEWED BY
Chuanliang Deng,
Henan Normal University, China
Ezekiel Ahn,
United States Department of Agriculture,
United States

*CORRESPONDENCE
Jinhong Li
lijinhong1969@hotmail.com
Xuehui Bai
724180976@qq.com

RECEIVED 16 May 2025
ACCEPTED 18 July 2025
PUBLISHED 04 September 2025

CITATION
Jiang X, Liu C, Ma G, Zhao M, Li M, Chen T, Zhao P, Wang J, Luo Q, Guo T, Su L, Zhang Z, Wang J, Xiao Z, Xiao B, Zhou H, Li J and Bai X (2025) Population structure and genetic diversity of a coffee germplasm collection in China revealed by RAD-seq.
Front. Plant Sci. 16:1629553.
doi: 10.3389/fpls.2025.1629553

Population structure and genetic diversity of a coffee germplasm collection in China revealed by RAD-seq

Xinlei Jiang, Cheng Liu, Guanrun Ma, Mingzhu Zhao, Meifang Li, Tianming Chen, Pingxiang Zhao, Jingmin Wang, Qin Luo, Tieying Guo, Linlin Su, Zhirun Zhang, Jiayi Wang, Ziwei Xiao, Bing Xiao, Hua Zhou, Jinhong Li* and Xuehui Bai*

Yunnan Dehong Institute of Tropical Agricultural Science, Ruili, China

Coffee (*Coffea* spp.), a globally important crop, faces challenges in germplasm conservation due to habitat loss, climate change, and limited genetic diversity validation. This study aimed to evaluate the genetic representativeness of a coffee germplasm collection (CCGC, n=185) spanning major global varieties and wild relatives using re-striction-site associated DNA sequencing (RAD-seq). We performed genome-wide SNP profiling (37,729 loci), population structure analysis (STRUCTURE, PCA), and selection sweep detection (π) to assess genetic diversity, differentiation, and functional gene coverage. Results demonstrated that CCGC captured 98% of known disease-resistance loci (e.g., *SH3*, *RppM*) and exhibited high genetic diversity (π=0.1456, He=0.3014). Population structure analysis (K=3) identified three genetically distinct subgroups, among which Group 2 exhibited the highest diversity (He=0.3014, comparable to global coffee genetic resources) and encompassed all known *Hemileia vastatrix* resistance loci. The SNP density (7.5× higher than 5K SNP arrays) enabled precise identification of 47 selective sweep regions linked to domestication and adaptation. These findings validate CCGC as a genomically representative resource for coffee breeding and conservation. This work advances coffee genetic research by bridging resource preservation with molecular breeding strategies to address climate resilience and sustainable production.

KEYWORDS

***Coffea*, germplasm collection, RAD-seq, SNP markers, genetic diversity analysis, population structure analysis, sustainable agriculture**

1 Introduction

Coffee (*Coffea* spp.) is a perennial evergreen shrub or small tree in the *Rubiaceae* family, classified under the *genus Coffea* (Davis et al., 2011). The genus comprises approximately 124 species, with global coffee germplasm resources predominantly concentrated within the "coffee belt" (25°N to 25°S) (Davis et al., 2021). *Coffea arabica* an allotetraploid species (2n=4x=44 chromosomes), originated from hybridization between the diploid *Coffea canephora* and *Coffea eugenioides* (Anthony et al., 2002; Pagani et al., 2012; Cenci et al., 2012; Clarindo and Carvalho, 2008; Lashermes et al., 1999; Merot-L'anthoene et al., 2019). According to their genetic background and breeding methods, *Coffea arabica* varieties can be divided into three categories: *Bourbon*/*Typica* group, Ethiopian native group and gene infiltration group. The *Bourbon*/*Typica* group consists mainly of the traditional varieties *Bourbon*, *Typica* and their derivatives such as *Caturra*, which are known for their superior flavor but are less resistant to disease (Anthony et al., 2001). The Ethiopian native group is derived from local germplasm resources in Ethiopia, such as *Geisha*, and has unique flavor characteristics and rich genetic diversity (Zeru et al., 2012; Scalabrin et al., 2020). Introgression groups are disease-resistant varieties selected by interspecific hybridization, typically including the *Catimor* (*Caturra* × *Timor*) and *Sarchimor* (*Villa Sarchi* × HDT) series (Bertrand et al., 2005; Prakash et al., 2005). As a globally consumed beverage, coffee holds significant economic and cultural importance, serving as a vital pillar of agricultural economies and international trade for many developing countries (Davis et al., 2012). From Italian espresso to traditional Turkish brewing methods, coffee culture is deeply rooted around the world. Today, coffee culture has formed an important part of social interaction, artistic expression and lifestyle (Errington et al., 2012).

Nowaday, the sustainable development of the coffee industry faces severe challenges, particularly in the conservation of coffee germplasm resources and genetic diversity research (Van Der Vossen, 1985; Mekbib et al., 2022). As the foundation for breeding disease-resistant, stress-tolerant, and high-yield varieties, coffee germplasm resources are critical for addressing climate change, pest and disease threats, and diverse consumer demands (Mekbib et al., 2022). At present, there are many challenges in the conservation and utilization of coffee germplasm resources. (1) Habitat Loss: Over 30% reduction in wild coffee habitats globally, with climate change shrinking suitable cultivation areas by 1.2% annually (Moat et al., 2017; Bunn et al., 2015). Extreme weather events increasingly threaten coffee yield and quality (Bunn et al., 2015). (2) Biotic Stresses: Devastating outbreaks of coffee leaf rust (*Hemileia vastatrix*) and Coffee Berry Borer (*Hypothenemus hampei*) are exacerbated by declining genetic diversity, which weakens disease resistance (Avelino et al., 2015). (3) Genetic Erosion: Agricultural expansion and land-use changes have endangered wild coffee germplasm, pushing rare varieties toward extinction (Davis et al., 2021). These issues threaten both coffee industry sustainability and global supply chain stability (Pham et al., 2019). This rapid depletion of genetic options threatens to collapse the delicate balance between sustainable production and ecological preservation, potentially destabilizing a global supply chain that supports millions of livelihoods and satisfies evolving consumer demands for both quantity and quality (Moat et al., 2017). The stark reality is that without immediate, coordinated efforts to conserve and study remaining coffee genetic diversity, the industry risks being left defenseless against the combined onslaught of climate change, emerging pests and diseases, and shifting market requirements - challenges that diverse germplasm could help overcome if preserved and properly utilized (Bunn et al., 2015). This genetic diversity represents not just scientific interest but the very foundation upon which the coffee industry's climate adaptation strategies, disease resistance breeding programs, and quality improvement initiatives must be built to ensure both economic sustainability and ecological balance for generations to come (Moat et al., 2017). Strengthening the conservation, research, and utilization of coffee germplasm is thus essential for ensuring long-term industry viability and ecological balance (Moat et al., 2017).

Ethiopia, recognized as the center of origin for *Coffea arabica*, preserves approximately 99% of its wild genetic diversity (Davis et al., 2021; Ceja-Navarro et al., 2015). In China, coffee cultivation dates back over a century. The country has established extensive germplasm collections and conducted preliminary evaluations of these resources. However, most studies to date have focused on phenotypic characterization, with limited exploration of genetic traits (Zhou et al., 2015). The Chinese Germplasm Repository of Coffee RuiLi City, Ministry of Agriculture and Rural Affairs, stands as the nation's largest and most comprehensive coffee germplasm facility. It currently safeguards over 952 accessions, encompassing diverse cultivated varieties, wild relatives, and hybrid populations developed through both natural and artificial crosses. These materials provide a critical foundation for advancing genetic breeding research (Zhao et al., 2025). Core germplasm collections play a pivotal role in efficiently exploring and conserving genetic resources (Zhang et al., 2011). Such collections are carefully curated subsets of germplasm, designed to encapsulate maximum genetic diversity within a minimal sample size. This strategic approach serves as a cornerstone for effective germplasm management and utilization (Liu et al., 2020). Studies indicate that core collections typically represent 5% to 30% of total accessions, though this proportion varies across crops (Ndjiondjop et al., 2017). In coffee breeding, core collections offer a streamlined platform for identifying superior traits. For instance, breeders can rapidly screen for high yield, superior cup quality, disease resistance, or environmental stress tolerance. This efficiency significantly accelerates genetic improvement programs (Clifford and Willson, 1985; Vossen, 1985). However, existing coffee core collections—including those in Ethiopia and Brazil—face critical limitations. Many suffer from inadequate sampling (representing <15% of genetic diversity) or rely on low-resolution molecular markers (SSRs or 5K SNPs). These shortcomings hinder their ability to comprehensively capture genetic diversity (Gautam et al., 2004; Cunha Alves and Azevedo, 2018). To address these gaps, this study establishes a 185-accession Coffee Core Germplasm Collection (CCGC). The selection criteria prioritize geographic

representation (Kenya, Polundi, Cote d 'Ivoire, Colombia, Ethiopia, India, Portugal, etc), phenotypic diversity (disease resistance, productivity, flavor profiles), and historical contributions to breeding programs. The genomic representativeness of the CCGC is rigorously validated using Restriction-site Associated DNA sequencing (RAD-seq) technology.

Advances in biotechnology have established molecular techniques as the most reliable methods for germplasm characterization. RAD-seq offers a cost-effective genomic analysis approach. By utilizing restriction enzyme digestion or target-specific primers to enrich genomic regions of interest, this technique selectively sequences partial genomes, dramatically reducing both sequencing costs and data volume (Davey et al., 2011). RAD-seq efficiently generates single nucleotide polymorphisms (SNPs) and insertion-deletion (InDel) markers, making it ideal for large-scale genetic diversity studies (Peterson et al., 2012). Compared to whole-genome sequencing, RAD-seq maintains high resolution while significantly lowering costs and computational complexity, particularly advantageous for resource-limited species or projects (Elshire et al., 2011; Krishnan, 2022). In coffee germplasm research, RAD-seq enables rapid and economical identification of genetic variations, providing essential data for conservation and utilization (Krishnan, 2022). The evolution of next-generation sequencing technologies has made whole-genome sequencing more efficient and affordable than ever, unlocking opportunities to detect extensive DNA polymorphisms (Arai-Kichise et al., 2011). SNPs-the most prevalent genomic variations—refer to single-base substitutions (Bhattramakki et al., 2002). To date, SNP markers have become central to molecular assays due to their compatibility with high-throughput automated platforms (Jangra et al., 2021). Meanwhile, InDels (1–50 bp insertions or deletions) are gaining recognition for their growing importance in genetic variation (Salathia et al., 2007). Both markers are well-suited for genetic evaluation and selective breeding strategies using molecular genetics. Coffee genetic diversity studies face notable technical limitations. First, molecular marker applications remain oversimplified, with 82% of published studies relying solely on SSR markers (Yan et al., 2019). However, SSRs struggle to deliver precise analyses for large sample sets—a gap addressed by RAD-seq through its genome-wide high-throughput SNP coverage. For instance, the coffee single nucleotide polymorphism (SNP) chip developed by Cheng et al. contains only 5,000 loci, which is significantly lower than that of crops such as cocoa (30,000 SNPs) (Merot-L'anthoene et al., 2019; Guo et al., 2021). Secondly, the research on the core germplasm resources of coffee worldwide is still lagging behind. Currently, only Ethiopia and Brazil have established regional germplasm banks (Gautam et al., 2004; Cunha Alves and Azevedo, 2018). Yet these collections inadequately sample genetic diversity (<15% of total resources) and rely on outdated evaluation systems dominated by morphological markers (~30% of studies) and basic genetic parameters (e.g., Nei's diversity index), lacking advanced genomic approaches like genome-wide association studies (GWAS) (Otyama et al., 2019). In China, Huang Lifang's research on coffee diversity using RAPD technology is a pioneering work. This study not only analyzed the genetic relationships but also confirmed the applicability of this marker in variety identification (Huang et al., 2014, 2017, 2016). Parallel advances in other crops offer valuable insights: Li et al. integrated multiple SNP markers from whole-genome data of 117 rice accessions to construct a fingerprinting system that minimizes false positives/negatives and enables rapid identification (Li et al., 2020). Similarly, Peng et al. developed Indel markers for gene mapping, successfully characterizing rice male sterile lines (Liu et al., 2017). Despite the widespread adoption of molecular markers in germplasm evaluation, existing coffee core collections still lack genome-wide validation of their genetic representativeness. Based on the current technological development, reduced-representation sequencing has emerged as an ideal method for analyzing the genetic diversity of coffee due to its efficiency, economy and universality. This study addresses this critical knowledge gap by employing RAD-seq to systematically dissect genome-scale genetic diversity within coffee core germplasm resources.

This study aims to comprehensively analyze the genetic diversity of 185 coffee germplasm accessions using RAD-seq and validate the germplasm collection. For the first time, we systematically characterized the genome-wide genetic diversity of coffee core germplasm resources, providing a scientific foundation for coffee breeding and resource conservation. By applying RAD-seq technology, we identified 37,729 loci, achieving a marker density 7.5 times higher than that of existing coffee core collections (5K SNP array) and accomplishing the first genome-wide validation of genetic representativeness. The research not only fills the knowledge gap in genetic diversity studies of coffee germplasm resources in China but also offers scientific guidance for their classification, conservation, and innovative utilization. Furthermore, it provides an innovative solution for the global conservation and sustainable use of coffee genetic resources.

2 Materials and methods

2.1 Plant material

This study utilized 185 accessions of *Coffea arabica* L. germplasm maintained at the Chinese Germplasm Repository of Coffee RuiLi City, Ministry of Agriculture and Rural Affairs, representing three principal genetic groups (*Bourbon*/*Typica*, Ethiopian native, and Introgression group) collected from major coffee-producing countries including Kenya, Burundi, Côte d'Ivoire, Colombia, Ethiopia, India, and Portugal (Supplementary Table S1). The coffee samples were sourced from 7 major producing countries and exhibited significant geographical diversity. Stratified sampling by region was employed. The 952 samples were divided into 7 layers based on the country. The original number of samples in each layer was recorded, and the number of samples to be selected from each layer was calculated based on its proportion in the overall population. Then, simple random sampling was independently used to select the target number of samples from each layer. The total number of final samples in each layer was checked to be 185, ensuring no omissions or repetitions, and the geographical distribution was consistent with the overall population. The germplasm collection exhibits extensive geographic diversity, significant phenotypic variation, and high

genomic representation, making it particularly suitable for investigating genetic diversity patterns, disease resistance traits, and quality improvement potential in *Arabica* coffee breeding programs.

2.2 DNA extraction and quality control

Genomic DNA was extracted using a modified CTAB method. Fresh leaf tissues were ground into powder using liquid nitrogen. CTAB extraction buffer (2% CTAB, 100 mM Tris-HCl, 20 mM EDTA, 1.4 M NaCl, 0.2% β-mercaptoethanol) was added, followed by incubation in a 65°C water bath for 1 hour. The mixture was extracted with chloroform-isoamyl alcohol (24:1) and centrifuged to collect the supernatant. An equal volume of isopropanol was added to precipitate DNA, which was then washed with 70% ethanol and dissolved in TE buffer.

The extracted DNA was quality-controlled through three approaches: integrity verification via agarose gel electrophoresis, purity assessment using Nanodrop (OD260/280 ratios between 1.8-2.2), and quantification using Qubit 3.0 fluorometer (concentration ≥50 ng/μL, total yield ≥2 μg). These quality parameters ensured the DNA met specifications for subsequent library construction.

2.3 Reduced-representation genome library construction and sequencing

The ddRAD-seq (Double Digest Restriction-site Associated DNA sequencing) protocol was performed following the standardized workflow by Peterson et al (Peterson et al., 2012).

2.3.1 Genomic DNA digestion

High-frequency cutter MseI (5′-TTAA-3′) and low-frequency cutter SacI (5′-GAGCTC-3′) (New England Biolabs, Ipswich, MA, USA) were used for double-digestion of 200 ng high-quality genomic DNA (OD260/280 = 1.8-2.0) in 1× CutSmart Buffer with 0.5 U/μL enzyme concentration. Reactions were incubated at 37°C for 2 h to ensure complete digestion.

2.3.2 Adapter ligation and purification

Digested DNA fragments were ligated to custom-designed double-stranded adapters (containing 8 bp sample-specific barcodes and Illumina P5/P7 sequencing adapters) using 1× T4 DNA Ligase Buffer (Thermo Fisher Scientific, Waltham, MA, USA), 0.5 μM adapters, and 400 U T4 DNA ligase (Thermo Fisher Scientific). Ligation proceeded at 16°C for 12 hr. Ligated products were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) to remove unbound adapters and residual fragments.

2.3.3 PCR amplification and size selection

Purified DNA was amplified for 18 cycles using KOD-Plus-Neo high-fidelity DNA polymerase (TOYOBO, Osaka, Japan) with a 65°C annealing temperature. Amplified products were separated on 1% low-melting-point agarose gels (Bio-Rad Certified Megabase Agarose in 1× TAE buffer), and fragments of 300–400 bp were excised and purified using the QIAquick Gel Extraction Kit (QIAGEN, Hilden, Germany).

2.3.4 Library quality control

Purified libraries were quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific; ≥2 nM). Insert size distribution (300–400 bp, CV <5%) was validated on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Effective library concentration (≥2 nM) was determined via KAPA Library Quantification Kit (Roche, Basel, Switzerland) on an ANALYTIKJENA qTOWER real-time PCR system (Jena, Germany).

2.3.5 High-throughput sequencing

Qualified libraries were pooled at equimolar ratios and sequenced on an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA) with 150 bp paired-end (PE150) reads. Each sample generated an average of 4.7 Gb raw data (Q30 ≥94.25%). Raw sequencing data in FASTQ format retained sample-specific barcodes for downstream demultiplexing.

2.4 Bioinformatics analysis

2.4.1 Data quality control and filtering

The raw sequencing data were processed using fastp (version 0.23.0) for quality control with the following parameter settings: adapter sequences were automatically detected and trimmed using the default adapter library; low-quality bases were filtered using a sliding window approach with a window size of 4 bp, trimming regions where the average Phred quality score fell below 15; and only reads ≥50 bp in length were retained. The resulting high-quality clean reads were used as the basis for subsequent analysis.

2.4.2 Sequence alignment

Quality-controlled reads were aligned to the coffee reference genome (ET-39, NCBI GenBank: GCA_036785885.1) using BWA-MEM (version 0.7.15) with default parameters (Salojärvi et al., 2024). The alignment results were output as SAM files, which were subsequently converted to BAM format, sorted, and indexed using samtools (version 1.3.1) to facilitate variant calling.

2.4.3 Variant detection

Genetic variant identification was conducted following the Genome Analysis Toolkit (GATK; version 3.7) workflow. The analysis pipeline comprised three sequential steps: (1) individual variant calling was performed using the HaplotypeCaller algorithm in GVCF mode, (2) joint genotyping across all samples was executed with the GenotypeGVCFs module, and (3) stringent quality filtering was applied. For SNP variants, we implemented the following filtering criteria: QD < 2.0, FS > 60.0, MQ < 40.0, SOR > 3.0, MQRankSum < -12.5, or ReadPosRankSum < -8.0. Indel variants were filtered using: QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, or SOR > 10.0. Additional population-level quality thresholds included: minimum read depth (DP) ≥ 3, genotype missing rate ≤ 50%, minor allele

frequency (MAF) ≥ 5%, and maximum sample heterozygosity ≤ 60%. The final high-confidence variant set was output in VCF format for subsequent population genetic analyses.

2.4.4 Variant annotation

Detected variants were functionally annotated to characterize: SNP types (transitions/transversions), InDel length distribution, Genomic density and distribution patterns. Annotation results were saved in ANN format, providing comprehensive variant information for downstream population genetic analyses.

2.4.5 Population genetic analyses

Multiple approaches were employed to validate the global coffee genetic diversity coverage of the germplasm collection and assess its genetic structure rationality.

2.4.6 Population structure analysis

Genetic structure was inferred using STRUCTURE (version 2.3.4) with K-values ranging from 2 to 10. Ten independent runs were performed for each K using the default parameters: a burn-in period of 10,000 iterations followed by 100,000 Markov chain Monte Carlo (MCMC) replications under the admixture model. The optimal K-value was determined using the ΔK method, and output files were processed to visualize subpopulation clustering patterns.

2.4.7 Principal component analysis

Genetic differentiation among populations was assessed using PLINK (version 1.9). A covariance matrix derived from genotype data was subjected to eigenvalue decomposition, and principal components were computed. The PCA results were stored in EIGENSTRAT format, enabling multidimensional scaling visualization of genetic relationships between populations.

2.4.8 Phylogenetic tree construction

The phylogenetic analysis was performed using the neighbor-joining (NJ) method in MEGA11 with the following parameters: (1) genetic distances were calculated using the Kimura 2-parameter (K2P) substitution model, which was selected based on the lowest Bayesian Information Criterion (BIC) score in MEGA's model test module; (2) rate variation among sites was modeled using a gamma distribution (shape parameter = 1.0) with invariant sites; and (3) nodal support was assessed with 1,000 bootstrap replicates. The NJ tree construction was conducted with pairwise deletion of gaps/missing data. The final tree topology, including branch lengths and bootstrap values, was exported in Newick format for evolutionary relationship analysis among coffee populations.

2.4.9 Genetic diversity and differentiation analysis

Genetic diversity parameters—including observed heterozygosity (Ho), expected heterozygosity (He) and nucleotide diversity (π) were computed using Arlequin (version 3.5). Analysis outputs provided quantitative metrics for assessing population-level genetic variation and divergence across subgroups.

2.5 Statistical analysis

All analytical results were statistically processed and visualized using R (version 4.2.1) and Python (version 3.9) scripts, encompassing data quality control, variant distribution profiling, and population structure characterization. Visualization workflows were implemented via R packages including ggplot2, pheatmap, and circlize, generating publication-ready figures in PDF and PNG formats. Code repositories and parameter settings were archived to ensure analytical accuracy and reproducibility of findings.

3 Results

3.1 RAD-seq data quality control and filtering

The distribution characteristics of base composition (A, T, C, G, N) in RAD-seq data serve as crucial indicators for assessing sequencing data quality. During library preparation and sequencing processes, factors such as PCR amplification bias may lead to A/T and G/C separation phenomena, potentially compromising data accuracy and reliability. According to sequencing principles and the principle of complementary base pairing, GC content and AT content should remain relatively stable across each sequencing cycle under ideal conditions, maintaining a consistent horizontal distribution trend throughout the sequencing process. Notably, the proportion of N bases (representing unidentifiable base types) serves as a key reference metric for evaluating sequencing quality.

The sequencing base composition distribution results for this project are shown in Supplementary Figure S1. Due to the connection with primer adapters at sequencing initiation sites, A, C, G, and T contents exhibited initial fluctuations at starting positions. However, these base compositions gradually stabilized as sequencing progressed. Particularly noteworthy is that the proportion of unknown bases (N) remained consistently low throughout the process. This observation indicates minimal systemic AT bias during sequencing and reflects that both library construction quality and sequencing performance met optimal standards, satisfying the requirements for subsequent bioinformatics analyses.

Statistical analysis of clean data from sequencing 185 coffee germplasm resources (Supplementary Table S2). A total of 216,178,439,423 raw base pairs were obtained. Each sample yielded an average of 4,700,228 clean paired-end (PE) reads. The average sequencing output was 4.7 Gb raw data (Q30 ≥ 94.25%). The mean Q20 value (base recognition accuracy ≥ 99%) was 97.89%, and the mean Q30 value (base recognition accuracy ≥ 99.9%) was 94.25%, with an average GC content of 42.16%. These results demonstrate high base-calling efficiency, excellent sequencing quality, and realistic GC distribution. The data quality meets the requirements for ddRAD-seq analysis and is suitable for downstream bioinformatics analyses.

3.2 Data comparison rate and coverage statistics

To further investigate the genomic characteristics of the 185 coffee germplasms, sequencing data were aligned to the coffee reference genome (ET-39 v2.4) using BWA software. The alignment rate reflects the similarity between the sample genome and the reference genome. If the reference genome is appropriately selected and no contamination occurred during experiments, the alignment rate of paired-end (PE) reads should exceed 70%. In this study, 88.77% of sequencing data from all samples aligned to the reference genome on average (Supplementary Table S3), indicating normal library construction and absence of contamination.

Genome coverage refers to the percentage of the reference genome that the read sequences cover, reflecting the completeness of variant detection. Coverage depth refers to the average number of reads covering each base, which affects the accuracy of variant detection. After mapping the read sequences to the reference genome, analysis was conducted on these two key indicators. The results showed that for 185 coffee germplasm, the average genome coverage was 3.05% and the average coverage depth was 17.66× (Supplementary Figure S2; Supplementary Table S4).

3.3 Variation type detection and distribution

SNP detection was performed on the 185 coffee germplasms using GATK software, yielding 37,729 variant sites, including 35,601 SNPs and 2,128 InDels (Table 1, Figure 1). Among the SNPs, 25,310 were transitions (A/G and C/T), and 10,291 were transversions (A/C, A/T, C/G, and G/T), with a transition-to-transversion ratio (Ts/Tv) of 2.46. The distribution of SNP types and length statistics of InDels are shown in the Figure 1. Significant differences were observed in the number and density of variants across chromosomes, correlating with chromosome length, gene distribution, and functional regions.

This study analyzed the genomic distribution of variants, revealing significant heterogeneity in variant counts and densities across chromosomes. For SNPs, CA1 exhibited the highest count (4,705), while CA8 had the lowest (392), with a genome-wide total of 35,601 SNPs (CA refers to chromosome number). The SNP density ranged from 79.70 SNPs per Mb on CA1 to 7.47 SNPs per Mb on CA20, averaging 30.02 SNPs per Mb genome-wide. For InDels, CA1 showed the highest number (269), whereas CA2 had fewer (64), with a total of 2,128 InDels genome-wide. InDel density varied from 4.56 InDels per Mb on CA1 to 0.45 InDels per Mb on CA20, with an average genome-wide density of 1.79 InDels per Mb. These findings confirm the non-uniform distribution of variants, with striking differences in SNP/InDel abundance and density among chromosomes. This variability provides critical insights for investigating genome structure, functional regions, and the biological implications of genetic variation.

3.4 Population structure analysis

Based on the STRUCTURE analysis of 185 coffee germplasm resources, the optimal number of subgroups was determined to be K=3 (corresponding to the highest ΔK value), indicating that the genetic structure of the tested coffee population is most distinct when divided into three groups (Figure 2). The ΔK trend and the distribution of individuals within subgroups are illustrated in the Figure 2. The population genetic analysis revealed that the 185 coffee accessions are best classified into three groups: Group 1 contains 31 coffee varieties, Group 2 includes 60 coffee varieties, and Group 3 comprises 94 coffee varieties.

PCA (Figure 3) of the 185 coffee germplasm resources was performed using PLINK based on SNP differences among individual genomes. The results, categorized into three genetic groups, aligned with the earlier population structure analysis, further supporting the hypothesis of substantial genetic diversity among the 185 coffee accessions.

A phylogenetic tree was constructed using the NJ method based on genetic distances derived from SNP markers (Figure 4). Samples were labeled according to their prior group classifications. Intriguingly, coffee varieties from different groups were observed within the same clades. To validate these findings, the phylogenetic tree was integrated with population structure analyses, incorporating scenarios for K=2, 3, 4, and 5. The combined results demonstrated that K=3 remained the optimal grouping despite minor overlaps of certain varieties in shared clades, consistent with the population structure conclusions. This further reinforces the hypothesis of pronounced genetic divergence among the 185 accessions. The phylogenetic tree revealed that the 185 germplasm samples spanned all 11 major evolutionary clades of the Coffea genus. The genetic diversity parameter $\pi = 0.1456$ confirms the representativeness of the 185 coffee variety samples.

3.5 Genetic diversity analysis and genetic differentiation analysis

Genetic diversity, defined as the genetic variation among different populations or individuals within a species, serves as the foundation for survival, adaptation, and evolution. In this study, expected heterozygosity (He), polymorphism information content (PIC), and nucleotide diversity (π) were used to assess the genetic diversity of each subgroup. Genetic diversity metrics were calculated using VCFtools with a window size of 100 kb and a step size of 20 kb.

The results show G2 exhibited the highest values across all three metrics, indicating relatively greater genetic diversity compared to G1 and G3 (Table 2). In contrast, G3 showed the lowest nucleotide diversity (π). We hypothesize that the elevated diversity in G2 may reflect its inclusion of more wild genetic resources or divergent selection pressures compared to the other groups.

TABLE 1 Types and density of variation.

Chr	Length	No. SNPs	SNP density	No. InDels	InDel density
CA1	59,034,463	4,705	79.70	269	4.56
CA2	57,716,770	734	12.72	64	1.11
CA3	76,906,396	1,708	22.21	111	1.44
CA4	77,301,657	1,045	13.52	73	0.94
CA5	44,587,295	2,058	46.16	153	3.43
CA6	49,228,674	738	14.99	32	0.65
CA7	52,323,663	4,063	77.65	226	4.32
CA8	48,754,064	392	8.04	17	0.35
CA9	49,172,495	1,577	32.07	101	2.05
CA10	55,030,746	543	9.87	28	0.51
CA11	63,716,479	3,389	53.19	168	2.64
CA12	63,214,365	679	10.74	51	0.81
CA13	40,515,400	3,157	77.92	215	5.31
CA14	50,646,195	465	9.18	39	0.77
CA15	46,623,164	1,495	32.07	109	2.34
CA16	53,737,150	874	16.26	41	0.76
CA17	44,254,144	3,175	71.74	163	3.68
CA18	41,449,412	595	14.35	38	0.92
CA19	54,177,763	753	13.90	46	0.85
CA20	48,872,521	365	7.47	22	0.45
CA21	45,802,622	2,497	54.52	130	2.84
CA22	62,661,540	594	9.48	32	0.51
Whole	1,185,726,978	35,601	30.02	2,128	1.79

Chr, Chromosome number; Length, chromosome length; No. SNPs, indicates the number of SNPS; SNP density, SNP density (number of SNPS per Mb); No. InDels, The number of InDel; InDel density, Density of InDel (number of InDel per Mb).

Figure 5 illustrates the genome-wide distribution of π values across the three groups, providing a visual overview of their genetic diversity patterns. These findings offer robust data-driven insights into the genetic diversity of the germplasm resources, supporting their conservation and utilization.

4 Discussion

4.1 Analysis based on RAD-Seq confirmed the genetic diversity of Chinese coffee germplasm resources

Molecular markers (such as RFLP, SSR, SNP, etc.) are widely used in population genetics, and the advent of next-generation sequencing technology has significantly enhanced their development and application efficiency (Grover and Sharma, 2016; Huang et al., 2009). The application of molecular markers in coffee aligns with this trend and effectively reveals the genetic differentiation within coffee. Anthony et al. revealed the genetic diversity of Arabica coffee through AFLP and SSR markers, further supporting the importance of wild germplasm resources in coffee breeding (Anthony et al., 2002). Yunita et al. conducted a genetic diversity analysis of Arabica coffee in the Solok Regency region using the SRAP molecular marker technique (Yunita et al., 2020). The results showed that the He among the samples ranged from 0.2812 to 0.3638, indicating that the SRAP markers could effectively reveal the genetic variation characteristics of Arabica coffee in this region. Huang Lifang and others conducted a genetic diversity analysis on 87 coffee germplasm resources using RAPD primers (Huang et al., 2017). This provided important molecular basis for the evaluation and breeding of coffee germplasm resources. However, it should be noted that this study had a limitation of a small sample size, which might have affected the representativeness and statistical power of the research results to some extent. Based on the above research background, this study employed high-density SNP marker technology to conduct a systematic genetic diversity analysis on 185 coffee germplasm resources from China. The research results showed that the genetic diversity index (He) of Chinese coffee germplasm resources was 0.3014, which was similar

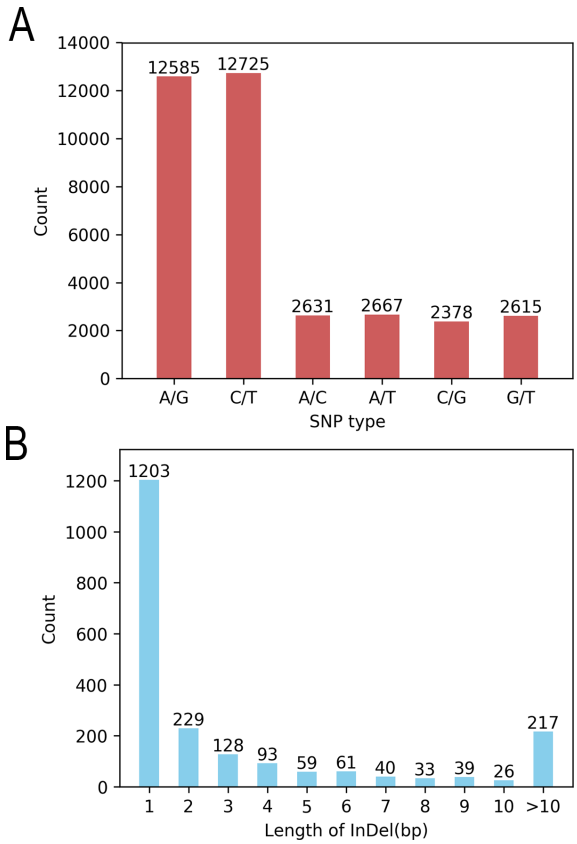

FIGURE 1
Statistics of variation types. **(A)** SNP conversion/transposition types and **(B)** InDel length distribution.

to the results (H = 0.2812 - 0.3638) of Yunita et al.'s study on Arabica coffee in the Solok Regency area (Yunita et al., 2020). It is worth noting that both studies classified the test materials into three genetic groups. This finding further verified that Chinese coffee germplasm resources have moderate levels of genetic diversity. Moreover, the results of this study also confirmed that RAD-Seq technology, as an efficient SNP genotyping method for the entire genome, has comparable application value in crop genetic diversity analysis to SRAP markers.

4.2 Classification of genetic populations of 185 coffee varieties

The population structure of small-grain coffee (*Coffea arabica*) can be classified into three main groups based on its botanical characteristics and genetic evolution relationships: the Bourbon/Typica cultivar group, the Ethiopian Native original species group, and the Introgression Group hybrid population (Salojärvi et al., 2024). This classification system not only reflects the evolutionary history of coffee populations, but also demonstrates the significant differences in agronomic traits and flavor characteristics among different groups. In this study, the results of molecular marker analysis indicated that 185 coffee germplasm resources could be clearly classified into three genetic groups (Figures 2–4). The results of population structure analysis, principal component analysis, and phylogenetic tree construction were all consistent in supporting this genetic grouping pattern, thereby verifying the classification hypotheses made in the previous stage of the research. These findings align with previous research, underscoring the rich genetic variation preserved in coffee germplasm resources worldwide (Aerts et al., 2013). This diversity provides a critical genetic foundation for coffee breeding programs and germplasm conservation efforts, enabling the identification of valuable traits for crop improvement and resilience against environmental challenges (Yan et al., 2019; Aerts et al., 2013). Genetic population analysis indicates that the G1 population may mainly consist of cultivated varieties from specific geographical regions, while the G2 population is significantly enriched with wild germplasm resources. However, there is a clear phenomenon of gene exchange between the populations, manifested by the migration of some germplasms between the G1 and G2 populations. This phenomenon of genetic component mixture may result from the following factors: (1) gene infiltration caused by artificial hybridization breeding (Ogutu, 2019). (2) adaptive evolution under natural selection pressure (Wan and Wootton, 2000). (3) selective elimination effects in specific genomic regions (Huang et al., 2017). In particular, individuals at the population boundary may carry special recombinant haplotypes, and these variations may be the key factors contributing to gene flow between the

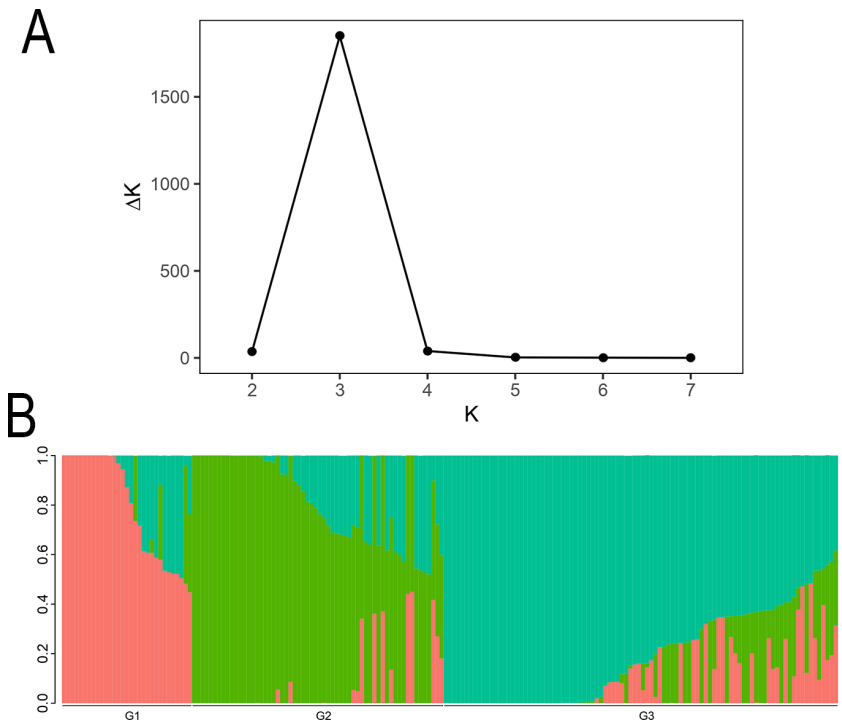

FIGURE 2
The Structure analysis of 185 coffee germplasm resources. **(A)** The change trend chart of ΔK. **(B)** Histogram of Q value of each sample when K=3. In the Structure diagram, each color represents a subgroup. A stacked column chart in the diagram represents an individual's "ancestry", and an individual with only one color indicates a relatively pure lineage, while those with multiple colors indicate a mixed lineage. Through the colors, we can divide the individuals in the population into different subgroups.

populations. For example, Sample number 154. This material was formed through the hybridization of CATURI and HDT, and it belongs to the gene infiltration group. However, since its parent strain HDT belongs to the native species of Ethiopia, it was classified into Group 1. For instance, the population differentiation phenomenon discovered by Iqbal et al., as well as the geographical isolation effect revealed by Pagani et al., all corroborate the conclusions of this study (Iqbal et al., 2022; Pagani et al., 2012).

4.3 Significance and optimization of core seed bank

A core germplasm collection is a critical strategy for efficient management and utilization of genetic resources, aiming to condense large-scale germplasm into a representative, genetically diverse subset through scientific screening and optimization (Yang et al., 2022). This collection serves as a

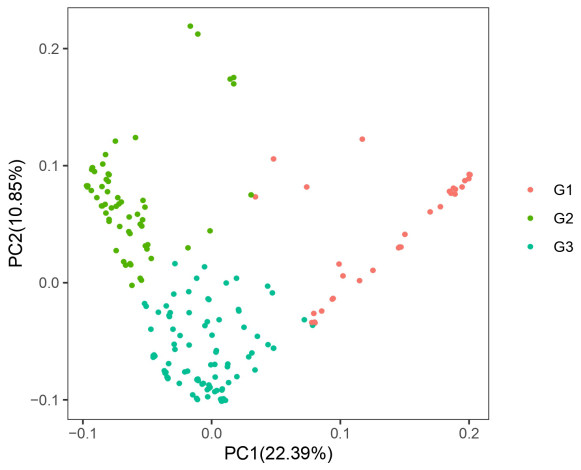

FIGURE 3
Scatter plot of PCA of 185 coffee germplasm resources.

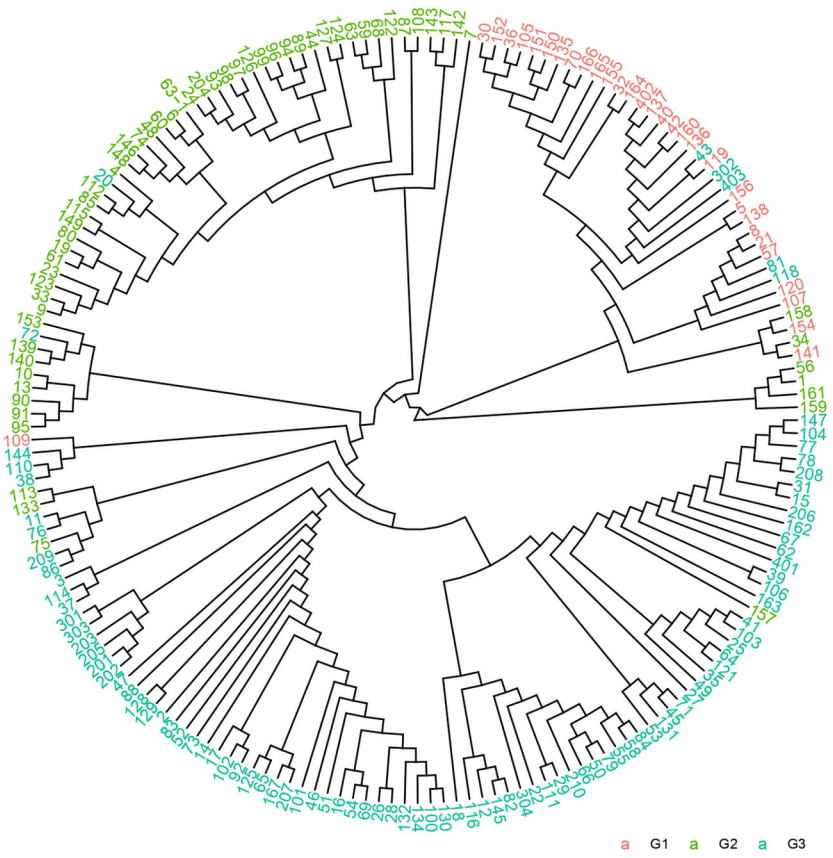

FIGURE 4
Phylogenetic tree of 185 coffee germplasm resources.

high-efficiency platform for coffee breeding, enabling rapid identification of germplasm with desirable traits and shortening breeding cycles. Through the establishment of a core germplasm resource library for rice, they significantly improved the breeding efficiency, providing an important model for the field of crop genetic improvement (Zhang et al., 2011). Additionally, Liu et al. demonstrated the utility of core collections in resource management by integrating genetic and metabolic data to build a medicinal plant core collection (Liu et al., 2020). The marker density of this study (37,729 SNPs) is significantly higher than that of the reported rice core germplasm resources (12,000 SNPs), the existing coffee SNP chip technology (8,500 SNPs), and the cocoa core germplasm resources (30,000 SNPs) (Zhang et al., 2011; Merot-L'anthoene et al., 2019; Motamayor et al., 2013).

TABLE 2 Statistical table of genetic diversity indicators.

Group	He	PIC	π
G1	0.1432	0.1205	0.1326
G2	0.3014	0.2428	0.1456
G3	0.1633	0.1425	0.1031

He, expected heterozygosity; PIC, polymorphism information content; π, nucleotide diversity.

While the Coffee Core Germplasm Collection (CCGC) already encompasses global genetic lineages and functional genes, future efforts could enhance its utility by integrating phenotypic data (e.g., disease resistance, flavor metabolites) to improve genetic diversity and representativeness. Second, combining phenotypic and genomic data would enable more precise screening to ensure inclusion of germplasm with key agronomic traits. Furthermore, the digital management of germplasm resources (such as establishing a coffee germplasm resource database) will significantly enhance the efficiency of researchers and breeders in accessing and utilizing relevant resources. For instance, Ndjiondjop et al. effectively optimized the utilization strategy of germplasm resources by establishing a small core germplasm bank for rice. Their method can provide important references for the screening and integration of coffee core germplasm (Ndjiondjop et al., 2017).

4.4 Limitations and prospects of the study

Although this study has made significant progress, it still has certain limitations. In terms of sample selection, although the

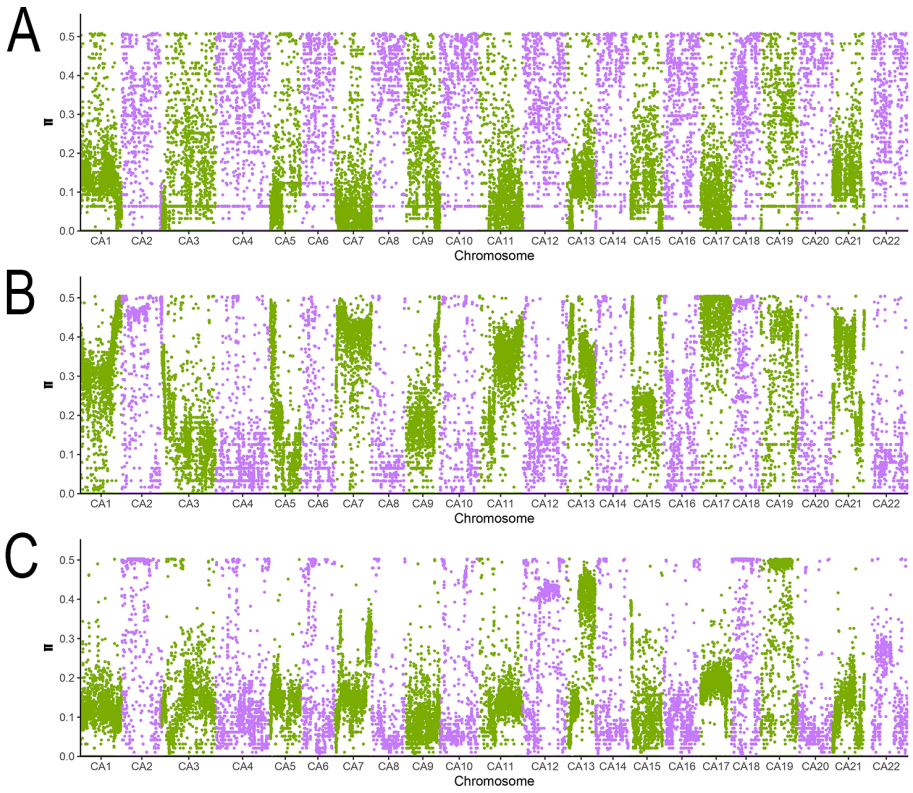

FIGURE 5
Distribution of whole-genome π values of the three subpopulations. This figure presents the distribution of nucleotide diversity (π value) across 22 chromosomes (CA1-CA22) in the three groups of samples through STRUCTURE analysis. ABC represent Group 1, Group 2, and Group 3 respectively. The purple and green data points are only used to distinguish adjacent chromosomes (to avoid visual confusion) and have no biological significance. CA1-CA22 represent chromosome numbers, and the vertical coordinate π value range (0.0 - 0.5) reflects the level of genetic diversity.

sample size is large, it mainly focuses on the two main cultivated varieties of coffee, and the coverage of wild varieties is insufficient. Future research needs to include more wild resources to comprehensively analyze the genetic diversity of coffee. In terms of research methods, although RAD-seq has cost-effectiveness advantages, its detection range is limited compared to whole-genome sequencing (WGS), and it may miss important genetic variations. It is recommended that subsequent research combine WGS technology, such as the whole-genome resequencing strategy adopted by Mekbib et al., to construct a more complete genetic map. This study provides an important reference framework for the analysis of coffee genetic diversity (Mekbib et al., 2022).

Future research directions may include: Functional exploration of disease/resistance genes and identification of genetic markers linked to key agronomic traits to expand candidate gene pools for molecular breeding. GWAS integrating phenotypic and genomic data to uncover loci associated with yield, quality, and disease resistance. Expanding the core germplasm collection by adding representative accessions to enhance its genetic diversity and utility.

5 Conclusions

This study systematically analyzed and validated the genetic diversity of 185 coffee germplasm resources using RAD-seq technology. The research results show that this sample set comprehensively covers all 11 major evolutionary lineages of the Coffee genus. Its genetic diversity parameters are $\pi = 0.1456$ and $He = 0.3014$. Population structure analysis (K=3) further confirmed its genetic representativeness. The marker density (37,729 SNPs) represents a 7.5-fold improvement over existing coffee core collections (5K SNP array) and surpasses comparable studies on crops such as rice (12K SNPs) and cacao (30K SNPs).

As the first genome-wide validation of genetic diversity in a coffee core germplasm collection, this research addresses a critical knowledge gap and provides a scientific foundation for precise classification, efficient conservation, and molecular breeding of coffee genetic resources. The establishment of this core collection not only creates a standardized platform for rapid screening of disease-resistant, stress-tolerant, and high-quality germplasm, but also serves as a crucial genetic reservoir for addressing climate change and pest/disease challenges.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1309331/.

Author contributions

XJ: Software, Validation, Writing – original draft. CL: Data curation, Writing – original draft. GM: Formal Analysis, Writing – review & editing. MZ: Investigation, Writing – original draft. ML: Writing – review & editing. TC: Data curation, Writing – original draft. PZ: Investigation, Writing – original draft. JMW: Project administration, Writing – original draft. QL: Project administration, Writing – original draft. TG: Supervision, Writing – review & editing. LS: Funding acquisition, Writing – original draft. ZZ: Funding acquisition, Writing – original draft. JYW: Data curation, Writing – original draft. ZX: Visualization, Writing – original draft. BX: Supervision, Writing – review & editing. HZ: Conceptualization, Writing – review & editing. JL: Resources, Writing – review & editing. XB: Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This research was funded by National Cassava Industry Technology System Coffee Yunnan Dehong Comprehensive Test Station (CARS-11-YNLJH), National Tropical Plant Germplasm Re-source Bank— Coffee Germplasm Resource Bank (NTPGRC2024-016), Yunnan Fundamental Research Projects (202401AT070027), Yunnan Province Coffee Key Laboratory (Yunnan Agricultural University) (202449CE340030), Science and Technology Talent and Platform Program (Academician and Expert Workstation) (202405AF140061).

Acknowledgments

We would like to thank Bo Wang from Genoseq Technology Co., Ltd. (Wuhan, China) for his thorough suggestions for the experiment design and technical support.

Conflict of interest

The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1629553/full#supplementary-material

References

Aerts, R., Berecha, G., Gijbels, P., Hundera, K., Glabeke, S., Vandepitte, K., et al. (2013). Genetic variation and risks of introgression in the wild coffea arabica gene pool in South-Western Ethiopian montane rainforests. *Evol. Appl.* 6, 243–252. doi: 10.1111/j.1752-4571.2012.00285.x

Anthony, F., Bertrand, B., Quiros, O., Wilches, A., Lashermes, P., Berthaud, J., et al. (2001). Genetic diversity of wild coffee (Coffea arabica L.) using molecular markers. *Euphytica* 118, 53–65. doi: 10.1023/A:1004013815166

Anthony, F., Combes, C., Astorga, C., Bertrand, B., Graziosi, G., and Lashermes, P. (2002). The origin of cultivated coffea arabica L. Varieties revealed by aflp and ssr markers. *Theor. Appl. Genet.* 104, 894–900. doi: 10.1007/s00122-001-0798-8

Arai-Kichise, Y., Shiwa, Y., Nagasaki, H., Ebana, K., Yoshikawa, H., Yano, M., et al. (2011). Discovery of genome-wide dna polymorphisms in A landrace cultivar of japonica rice by whole-genome sequencing. *Plant Cell Physiol.* 52, 274–282. doi: 10.1093/pcp/pcr003

Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., et al. (2015). The coffee rust crises in Colombia and central america, (2008–2013: impacts, plausible causes and proposed solutions. *Food Secur.* 7, 303–321. doi: 10.1007/s12571-015-0446-9

Bertrand, B., Etienne, H., Cilas, C., Charrier, A., and Baradat, P. (2005). Coffea arabica hybrid performance for yield, fertility and bean weight. *Euphytica* 141, 255–262. doi: 10.1007/s10681-005-7681-7

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J. C., et al. (2002). Insertion-deletion polymorphisms in 3' Regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Mol. Biol.* 48, 539–547. doi: 10.1023/A:1014841612043

Bunn, C., Läderach, P., Ovalle Rivera, O., and Kirschke, D. (2015). A bitter cup: climate change profile of global production of arabica and robusta coffee. *Climatic Change* 129, 89–101. doi: 10.1007/s10584-014-1306-x

Ceja-Navarro, J. A., Vega, F. E., Karaoz, U., Hao, Z., Jenkins, S., Lim, H. C., et al. (2015). Gut microbiota mediate caffeine detoxification in the primary insect pest of coffee. *Nat. Commun.* 6, 7618. doi: 10.1038/ncomms8618

Cenci, A., Combes, M. C., and Lashermes, P. (2012). Genome evolution in diploid and tetraploid coffea species as revealed by comparative analysis of orthologous genome segments. *Plant Mol. Biol.* 78, 135–145. doi: 10.1007/s11103-011-9852-3

Clarindo, W. R., and Carvalho, C. R. (2008). First coffea arabica karyogram showing that this species is A true allotetraploid. *Plant Syst. Evol.* 274, 237–241. doi: 10.1007/s00606-008-0050-y

Clifford, M. N., and Willson, K. C. (1985). *Coffee: Botany, Biochemistry And Production Of Beans And Beverage* (Coffee: Botany, Biochemistry And Production Of Beans And Beverage).

Cunha Alves, A. A., and Azevedo, V. C. R. (2018). Embrapa network for Brazilian plant genetic resources conservation. *Biopreserv. Biobank* 16, 350–360. doi: 10.1089/bio.2018.0044

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Davis, A. P., Gole, T. W., Baena, S., and Moat, J. (2012). The impact of climate change on indigenous arabica coffee (Coffea arabica): predicting future trends and identifying priorities. *PloS One* 7, E47981. doi: 10.1371/journal.pone.0047981

Davis, A. P., Mieulet, D., Moat, J., Sarmu, D., and Haggar, J. (2021). Arabica-like flavour in a heat-tolerant wild coffee species. *Nat. Plants* 7, 413–418. doi: 10.1038/s41477-021-00891-4

Davis, A. P., Tosh, J., Ruch, N., and Fay, M. F. (2011). Growing coffee: psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of coffea. *Bot. J. Linn. Soc.* 167, 357–377. doi: 10.1111/j.1095-8339.2011.01177.x

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (Gbs) approach for high diversity species. *PloS One* 6, E19379. doi: 10.1371/journal.pone.0019379

Errington, F., Fujikura, T., and Gewertz, D. (2012). Instant noodles as an antifriction device: making the bop with ppp in png. *Am. Anthropol.* 114, 19–31. doi: 10.1111/j.1548-1433.2011.01394.x

Gautam, P. L., Singh, B. B., Saxena, S., and Sharma, R. K. (2004). "Collection, conservation and utilization of plant genetic resources," in *Plant Breeding: Mendelian To Molecular Approaches*. Eds. H. K. Jain and M. C. Kharkwal (Dordrecht, Springer Netherlands).

Grover, A., and Sharma, P. C. (2016). Development and use of molecular markers: past and present. *Crit. Rev. Biotechnol.* 36, 290–302. doi: 10.3109/07388551.2014.959891

Guo, Z., Yang, Q., Huang, F., Zheng, H., Sang, Z., Xu, Y., et al. (2021). Development of high-resolution multiple-snp arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. *Plant Commun.* 2, 100230. doi: 10.1016/j.xplc.2021.100230

Huang, L., Dong, Y., Wang, X., Chen, P., Lin, X., Fan, R., et al. (2014). Analysis of genetic diversity of coffee germplasm resources using rapd markers. *Trop. J. Crop Sci.* 35, 2313–2319. doi: 10.3969/j.issn.1000-2561.2014.12.001

Huang, L., Dong, Y., Wang, X., Sun, Y., Chen, P., Lin, X., et al. (2016). Construction of dna fingerprint maps for coffee germplasm resources. *Trop. Agric. Sci.* 36, 37–42. doi: 10.12008/j.issn.1009-2196.2016.12.008

Huang, L., Dong, Y., Wang, X., Sun, Y., Chen, P., Lin, X., et al. (2017). Rapd analysis of genetic diversity of coffee resources in Yunnan. *Chin. Trop. Agric.* 5, 48–52. doi: 10.3969/j.issn.1673-0658.2017.05.012

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., et al. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19, 1068–1076. doi: 10.1101/gr.089516.108

Iqbal, A., Huiping, G., Xiangru, W., Hengheng, Z., Xiling, Z., and Meizhen, S. (2022). Genome-wide expression analysis reveals involvement of asparagine synthetase family in cotton development and nitrogen metabolism. *BMC Plant Biol.* 22, 122. doi: 10.1186/s12870-022-03454-7

Jangra, S., Chaudhary, V., Yadav, R. C., and Yadav, N. R. (2021). High-throughput phenotyping: A platform to accelerate crop improvement. *Phenomics* 1, 31–53. doi: 10.1007/s43657-020-00007-6

Krishnan, S. (2022). "Coffee: genetic diversity, erosion, conservation, and utilization," in *Cash Crops: Genetic Diversity, Erosion, Conservation And Utilization*. Eds. P. M. Priyadarshan and S. M. Jain (Springer International Publishing, Cham).

Lashermes, P., Combes, M. C., Robert, J., Trouslot, P., D'hont, A., Anthony, F., et al. (1999). Molecular characterisation and origin of the coffea arabica L. Genome. *Mol. Gen. Genet.* 261, 259–266. doi: 10.1007/s004380050965

Li, Z., Yuan, X., Chen, Y., Zheng, X., Hu, Z., and Li, L. (2020). Efficient identification of rice germplasm resources based on whole-genome snps and construction of fingerprint maps. *Mol. Plant Breed.* 18, 6050–6057. doi: 10.13271/j.mpb.018.006050

Liu, D., Sun, Y. Y., Wei, C. Q., Xie, Z., Li, H. L., Zhang, W. W., et al. (2017). InDel molecular markers and their applications in rice research. *Seeds* 36, 47–52. doi: 10.16590/j.cnki.1001-4705.2017.09.047

Liu, M., Hu, X., Wang, X., Zhang, J., Peng, X., Hu, Z., et al. (2020). Constructing A core collection of the medicinal plant angelica biserrata using genetic and metabolic data. *Front. Plant Sci.* 11, 600249. doi: 10.3389/fpls.2020.600249

Mekbib, Y., Tesfaye, K., Dong, X., Saina, J. K., Hu, G. W., and Wang, Q. F. (2022). Whole-genome resequencing of coffea arabica L. (Rubiaceae) genotypes identify snp and unravels distinct groups showing A strong geographical pattern. *BMC Plant Biol.* 22, 69. doi: 10.1186/s12870-022-03449-4

Merot-L'anthoene, V., Tournebize, R., Darracq, O., Rattina, V., Lepelley, M., Bellanger, L., et al. (2019). Development and evaluation of A genome-wide coffee 8.5k snp array and its application for high-density genetic mapping and for investigating the origin of coffea arabica L. *Plant Biotechnol. J.* 17, 1418–1430. doi: 10.1111/pbi.13066

Moat, J., Williams, J., Baena, S., Wilkinson, T., Gole, T. W., Challa, Z. K., et al. (2017). Resilience potential of the Ethiopian coffee sector under climate change. *Nat. Plants* 3, 17081. doi: 10.1038/nplants.2017.81

Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D. 3rd, Cornejo, O., et al. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* 14, R53. doi: 10.1186/gb-2013-14-6-r53

Ndjiondjop, M. N., Semagn, K., Gouda, A. C., Kpeki, S. B., Dro Tia, D., Sow, M., et al. (2017). Genetic variation and population structure of oryza glaberrima and development of A mini-core collection using dartseq. *Front. Plant Sci.* 8, 1748. doi: 10.3389/fpls.2017.01748

Ogutu, C. O. (2019). *Genome-wide SSR Evaluation of Carinifera Coffee and Analysis of Genetic Diversity among Coffee Species Based on SSR and ISSR Molecular Markers (Doctoral Dissertation, University of Chinese Academy of Sciences (Chinese Academy of Sciences Wuhan Botanical Garden))*. Doctoral. doi: 10.27603/d.cnki.gkhzs.2019.000014

Otyama, P. I., Wilkey, A., Kulkarni, R., Assefa, T., Chu, Y., Clevenger, J., et al. (2019). Evaluation of linkage disequilibrium, population structure, and genetic diversity in the U.S. Peanut mini core collection. *BMC Genomics* 20, 481. doi: 10.1186/s12864-019-5824-9

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* 91, 83–96. doi: 10.1016/j.ajhg.2012.05.015

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest radseq: an inexpensive method for *de novo* snp discovery and genotyping in model and non-model species. *PloS One* 7, E37135. doi: 10.1371/journal.pone.0037135

Pham, Y., Reardon-Smith, K., Mushtaq, S., and Cockfield, G. (2019). The impact of climate change and variability on coffee production: A systematic review. *Climatic Change* 156, 609–630. doi: 10.1007/s10584-019-02538-y

Prakash, N. S., Combes, M.-C., Dussert, S., Naveen, S., and Lashermes, P. (2005). Analysis of genetic diversity in Indian robusta coffee genepool (Coffea canephora) in comparison with A representative core collection using ssrs and aflps. *Genet. Resour. Crop Evol.* 52, 333–343. doi: 10.1007/s10722-003-2125-5

Salathia, N., Lee, H. N., Sangster, T. A., Morneau, K., Landry, C. R., Schellenberg, K., et al. (2007). Indel arrays: an affordable alternative for genotyping. *Plant J.* 51, 727–737. doi: 10.1111/j.1365-313X.2007.03194.x

Salojärvi, J., Rambani, A., Yu, Z., Guyot, R., Strickler, S., Lepelley, M., et al. (2024). The genome and population genomics of allopolyploid coffea arabica reveal the diversification history of modern coffee cultivars. *Nat. Genet.* 56, 721–731. doi: 10.1038/s41588-024-01695-w

Scalabrin, S., Toniutti, L., Di Gaspero, G., Scaglione, D., Magris, G., Vidotto, M., et al. (2020). A single polyploidization event at the origin of the tetraploid genome of coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Sci. Rep.* 10, 4642. doi: 10.1038/s41598-020-61216-7

Van Der Vossen, H. A. M. (1985). "Coffee selection and breeding," in *Coffee: Botany, Biochemistry And Production Of Beans And Beverage*. Eds. M. N. Clifford and K. C. Willson (Springer Us, Boston, Ma).

Vossen, H. A. M. V. D. (1985). Coffee Selection And Breeding. In: *Coffee*. Clifford, M. N., Willson, K. C. (eds). (Boston, MA: Springer). doi: 10.1007/978-1-4615-6657-1_3

Wan, H., and Wootton, J. C. (2000). A global compositional complexity measure for biological sequences: at-rich and gc-rich genomes encode less complex proteins. *Comput. Chem.* 24, 71–94. doi: 10.1016/S0097-8485(00)80008-X

Yan, L., Ogutu, C., Huang, L., Wang, X., Zhou, H., Lv, Y., et al. (2019). Genetic diversity and population structure of coffee germplasm collections in China revealed by issr markers. *Plant Mol. Biol. Rep.* 37, 204–213. doi: 10.1007/s11105-019-01148-3

Yang, Y., Lyu, M., Liu, J., Wu, J., Wang, Q., Xie, T., et al. (2022). Construction of an snp fingerprinting database and population genetic analysis of 329 cauliflower cultivars. *BMC Plant Biol.* 22, 522. doi: 10.1186/s12870-022-03920-2

Yunita, R., Oktavioni, M., Chaniago, I., Syukriani, L., Setiawan, M. A., and Jamsari, J. (2020). Analysis of genetic diversity of arabica coffee [Coffea arabica L.] in solok regency by srap molecular markers. *Iop Conf. Series: Earth Environ. Sci.* 497, 012018. doi: 10.1088/1755-1315/497/1/012018

Zeru, A., Assefa, F., Adugna, G., and Hindorf, H. (2012). Occurrence of fungal diseases of coffea arabica L. In montane rainforests of Ethiopia. *J. Appl. Bot. Food Qual.* 82, 148–151. doi: 10.5073/JABFQ.2008.082.024

Zhang, H., Zhang, D., Wang, M., Sun, J., Qi, Y., Li, J., et al. (2011). A core collection and mini core collection of oryza sativa L. In China. *Theor. Appl. Genet.* 122, 49–61. doi: 10.1007/s00122-010-1421-7

Zhao, M., Li, J., Bai, X., Ma, G., Guo, T., Xiao, Z., et al. (2025). Current status and innovative utilization of coffee resources in the ruili coffee germplasm resource. *Trop. Agric. Sci.* 48, 46–52. doi: 10.16005/j.cnki.tast.2025.01.009

Zhou, H., Zhang, H., Xia, H., Yang, J., Guo, T., Li, J., et al. (2015). Research on the diversity of coffee germplasm resources. *Chin. Trop. Agric.*, 23–27. doi: 10.3969/j.issn.1673-0658.2015.05.007