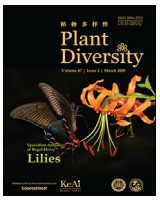
Plant Diversity 47 (2025) 214–228

Contents lists available at ScienceDirect

Plant Diversity

journal homepage: http://www.keaipublishing.com/en/journals/plant-diversity/
http://journal.kib.ac.cn



Research paper

Single-nucleotide polymorphisms and copy number variations drive adaptive evolution to freezing stress in a subtropical evergreen broad-leaved tree: Hexaploid wild *Camellia oleifera*

Haoxing Xie [a], Kaifeng Xing [a], Jun Zhou [a], Yao Zhao [a, b], Jian Zhang [a], Jun Rong [a, b, *]

[a] *Key Laboratory of Poyang Lake Environment and Resource Utilization, Ministry of Education, Center for Watershed Ecology, School of Life Sciences, Nanchang University, Nanchang 330031, China*
[b] *Lushan Botanical Garden, Chinese Academy of Sciences, Lushan 332999, China*

ARTICLE INFO

Article history:
Received 24 April 2024
Received in revised form
21 July 2024
Accepted 24 July 2024
Available online 27 July 2024

Keywords:
Adaptive evolution
Camellia oleifera
Copy number variations
Freezing stress
Polyploid
Single-nucleotide polymorphisms

ABSTRACT

Subtropical evergreen broad-leaved trees are usually vulnerable to freezing stress, while hexaploid wild *Camellia oleifera* shows strong freezing tolerance. As a valuable genetic resource of woody oil crop *C. oleifera*, wild *C. oleifera* can serve as a case for studying the molecular bases of adaptive evolution to freezing stress. Here, 47 wild *C. oleifera* from 11 natural distribution sites in China and 4 relative species of *C. oleifera* were selected for genome sequencing. "Min Temperature of Coldest Month" (BIO6) had the highest comprehensive contribution to wild *C. oleifera* distribution. The population genetic structure of wild *C. oleifera* could be divided into two groups: in cold winter (BIO6 $\leq$ 0 °C) and warm winter (BIO6 > 0 °C) areas. Wild *C. oleifera* in cold winter areas might have experienced stronger selection pressures and population bottlenecks with lower N_e than those in warm winter areas. 155 single-nucleotide polymorphisms (SNPs) were significantly correlated with the key bioclimatic variables (106 SNPs significantly correlated with BIO6). Twenty key SNPs and 15 key copy number variation regions (CNVRs) were found with genotype differentiation > 50% between the two groups of wild *C. oleifera*. Key SNPs in cis-regulatory elements might affect the expression of key genes associated with freezing tolerance, and they were also found within a CNVR suggesting interactions between them. Some key CNVRs in the exon regions were closely related to the differentially expressed genes under freezing stress. The findings suggest that rich SNPs and CNVRs in polyploid trees may contribute to the adaptive evolution to freezing stress.

1. Introduction

As a major abiotic stress, cold stress includes chilling stress (0–15 °C) and freezing stress (< 0 °C), which poses a significant threat to plants due to their sessile nature, limiting the distribution of plants, disrupting their growth and development, and reducing crop productivity (Liu et al., 2018; Zhang et al., 2022). Perennial woody plants overwinter with prolonged exposure to cold stress and some can survive under much lower freezing temperatures than annual herbaceous plants (Strimbeck et al., 2015). However, the molecular mechanisms of freezing tolerance in woody plants are largely unresolved compared to those in herbaceous plants (Wisniewski et al., 2014). On the other hand, subtropical evergreen broad-leaved trees are often sensitive to cold stress, especially for freezing stress, which may cause more serious damage or even death compared with chilling stress (Aslam et al., 2022). As an exception, wild *Camellia oleifera* of the genus *Camellia* (Theaceae) can exhibit strong freezing tolerance.

Wild *Camellia oleifera* is widely distributed in the subtropical low mountains and hilly areas in China (Cui et al., 2016). Wild *C. oleifera* in the Lu Mountain is in the northern distribution range of wild *C. oleifera* and could survive under severe freezing stress (below −10 °C) in winter (Xie et al., 2023). Xie et al. (2023) integrated field and lab experiments to reveal that wild *C. oleifera* in the

* Corresponding author. Key Laboratory of Poyang Lake Environment and Resource Utilization, Ministry of Education, Center for Watershed Ecology, School of Life Sciences, Nanchang University, Nanchang 330031, China.
E-mail address: rong_jun@hotmail.com (J. Rong).
Peer review under responsibility of Editorial Office of Plant Diversity.

https://doi.org/10.1016/j.pld.2024.07.009

Lu Mountain had strong freezing tolerance and some genes associated with the responses to freezing stress were identified with transcriptome analyses. Using microsatellite markers, Cui et al. (2022) uncovered rich genetic diversity and significant differentiation among wild *C. oleifera* populations across different latitudes and longitudes. The wild *C. oleifera* population in the Lu Mountain was genetically distinct which might be due to adaption isolation by cold climate conditions together with geographical isolation (Cui et al., 2022). In the genus *Camellia*, polyploids such as tetraploids and hexaploids are common, especially in the section *Paracamellia* (Ming, 2000). As the type species of the section *Paracamellia*, *C. oleifera* is hexaploid and may originate from inter-specific hybridization followed by whole genome duplication during the Quaternary under extreme cold stress (Qin et al., 2023). Owing to the complex hybrid origin history associated with harsh environmental conditions, allopolyploid plants usually evolve strong environmental adaptability (Heslop-Harrison et al., 2022). Therefore, the hexaploid wild *C. oleifera* can serve as a special case for studying the molecular bases of adaptive evolution to freezing stress in subtropical evergreen broad-leaved trees.

Moreover, cultivated *Camellia oleifera* is one of the four major woody oil crops in the world, together with *Elaeis guineensis*, *Olea europaea* and *Cocos nucifera* (Gao et al., 2024). Camellia oil contains high oleic acid content, making up over 80% of the fatty acid composition, known as “oriental olive oil”. It also contains many biological active components such as squalene, vitamin E, saponin and so on (Gao et al., 2024). Camellia oil is the premium edible vegetable oil recommended by the Food and Agriculture Organization of the United Nations (FAO) (Ke, 2019). Currently, cultivated *C. oleifera* is the dominant woody oil crop in southern China. To the end of 2022, the planting areas of *C. oleifera* reached about 4.67 million hectares with the annual oil production of about 1 million tons. Wild *C. oleifera* is a valuable genetic resource for cultivated *C. oleifera* breeding, especially for the tolerance to abiotic stresses.

The classic method for detecting the genetic basis of adaptation is mainly based on population genetics with limited molecular markers (Wright and Gaut, 2004). With the development of sequencing technologies, obtaining huge amounts of high-quality sequencing data at a reasonable cost has become more and more easier (Rellstab et al., 2015). This creates an opportunity for landscape genomic methods, which combine information on phenotype, genotype and local environment in many individuals or populations to study the impact of environmental factors on genetic variation pattern (Sork and Waits, 2010; Sork et al., 2013). Trees have long life history and complex genome characteristics, which are conducive to the application of landscape genomic methods, providing comprehensive insights into the molecular bases of environmental adaptation (Sork et al., 2013). Single-nucleotide polymorphisms (SNPs) are frequently utilized in the analyses due to their ubiquity in genomes and their value in identifying functional genes and regulatory regions correlated with environmental factors (Sang et al., 2022; Wang et al., 2022; Xiang et al., 2023). Furthermore, copy number variations (CNVs) may play important roles in environmental adaptation of plants, and they are especially prevalent in polyploid crops resulting from hybridization and polyploidization. Genes affected by CNVs have been identified to be associated with responses to biotic or abiotic stresses in some polyploid crops, but the relevant research in polyploid trees is still limited (Lye and Purugganan, 2019; Schiessl et al., 2019).

In this study, we performed genome sequencing of 47 wild *Camellia oleifera* in China. Some were from 7 natural distribution sites in the northern distribution edge of wild *C. oleifera*, and the others were from 4 natural distribution sites in the southern distribution region. For comparison, 4 relative species of *C. oleifera* were also included in the genome sequencing. Based on a large number of SNPs and 19 bioclimatic variables, the analyses of population genetics and landscape genomics were conducted to uncover population genetic structure, demographic history and patterns of genetic differentiation in wild *C. oleifera*. Besides, we attempted to detect the CNVs in the hexaploid wild *C. oleifera* for the first time. Finally, key SNPs and CNVs and their related genes were identified to uncover the molecular bases of adaptive evolution to freezing stress in wild *C. oleifera*. This study may contribute to the exploration and utilization of valuable genetic resources of wild *C. oleifera* for breeding, and facilitate our understanding on the molecular bases of adaptive evolution to freezing stress in polyploid trees.

2. Materials and methods

2.1. Plant material collection

To study the molecular bases of adaptive evolution to freezing stress in wild *Camellia oleifera*, 11 natural distribution sites of wild *C. oleifera* in China were selected according to Cui et al. (2016), including 7 natural distribution sites (LSG, LSD, DLC, DCP, HLT, HK, and QLZ) in the northern edge and 4 natural distribution sites (LFS, TM, YBS, and QLS) in the southern part of the distribution areas of wild *C. oleifera* in China (Fig. 1a and Table S1). At each natural distribution site, longitude and latitude information was recorded, at least 3 well-grown wild *C. oleifera* were selected and fresh young leaves were collected. A total of 47 leaf samples of wild *C. oleifera* were collected from the 11 natural distribution sites (Fig. 1a and Table S1). For comparison, 4 leaf samples of relative species of *C. oleifera* (*C. brevistyla*, *C. miyagii*, *C. confusa* and *C. kissi*) were obtained from the Jinhua International Camellia Species Garden, Jinhua, Zhejiang, China (Table S1). Leaf samples of wild *C. oleifera* were collected in 2015, 2016, and 2021, and leaf samples of relative species of *C. oleifera* were collected in 2013 (Table S1). After collection, each leaf sample was covered with aluminum foil, and placed into liquid nitrogen immediately, and then stored at −80 °C refrigerator in the laboratory for subsequent DNA extraction.

2.2. DNA extraction and sequencing

Genomic DNA was extracted from leaf samples using the DNAsecure Plant Kit DP320 (Tiangen biotech, Beijing, China) according to the manufacturer's instructions. The concentration and quality of DNA samples were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). High-quality DNA samples were used for DNA library preparation for genome sequencing at Beijing Genomics Institute (BGI, Wuhan, China).

The DNA library with 300–400 bp short-inserts were constructed and sequenced on the MGISEQ-2000 platform (BGI, Wuhan, China) with paired-end sequencing (2 × 150 bp) methods. Clean reads were obtained by removing reads containing adapters and/or PCR duplication and/or low qualified reads from raw reads with SOAPnuke v.2.1.7 (Chen et al., 2017).

2.3. Variants calling and annotation

During this study, high-quality reference genome of hexaploid wild *Camellia oleifera* was not available. So the reference genome of diploid *C. lanceoleosa* (Gong et al., 2022), close relative species of *C. oleifera* (Qin et al., 2023), was used in this study. Clean reads from genome sequencing were aligned to the reference genome of *C. lanceoleosa* using BWA v.0.7.17-r1188 (Li, 2013). Based on SAMtools v.1.18, SAM files were converted into BAM files and sorted

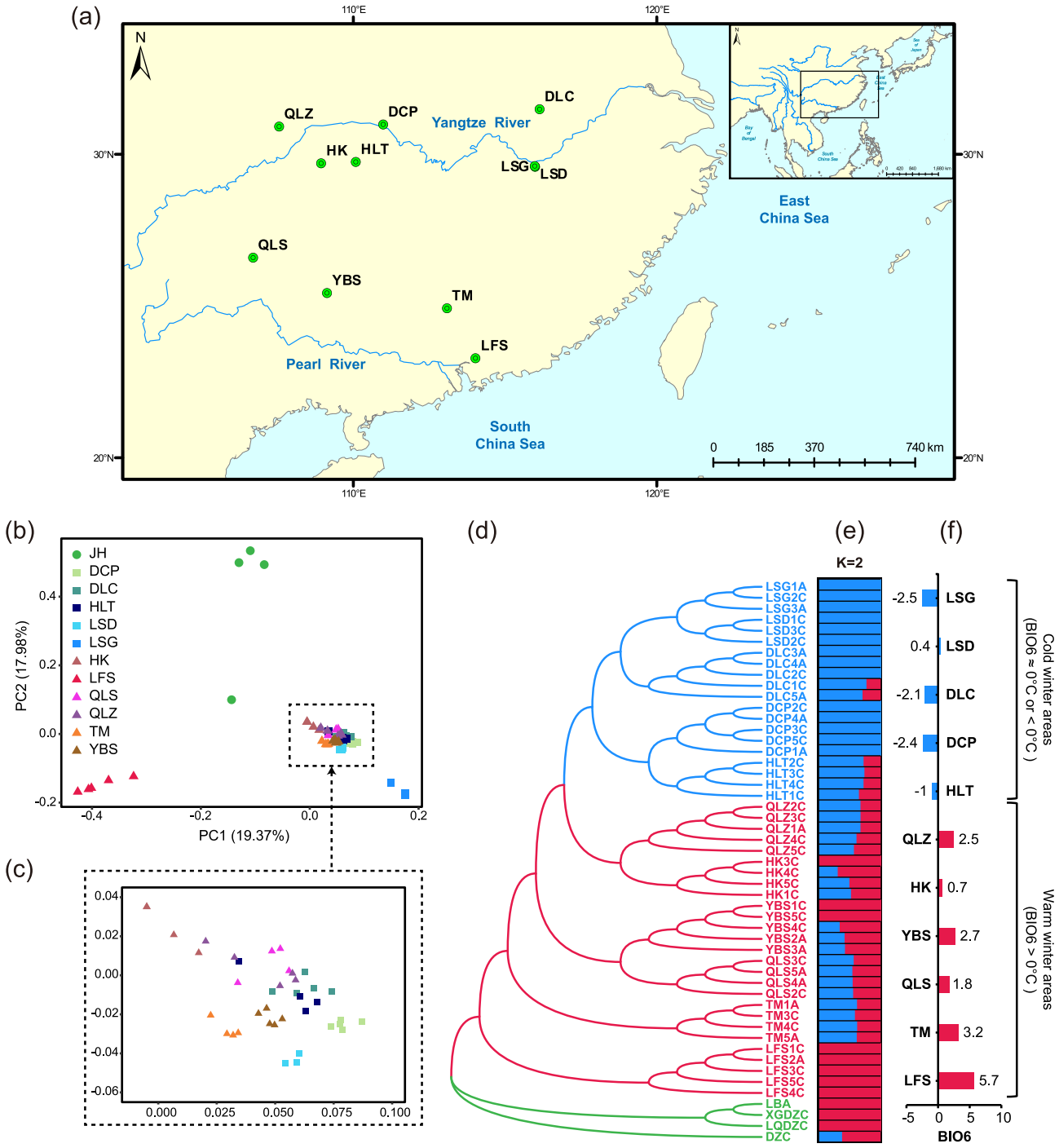


Fig. 1. Natural distribution sites and population genetics analyses of wild *Camellia oleifera* and relative species. (a) Geographical distribution of 11 natural distribution sites of wild *C. oleifera* (green dots). (b) Principal component analysis (PCA) for the 47 samples of wild *C. oleifera* and 4 samples of relative species of *C. oleifera*. (c) Zooming in a mixing region within smaller scales. Colored circles represent the relative species of *C. oleifera*, colored squares represent wild *C. oleifera* in the cold winter areas, colored triangles represent wild *C. oleifera* in the warm winter areas. (d) Neighbor-joining phylogenetic tree for the 47 samples of wild *C. oleifera* and 4 samples of relative species of *C. oleifera*, with blue clades indicating wild *C. oleifera* in the cold winter areas, red clades indicating wild *C. oleifera* in the warm winter areas, and green clades indicating relative species of *C. oleifera*. (e) Model-based population genetic structure with the best K value (K = 2). (f) Key bioclimatic variable BIO6 values at the 11 natural distribution sites of wild *C. oleifera*. BIO6: Min Temperature of Coldest Month.

(Danecek et al., 2021). Polymerase chain reaction duplicates were marked using Sambamba v.1.0.0 (Tarasov et al., 2015). The HaplotypeCaller, CombineGVCFs, GenotypeGVCFs, MergeVcfs and SelectVariants functions from GATK v.4.4.0.0 were used to perform joint SNP calling (Van der Auwera et al., 2013). Then the VariantFiltration and SelectVariants functions from GATK v.4.4.0.0 were used to mark and filter unqualified SNPs, with parameters "QD < 2.0||MQ < 40.0||FS > 60.0||SOR > 3.0||MQRankSum < −12.5||ReadPosRankSum < −8.0". And bi-allelic SNPs were preserved by BCFtools v.1.17 (Danecek et al., 2021). SNP annotation was performed using ANNOVAR (Wang et al., 2010). The range of upstream and downstream was defined as 2000 bp.

CNVcaller was used for detecting the integrated copy number variation regions (CNVRs) (Wang et al., 2017). Firstly, the reference genome was segmented into overlapping sliding windows based on default parameters; secondly, the absolute copy number of all windows in each sample genome sequencing data was calculated; then, the CNVRs were determined according to Pearson correlation coefficient with significant level of 0.01; finally, clustering the input samples into genotypes used Gaussian mixture modes. CNVRs with silhouette score >0.5 were considered reliable and accurate. CNVR annotation was performed using ANNOVAR (Wang et al., 2010). The range of upstream and downstream was defined as 2000 bp as well.

2.4. Detection of key bioclimatic variables

The 19 bioclimatic variables were downloaded from WorldClim (https://www.worldclim.org) (Fick and Hijmans, 2017). They are the averages for the years 1970–2000 and the spatial resolution is 30 arc-seconds. According to the longitude and latitude information of 11 natural distribution sites of wild *Camellia oleifera*, ArcMap v.10.8 was used to extract the 19 bioclimatic variables corresponding to each natural distribution site. The comprehensive contribution (percent contribution and permutation importance) of these bioclimatic variables to wild *C. oleifera* distribution was calculated by Maxent v.3.4.3 based on the 19 bioclimatic variables at each natural distribution site (Phillips et al., 2006). Bioclimatic variables with top three comprehensive contribution were considered as key bioclimatic variables. The correlation coefficient between the 19 bioclimatic variables was calculated using the R base package stats v.4.3.0 with Spearman method. The R package *corrplot* v.0.92 was used to visualize the results.

2.5. Basic analyses of population genetics

To reduce false positives, SNPs with missing rate >0.1 and minor allele frequency (MAF) < 0.05 were removed using PLINK v.1.9 (Purcell et al., 2007). Principal component analysis (PCA) was conducted by GCTA v.1.94.1 (Yang et al., 2011), R package tidyverse v2.0.0 and export v.0.3.0 were used to visualize the results. The neighbor-joining tree was constructed using the PHYLIP v.3.697 (Felsenstein, 1989), and Interactive Tree Of Life (iTOL, https://itol.embl.de/, v6) website was used for phylogenetic tree display (Letunic and Bork, 2021). After missing rate and MAF filtering, SNPs were further pruned for linkage disequilibrium (LD) with parameters "–indep-pairwise 50 10 0.2" using PLINK v.1.9 (Purcell et al., 2007). Model-based population structure estimation was conducted using ADMIXTURE v.1.3.0 and pong v.1.5 was used to visualize the results (Alexander et al., 2009; Behr et al., 2016).

In addition, after missing rate and MAF filtering, all SNPs were used to estimate and compare LD decay patterns by PopLDdecay v.3.42, with parameter "MaxDist" of 500 kb (Zhang et al., 2018).

2.6. Demographic history inference

Stairway Plot v.2 and SMC++ v.1.15.4 were used to infer the history of effective population size changes (Terhorst et al., 2017; Liu and Fu, 2020). For stairway Plot, the folded site frequency spectrum was generated in ANGSD v0.940 with the filtering parameters "minMapQ 1 -minQ20", and the genome of *Camellia lanceoleosa* was used as reference (Korneliussen et al., 2014); in the subsequent analysis, parameters "mu: 6.5e-9" and "year_per_generation: 3" were used.

For SMC++ analysis, SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml) was used to perform the masking step to delineate the largely uncalled regions; using the SNPs after further LD pruned, mutation rate per site per generation was set to 6.5e-9, and generation time was set to 3 years.

2.7. Detection of selected variants

Utilizing the SNPs after missing rate and MAF filtering, fixation index (F_{ST}) with 100 kb nonoverlapping windows were calculated by VCFtools v.0.1.16 (Danecek et al., 2011). The python version of XP-CLR was used for selective sweep analysis with 100 kb nonoverlapping windows, and other parameters were "–rrate 1e-8 –ld 0.95 –maxsnps 5000" (Chen et al., 2010). The SNPs in the windows with both top 1% F_{ST} and top 1% XP-CLR scores were considered as selected SNPs.

CNVRs with silhouette score >0.5 were used to calculate V_{ST}. V_{ST} is conceptually similar to F_{ST}, and it ranges from 0 (undifferentiated) to 1 (fully differentiated) (Redon et al., 2006). V_{ST} was calculated as follows:

$$V_{ST} = \frac{V_{total} - (V_{pop1} \times N_{pop1} + V_{pop2} \times N_{pop2} + \cdots + V_{popx} \times N_{popx})/N_{total}}{V_{total}}$$

where based on the absolute copy number from each sample, V_{total} is the total variance, V_{popx} is the variance for each respective population (group), N_{total} is the total sample size, and N_{popx} is the sample size for each respective population (group) (Redon et al., 2006; Zhao and Gibbons, 2018). The CNVRs with top 1% V_{ST} were considered as selected CNVRs. All Manhattan plots were drawn using the R package CMplot v.4.5.0 (Yin et al., 2021).

2.8. Genome-environment association analysis

Key bioclimatic variables (top three comprehensive contribution) and selected SNPs (both top 1% F_{ST} and top 1% XP-CLR scores) were used to performed genome–environment association analysis, which included two methods: latent factor mixed models (LFMM) and redundancy analysis (RDA). LFMM was conducted in the R package *lfmm* v.1.1, and each key bioclimatic variable was analyzed separately (Frichot et al., 2013). The SNPs with p-value <0.01 were considered as SNPs significantly correlated with key bioclimatic variables. The Manhattan plots were drawn using the R package CMplot v.4.5.0.

RDA was used to identify the association between genetic variation and multiple environmental axes. RDA was performed with the R package vegan v.2.6-4 (https://cran.r-project.org/web/packages/vegan/index.html), and 2.5 standard deviation cutoff (two-tailed $p = 0.012$) was selected to identify SNPs significantly correlated with each key bioclimatic variable based on the strongest correlations (Forester et al., 2018). Finally, the intersection of the results between LFMM and RDA was the SNPs that were significantly correlated with each key bioclimatic variable, which were called adaptive SNPs.

To explore the influence of geography and environment on genetic variation of adaptive SNPs (the SNPs identified by both LFMM and RDA) and neutral SNPs (the SNPs after further LD pruned), based on the R package *vegan* v.2.6-4, isolation-by-distance (IBD) and isolation-by-environment (IBE) analyses were performed: Mantel and/or partial Mantel tests were separately used to test for associations between $F_{ST}/(1 - F_{ST})$ and geographical distance, and between $F_{ST}/(1 - F_{ST})$ and environmental distance. Significance was determined by 999 permutations. VCFtools v.0.1.16 was used to calculate the F_{ST} among different populations. Geographical distance was calculated based on longitude and latitude information, and environmental distance was represented by Euclidean distance of bioclimatic variables.

2.9. Identification and functional analysis of key variants

The genotypes of SNP were divided into three types: homozygous reference allele, homozygous alternative allele and heterozygous allele. Based on the adaptive SNPs, SNPs with genotype differentiation exceeding 50% between groups were considered key SNPs. LDBlockShow v.1.40 was used to generate LD heatmap (Dong et al., 2020). In this study, when a key SNP was located in a gene, this gene was considered as the related gene of the key SNP; when a key SNP was located in the intergenic region, both of the two genes near the key SNP were considered as the related genes of the key SNP. The full-length cDNA sequences of the related genes of key SNPs were extracted by gffread v.0.12.7 (Pertea and Pertea, 2020). Then, the cDNA sequences were compared to the key unigenes (full-length transcripts) associated with freezing tolerance in the leaf transcriptome data of *Camellia oleifera* from Xie et al. (2023) using BLAST v.2.13.0+ (Altschul et al., 1990), and E-value < 1e-50 were considered as a credible comparison result. The transcriptome data from Xie et al. (2023) were divided into the field and lab experiments. In the field experiment, leaf samples of wild *C. oleifera* from the wild *C. oleifera* distribution site in the Lu Mountain were collected in different air temperature periods: before cold acclimation (D1, D2 and D3), during cold acclimation (D4 and D5), and under freezing temperature (D6). In the lab experiment, leaf samples of wild *C. oleifera* in the Lu Mountain (LSG) and a local *C. oleifera* cultivated variety *C. oleifera* variety "Ganwu 1" (GW1) were subjected to −10 °C freezing temperature treatment with time gradient (a: 0 h, b: 1 h, c: 3 h, d: 6 h, e: 16 h, f: 32 h). PlantCARE (https://bioinformatics.psb.ugent.be/webtools/plantcare/html/) was used to predict the cis-regulatory elements (CREs) where the key SNPs were located (Lescot et al., 2002).

According to absolute copy number ≈ 1, 2, 3, 4, 5 and 6 or >6, the genotypes of CNVR were divided into six types. Selected CNVRs (top 1% V_{ST}) located in the exon regions were used. CNVRs with genotype differentiation exceeding 50% between groups were considered as the key CNVRs. The related genes of the key CNVRs located in the exon regions were annotated using eggNOG-mapper v.2.1.12, and the functional annotation results from Kyoto Encyclopedia of Genes and Genomes (KEGG) database were plotted by SRplot (http://www.bioinformatics.com.cn/SRplot) (Kanehisa et al., 2004; Cantalapiedra et al., 2021; Tang et al., 2023). Based on the same methods as for identifying the key SNPs, the related genes of the key CNVRs located in the exon regions were aligned to the key unigenes associated with freezing tolerance in the leaf transcriptome data of *Camellia oleifera* from Xie et al. (2023). Expression level heat maps of the related genes of the key SNP and the CNVRs were made using the R package pheatmap v.1.0.12.

During the review process of this study, the chromosome-scale genome of hexaploid *Camellia oleifera* has been published (Zhu et al., 2024). Due to time constraints, we are unable to redo all the analyses. To ensure the reliability of our analysis results, the full-length cDNA sequences of the *C. oleifera* genome were extracted by gffread v.0.12.7. Then, for the related genes of key SNPs and CNVRs with annotated functions, the full-length cDNA sequences of these genes in the *C. lanceoleosa* genome were compared to the cDNA sequences of the *C. oleifera* genome with BLAST v.2.13.0+, and E-value < 1e-50 was considered as a credible comparison result.

3. Results

3.1. Summary of genome sequencing, variants calling and annotation

A total of 1.69 Tb clean data were obtained from genome sequencing of the 47 samples of wild *Camellia oleifera*, with the average of 35.95 Gb per sample, and the average Q20, Q30 and GC content of the clean data were 97.76%, 93.27% and 39.75%, respectively (Table S2). For the 4 samples of relative species of *C. oleifera*, a total of 432.05 Gb clean data were obtained, with the average of 108.01 Gb per sample, and the average Q20, Q30 and GC content of the clean data were 98.25%, 94.50% and 38.58%, respectively (Table S2). Overall, high-quality sequencing data were obtained, providing solid bases of the subsequent analyses.

With aligning to the reference genome of *Camellia lanceoleosa*, the average coverage rate was 79.62% and the average sequencing depth was 11.48× for each sample of wild *C. oleifera*, and the average coverage rate and the average sequencing depth for each sample of the relative species of *C. oleifera* were 81.96% and 34.86×, respectively (Table S2). A total of 271,704,193 SNPs were obtained in the 47 samples of wild *C. oleifera* and the 4 samples of relative species of *C. oleifera* after the basic filtering of GATK (Table S3). The annotation result showed that most of the SNPs were located in the intergenic regions, accounting for 82.26%, and only 1.49% of the SNPs were located in the exon region (Table S3). A total of 269,113 CNVRs were obtained in the 47 samples of wild *C. oleifera*, and 117,501 CNVRs had silhouette score >0.5 (Table S3). More than 70% of the CNVRs were located in the intergenic regions, followed by the exon regions, approaching 20% (Table S3). In sum, the genetic diversity in the genomes of wild *C. oleifera* was high especially in the intergenic regions, containing huge amounts of SNPs and CNVRs.

3.2. Group division of wild Camellia oleifera based on key bioclimatic variables and population genetics analyses

The results of Maxent showed that bioclimatic variable "Min Temperature of Coldest Month" (BIO6) had the highest comprehensive contribution, and the percent contribution and permutation importance were 45.7% and 51% respectively, and comprehensive contribution ranked the second and the third were bioclimatic variable "Mean Diurnal Range" (BIO2) and "Precipitation Seasonality" (BIO15) respectively (Tables S4 and S5). The three key bioclimatic variables (BIO2, BIO6 and BIO15) had the absolute value of pairwise correlation coefficients <0.8 (Fig. S1).

In the PCA results, overall, the relative species (JH) of *Camellia oleifera* were separated from the wild *C. oleifera* (Fig. 1b). Based on the PC1 axis, the wild *C. oleifera* from the 11 natural distribution sites were divided into three parts: LFS population, LSG population and other populations (Fig. 1b). Within smaller scales, the other populations could be divided into two parts: HK, QLS, QLZ, TM and YBS populations, and DCP, DLC, HLT and LSD populations (Fig. 1c). And HK, QLS, QLZ, TM and YBS populations were closer to LFS population, DCP, DLC, HLT and LSD populations were closer to LSG population (Fig. 1c). The phylogenetic tree showed that the 4 relative species of *C. oleifera* were grouped together as an independent clade (green clades), and wild *C. oleifera* from the 11

populations were grouped together in a large clade with wild *C. oleifera* from the same population clustered together (Fig. 1d). In the clade of wild *C. oleifera* from the 11 populations, DCP, DLC, HLT, LSD and LSG populations were grouped together (blue clades), separated from HK, LFS, QLS, QLZ, TM and YBS populations (red clades). The best result of the population genetic structure analysis was of K = 2 (Table S6), and wild *C. oleifera* from the 11 populations could be divided into two groups: one including DCP, DLC, HLT, LSD and LSG populations, and the other including HK, LFS, QLS, QLZ, TM and YBS populations (Fig. 1d and e). The results of the PCA, the phylogenetic tree and the population genetic structure were in close agreement with each other.

In addition, based on the key bioclimatic variable BIO6 with the highest comprehensive contribution to wild *Camellia oleifera* distribution, wild *C. oleifera* from the 11 populations could also be divided into two groups: the group in the cold winter areas with BIO6 ≈ 0 °C or < 0 °C, including DCP, DLC, HLT, LSD and LSG populations; and the group in the warm winter areas with BIO6 > 0 °C, including HK, LFS, QLS, QLZ, TM and YBS populations (Fig. 1f and Table S4). This grouping result based on the key bioclimatic variable was in line with the above results of population genetic structure, indicating a close correlation between them.

3.3. *Demographic history of two groups of wild* Camellia oleifera

The LD decay patterns of the two groups of wild *Camellia oleifera* showed relatively similar trends, but wild *C. oleifera* in the warm winter areas exhibited faster LD decay, suggesting that the nucleotide diversity of wild *C. oleifera* in the warm winter areas was higher than that of wild *C. oleifera* in the cold winter areas (Fig. S2). Stairway Plot results indicated that the overall trend of changes in the effective population size (N_e) of wild *C. oleifera* in the cold winter areas and the warm winter areas were similar (Fig. 2). Before the last glaciation (LG), the N_e of the two groups of wild *C. oleifera* were relatively stable, and their maximum N_e were the same; since LG, the N_e of the two groups of wild *C. oleifera* declined dramatically (Fig. 2). The N_e of wild *C. oleifera* in the cold winter areas dropped several stairs, and it might have gone through several serious bottleneck events; the N_e of wild *C. oleifera* in the warm winter areas declined at a relatively steady rate. At present, the N_e of wild

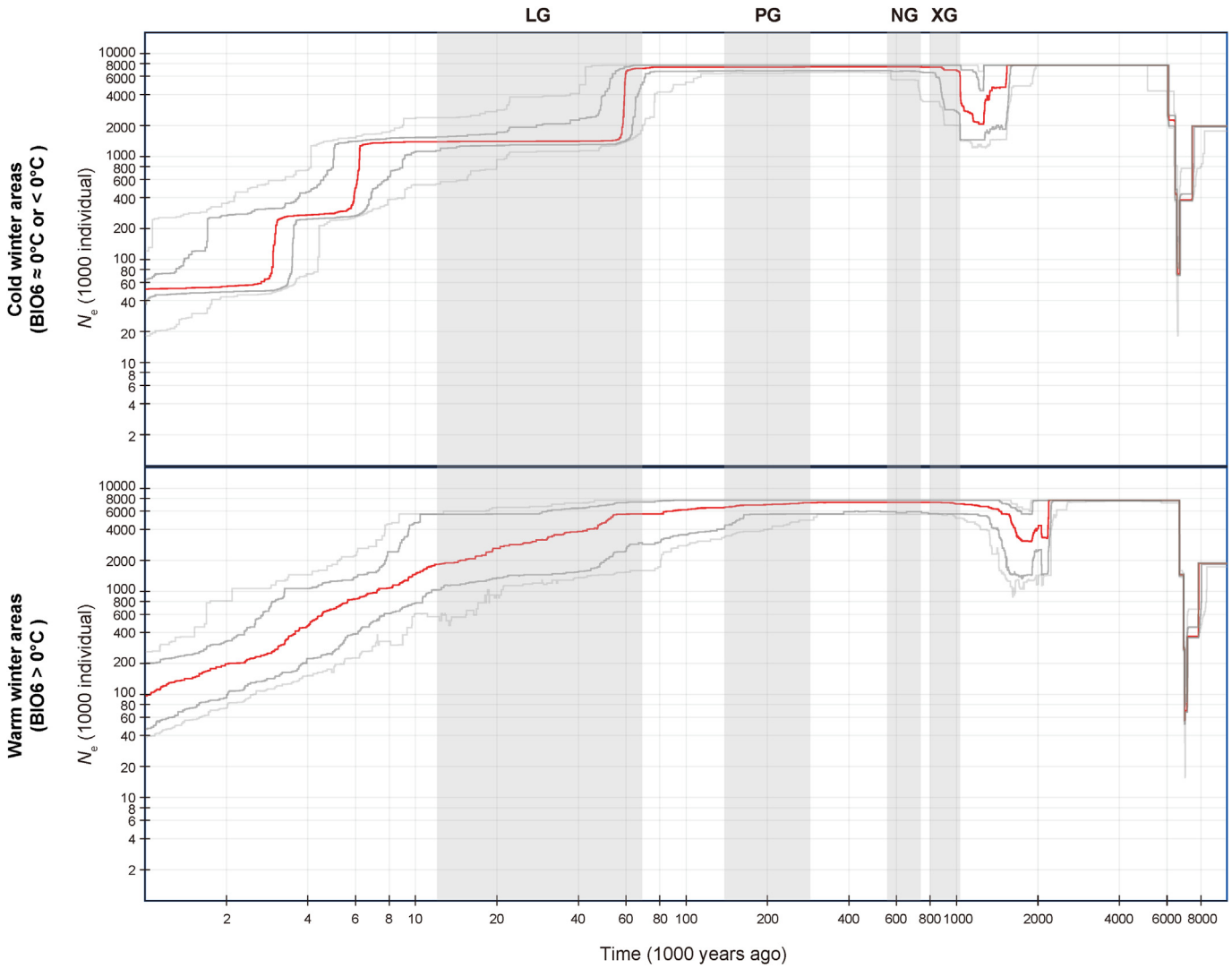


Fig. 2. Demographic history of the two groups of wild *Camellia oleifera* in the cold winter areas and in the warm winter areas from Stairway Plot. In the plots, the red line is the estimation (median), the thick gray lines define the 75% confidence interval, and the light gray lines define the 95% confidence interval. Four gray rectangle regions indicate the last glaciation (LG, 11–70 kya), penultimate glaciation (PG, 130–300 kya), Naynayxungla Glaciation (NG, 500–780 kya), and Xixiabangma Glaciation (XG, 800–1170 kya).

C. oleifera in the cold winter areas was lower than that of wild *C. oleifera* in the warm winter areas, suggesting lower genetic diversity in the former (Fig. 2). Despite the fact that SMC++ and Stairway Plot employ different approaches to detect demographic history, the SMC++ results also showed that the overall trends of the N_e of the two groups of wild *C. oleifera* were decreasing, and at present the N_e of wild *C. oleifera* in the warm winter areas was larger than that of wild *C. oleifera* in the cold winter areas (Fig. S3). These results imply that the two groups of wild *C. oleifera* might have experienced different selection pressures and/or population

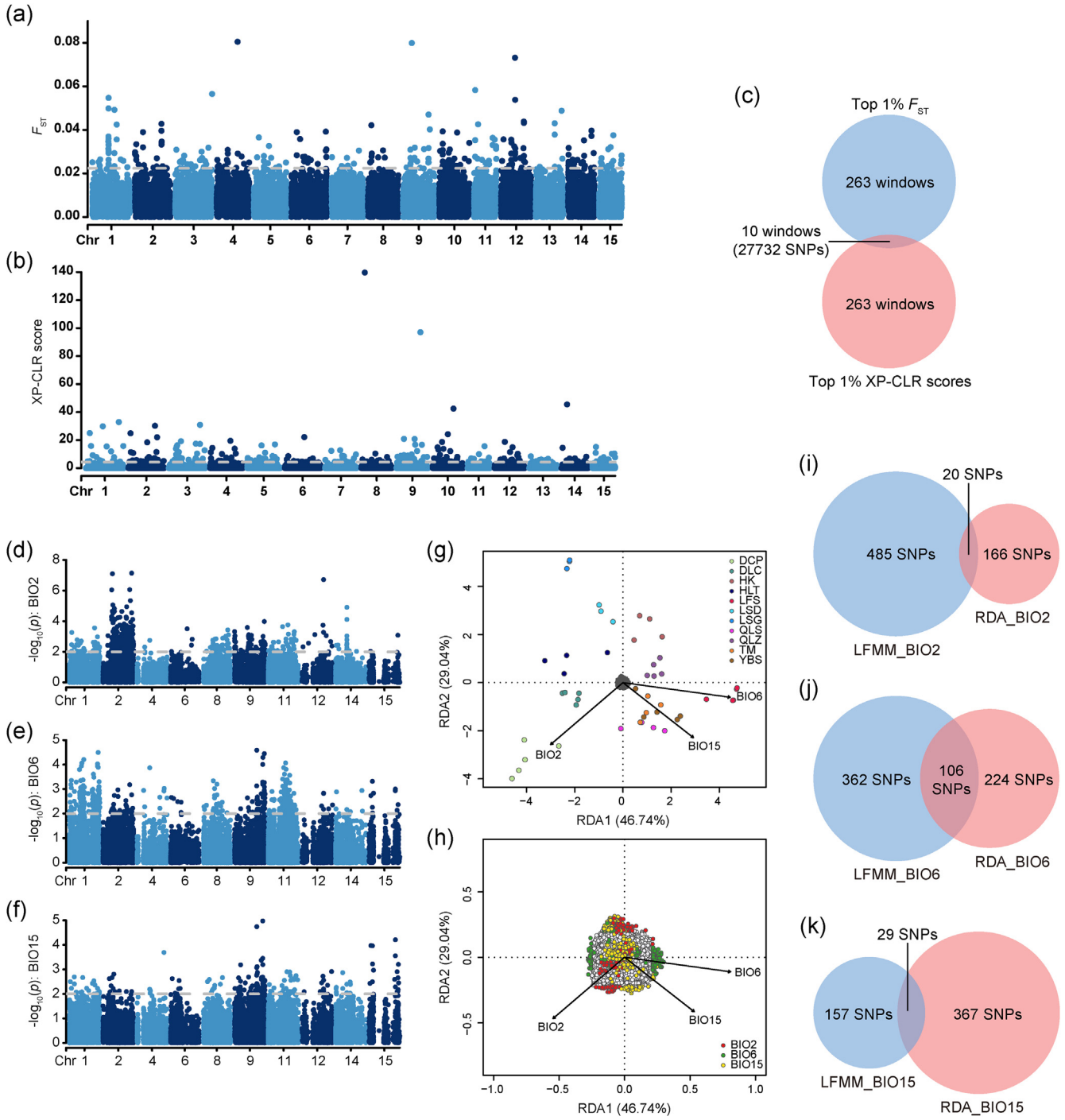


Fig. 3. Detection of selected SNPs and environment association analyses. Between the wild *Camellia oleifera* in the cold winter areas and the warm winter areas, (a) Manhattan plot of the fixation index (F_{ST}), the gray horizontal dashed line represents the top 1% F_{ST}, and (b) Manhattan plot of the XP-CLR, the gray horizontal dashed line represents the top 1% XP-CLR scores. (c) Venn diagram of windows in the top 1% F_{ST} and top 1% XP-CLR scores. Based on the latent factor mixed models (LFMM), Manhattan plots for SNPs associated with (d) BIO2, (e) BIO6, and (f) BIO15. All gray horizontal dashed lines represent significance threshold of $p = 0.01$. (g) The ordination plots of redundancy analysis (RDA), where individuals are colored circles, SNPs are in the center and black vectors are the key bioclimatic variables, and (h) this plot focuses on the SNPs, coloring the SNPs according to the key bioclimatic variables the most strongly correlated with, and 2.5 standard deviation cutoff (two-tailed $p = 0.012$) was used. In the LFMM and RDA, Venn diagram of SNPs significantly correlated with (i) BIO2, (j) BIO6, and (k) BIO15. BIO2: Mean Diurnal Range, BIO6: Min Temperature of Coldest Month, BIO15: Precipitation Seasonality.

bottlenecks in the past, with wild *C. oleifera* in the cold winter areas experiencing stronger selection pressure and/or population bottlenecks than wild *C. oleifera* in the warm winter areas.

3.4. Key SNPs in the adaptive SNPs

Between the two groups of wild *Camellia oleifera* in the cold winter areas and in the warm winter areas, both top 1% F_{ST} and top 1% XP-CLR scores detected 273 windows (Fig. 3a and b). There were 10 windows at their intersection, which involved chromosomes 1, 2, 4, 6, 8, 9, 11, 12, 14 and 15, and a total of 27,732 SNPs within these windows (Fig. 3c). The 27,732 selected SNPs and the three key bioclimatic variables (BIO2, BIO6 and BIO15) were used to performed LFMM and RDA analyses. In the LFMM results, 505 SNPs, 468 SNPs and 186 SNPs were significantly correlated with BIO2, BIO6 and BIO15, respectively (Fig. 3d–f). The RDA results showed that the RDA1 axis and the RDA2 axis explained 46.74% and 29.04% of the genomic variation in wild *C. oleifera*, and wild *C. oleifera* in the cold winter areas and the warm winter areas were completely separated based on the RDA1 axes (Fig. 3g). The RDA results indicated that 186 SNPs, 330 SNPs and 396 SNPs were significantly correlated with BIO2, BIO6 and BIO15, respectively (Fig. 3h). A total of 155 SNPs were found in common between the LFMM results and the RDA results, and 20 SNPs were significantly correlated with BIO2, 106 SNPs were significantly correlated with BIO6, and 29 SNPs were significantly correlated with BIO15 (Fig. 3i–k and Table S7).

These 155 adaptive SNPs displayed moderate and significant patterns of IBD ($r = 0.4228$, $p = 0.002$), and strong and significant patterns of IBE (for the three key bioclimatic variables) ($r = 0.7035$, $p = 0.001$) (Fig. 4a and b). After controlling for the effect of geographical distance, the adaptive SNPs still showed strong and significant patterns of IBE (for the three key bioclimatic variables) ($r = 0.637$, $p = 0.001$), suggesting that genetic variation of the adaptive SNPs was mainly influenced by the three key bioclimatic variables. The neutral SNPs (the SNPs after further LD pruned) exhibited different patterns from the adaptive SNPs, with weak and

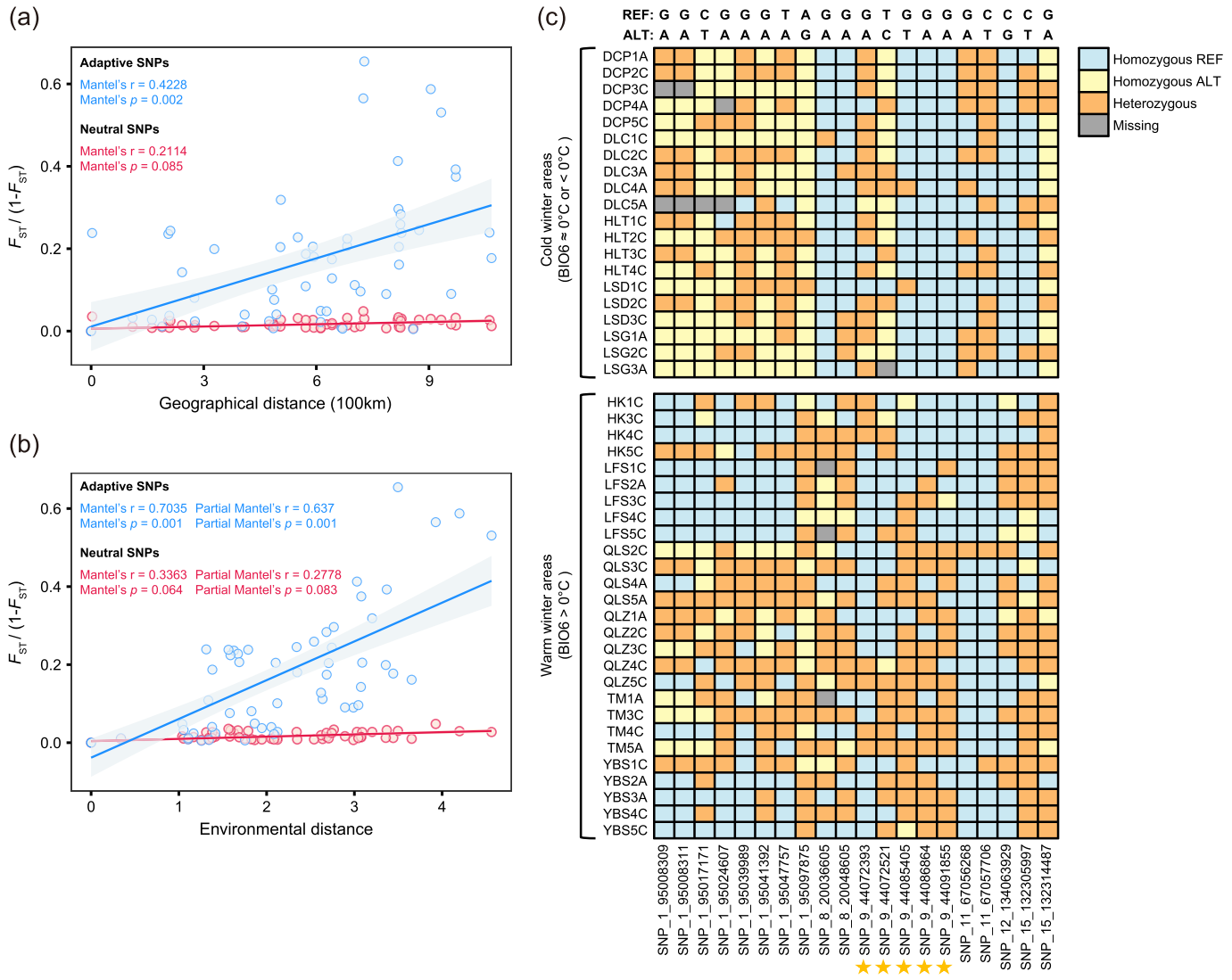


Fig. 4. Isolation-by-distance (IBD) and isolation-by-environment (IBE) analyses and genotypes of key SNPs. Based on the adaptive SNPs (blue dots and blue line) and neutral SNPs (red dots and red line), (a) IBD analyses (Mantel test, two-sided), and (b) IBE analyses (Mantel test, two-sided; partial Mantel test, two-sided, controlling for the effect of geographical distance) were performed, and the three key bioclimatic variables (BIO2, BIO6 and BIO15) were used for the IBE analysis. All shadows of linear regression denote the 95% confidence interval. (c) Genotypic heatmap of 20 key SNPs, the yellow five-pointed star indicates that the related gene of this SNP successfully matched with key unigenes associated with freezing tolerance in the transcriptome data from Xie et al. (2023). REF: reference allele, ALT: alternative allele. BIO2: Mean Diurnal Range, BIO6: Min Temperature of Coldest Month, BIO15: Precipitation Seasonality.

insignificant patterns of IBD ($r = 0.2114$, $p = 0.085$) and IBE (for the three key bioclimatic variables) ($r = 0.2778$, $p = 0.083$) (Fig. 4a and b). Furthermore, for the IBE analyses based on the 19 bioclimatic variables, with or without controlling for the effect of geographical distance, the adaptive SNPs showed strong and significant patterns of IBE, further demonstrating the close relationship between the genetic variation of adaptive SNPs and the environment (Fig. S4).

Among these 155 adaptive SNPs, 20 key SNPs with genotype differentiation exceeding 50% between wild *Camellia oleifera* in the cold winter areas and the warm winter areas, with 18 SNPs significantly correlated with BIO6, and 2 SNPs significantly correlated with BIO15, and they were all located in the intergenic regions (Fig. 4c and Table S7). A total of 12 genes were associated with these 20 key SNPs. *LOK49_LG08G00640* encodes "cytochrome P450 82A3", and this gene successfully matched two key unigenes (*isoform_222481* and *isoform_284377*) associated with freezing tolerance in the transcriptome data from Xie et al. (2023) (Fig. 5a and Table 1). Under freezing temperature, the expressions of *isoform_222481* and *isoform_284377* were greatly upregulated and showed higher expression in the LSG samples with strong freezing tolerance than those in the GW1 samples with weak freezing tolerance (Fig. 5a). A total of 5 key SNPs were found near *LOK49_LG08G00640*, and according to the LD heatmap, except for SNP_9_44072393, there was linkage disequilibrium among the other 4 SNPs (Fig. 5b). In these 5 key SNPs, SNP_9_44085405 was significantly correlated with BIO6 and located within a cis-regulatory element (CRE) involved in salicylic acid (SA) responsiveness; SNP_9_44086864 was significantly correlated with BIO15 and located within a CRE with unknown function (Fig. 5c and Table S7). In most of wild *C. oleifera* in the cold winter areas, the genotypes of SNP_9_44085405 and SNP_9_44086864 were homozygous of the reference allele, and there was no mutation in their CRE; but in most of wild *C. oleifera* in the warm winter areas the genotypes showed heterozygous alleles, which leads to mutation in their CRE (Fig. 4c). The expression level of *LOK49_LG08G00640* may be regulated by nearby CRE, and the function of these CRE may be influenced by the genotypes of SNP_9_44085405 and SNP_9_44086864.

3.5. Selected CNVRs and key CNVRs

A total of 1175 selected CNVRs were obtained in top 1% V_{ST} between the wild *Camellia oleifera* in the cold winter areas and in the warm winter areas (Fig. 6a and Table S8). Only one selected CNVR, CNVR_9_43980001_44136800, overlapped with 4277 selected SNPs (out of 27,732 selected SNPs), including 24 adaptive SNPs (out of 155 adaptive SNPs) and the 5 key SNPs near *LOK49_LG08G00640* (Fig. 5b). For CNVR_9_43980001_44136800, in most wild *C. oleifera* in the cold winter areas, the absolute copy number ≈ 2; whereas in wild *C. oleifera* in the warm winter areas, the absolute copy number ≈ 2 accounts for half, and the absolute copy number ≈ 3 accounts for the other half (Fig. 5b). There may be interactions between 5 key SNPs near *LOK49_LG08G00640* and CNVR_9_43980001_44136800, which may affect the expression level of *LOK49_LG08G00640* together.

Among the selected CNVRs, 297 CNVRs were located in the exon regions (Fig. 6a and Table S8). The functional annotation results of related genes of the 297 CNVRs showed that the functions of genes were divided into five classes with 70 terms, including multiple functional pathways involved in responses to cold stress, such as "starch and sucrose metabolism", "fatty acid biosynthesis", "plant hormone signal transduction", "MAPK signaling pathway-plant" and "circadian rhythm-plant" (Fig. S5). Among these 297 selected CNVRs, 15 key CNVRs were found with genotype differentiation exceeding 50% between wild *C. oleifera* in the cold winter areas and the warm winter areas (Fig. 6b and Table S8). A total of 23 genes were associated with these 15 key CNVRs. *LOK49_LG07G01297* encodes "AFG1-like ATPase", and this gene successfully matched a key unigene (*isoform_75968*) associated with freezing tolerance in the transcriptome data from Xie et al. (2023) (Fig. 6c and Table 2). Under freezing temperature, the expression of *isoform_75968* was upregulated and showed higher expression in the LSG sample with strong freezing tolerance than that in the GW1 sample with weak freezing tolerance (Fig. 6c). CNVR_7_80731601_80744000 covered partial exon regions of *LOK49_LG07G01297* (Fig. 6d), and its absolute copy number $\approx$ 5, 6 or >6 in most wild *C. oleifera* in the cold winter areas, while absolute copy number $\approx$ 4 in most wild *C. oleifera* in the warm winter areas (Fig. 6b). The expression level of *LOK49_LG07G01297* may be closely related to the absolute copy number of CNVR_7_80731601_80744000.

3.6. Related genes of key SNPs and CNVRs in the hexaploid Camellia oleifera genome

In our study, based on the diploid *Camellia lanceoleosa* genome, we found that 12 genes were associated with key SNPs, while another 23 genes were associated with key CNVRs. Among these genes, 19 genes were annotated with specific functions (not labeled as "hypothetical protein") (Tables 1 and 2). For the BLAST results of the full-length cDNA sequences of the 19 genes, most of them could be found in the hexaploid *C. oleifera* genome with high similarity (Table S9).

4. Discussion

4.1. Freezing stress is one of the key environmental factors promoting the adaptive evolution of wild Camellia oleifera

Camellia oleifera is a representative plant species in subtropical evergreen broad-leaved forests in China. Our study showed that BIO6 was the key bioclimatic variable with the greatest impact on the distribution pattern of the 11 wild *C. oleifera* populations (Table S5). BIO6 represents "min temperature of the coldest month", and the 11 wild *C. oleifera* populations could be divided into two groups based on BIO6, i.e. wild *C. oleifera* in the cold winter areas with BIO6 $\approx 0\ ^\circ C$ or $< 0\ ^\circ C$, and wild *C. oleifera* in the warm winter areas with BIO6 $> 0\ ^\circ C$ (Fig. 1f and Table S4). From the results of PCA, phylogenetic tree, and the optimal population genetic structure ($K = 2$), there was genetic differentiation between wild *C. oleifera* in the cold winter areas and the warm winter areas (Fig. 1b–e). The LD decay and demographic history indicated that wild *C. oleifera* in the cold winter areas and the warm winter areas might have experienced different degrees of selective pressure and/or population bottleneck in the past (Figs. 2, S2 and S3).

Previous studies showed that the climate underwent significant changes with the temperature plummeted during the ice ages in the Pleistocene, which had a profound impact on the distribution of subtropical evergreen broad-leaved forests (Yu et al., 2000; Ni et al., 2010). Subtropical evergreen broad-leaved forests not only had refugia in the south (south of 24°N) but also in the north (24°–33°N), conforming to the pattern of multiple refugia (Ye et al., 2017). Wild *Camellia oleifera* in the cold winter areas are distributed in the northern distribution edge of wild *C. oleifera* in China (Fig. 1a and Table S1). Demographic history inferred from Stairway Plot showed that wild *C. oleifera* in the cold winter areas might have gone through several serious bottleneck events since LG, leading to several rapid drops in N_e (Fig. 2). We speculated that this group might come from the northern refugia during LG, and the geographical isolation caused by habitat fragmentation and the strong selective pressure caused by extreme cold climate might

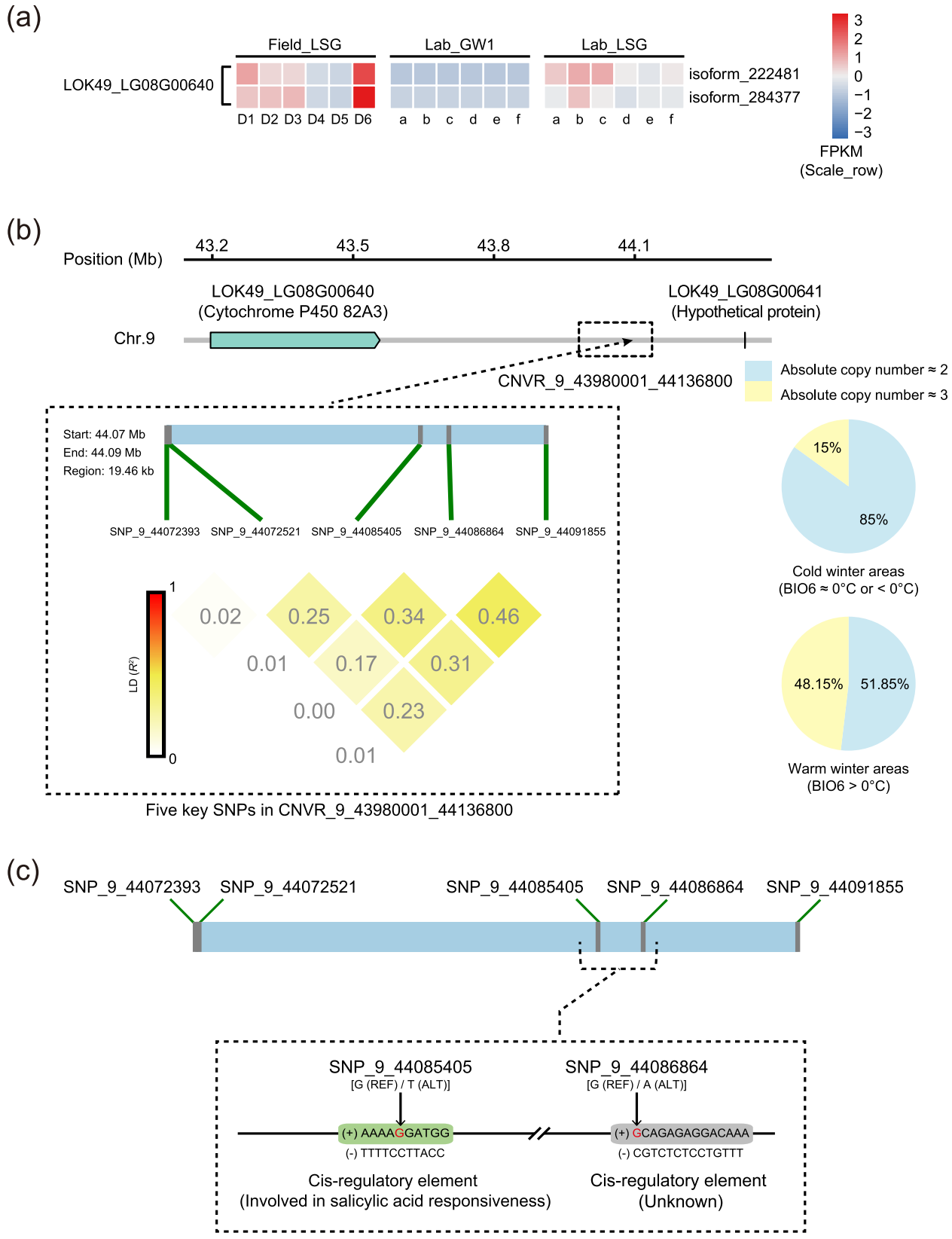


Fig. 5. Analyses of key SNPs and their related genes. (a) Based on the unigenes associated with freezing tolerance in the transcriptome data from Xie et al. (2023), heatmap of expression levels of unigenes corresponding to the related gene *LOK49_LG08G00640* of key SNPs. (b) The position of 5 key SNPs located within CNVR_9_43980001_44136800 near *LOK49_LG08G00640* in the genome, the degree of linkage disequilibrium (LD) degree between the 5 key SNPs, and genotypes of CNVR_9_43980001_44136800 for the two groups of wild *Camellia oleifera* in the cold winter areas and the warm winter areas. (c) Cis-regulatory elements (CREs) where the key SNPs are located, identified by PlantCARE website. REF: reference allele, ALT: alternative allele.

Table 1
Basic information of 12 related genes of 20 key SNPs with genotype differentiation exceeding 50% between wild *Camellia oleifera* in the cold winter areas and the warm winter areas. GENE: ID of SNPs related genes in the genome, CHR: the chromosome where the genes are located in the genome, START: the starting position of genes in the genome, END: the ending position of genes in the genome, LEN: the length of genes, ANN: annotation results of genes in the genome.

GENE	CHR	START	END	LEN	ANN
LOK49_LG01G01574	1	94990718	94993516	2799	Protein Root initiation defective 3
LOK49_LG01G01575	1	95140319	95142103	1785	ABC transporter G family member 8
LOK49_LG09G00427	8	20017521	20017679	159	Hypothetical protein
LOK49_LG09G00428	8	20073115	20073336	222	Hypothetical protein
LOK49_LG08G00640	9	43196289	43556954	360666	Cytochrome P450 82A3
LOK49_LG08G00641	9	44334649	44334855	207	Hypothetical protein
LOK49_LG15G01683	11	66860710	66861069	360	Ripening-related protein grip22
LOK49_LG15G01684	11	67087607	67096672	9066	Hypothetical protein
LOK49_LG11G01846	12	133880232	133881416	1185	UDP-glycosyltransferase 91A1
LOK49_LG11G01847	12	134127125	134134502	7378	Mitochondrial carrier protein CoAc1
LOK49_LG14G02272	15	132091777	132119069	27293	Helicase SEN1
LOK49_LG14G02273	15	132378538	132382958	4421	Hypothetical protein

contribute to the dramatic reduction in N_e. Wild *C. oleifera* in the warm winter areas are mostly distributed in the southern distribution part of wild *C. oleifera* in China (Fig. 1a and Table S1), which might come from the southern refugia during LG, with more suitable climate and weaker selective pressure compared to the northern refugia. The difference in selective pressure from climate change is an important prerequisite for adaptive evolution (Hu et al., 2023). For wild *C. oleifera*, extreme cold climate played an important role in driving adaptive evolution.

4.2. Key SNPs in cis-regulatory elements may affect the freezing tolerance of wild Camellia oleifera

In general, polyploids have more abundant genetic variations compared to their diploid relative species (Sattler et al., 2016). Xie et al. (2024) conducted genome survey of hexaploid wild *Camellia oleifera* from the Lu Mountain and found that the repeat content in the genome of hexaploid wild *C. oleifera* was similar to those of the diploid relative species while the heterozygosity in the genome of hexaploid wild *C. oleifera* was more than three times of that in the diploid relatives. Our study also found high genetic diversity in genomes of wild *C. oleifera* among populations with a total of 271,704,193 SNPs identified, mainly located in the intergenic regions. Both the results of IBD and IBE analyses proved that the 155 adaptive SNPs identified by both LFMM and RDA were closely related to the three key bioclimatic variables (Fig. 4a and b). Most of the adaptive SNPs were located in the intergenic regions, and 20 key SNPs with genotype differentiation exceeding 50% between wild *C. oleifera* in the cold winter areas and the warm winter areas were all located in the intergenic regions (Fig. 4c and Table S7), indicating that the evolution of environmental adaptation may be mainly based on the changes in regulatory regions rather than changes in protein coding sequences. Previous studies have mentioned that DNA sequence polymorphisms in the non-coding regions of the genome can endow plants with the ability to adapt to adverse environmental conditions, with particular emphasis on the crucial role of CREs (Schmitz et al., 2021; Gullotta et al., 2022).

We found that SNP_9_44085405 in the 20 key SNPs was significantly correlated with the key bioclimatic variable BIO6 and located within a CRE involved in SA responsiveness (Fig. 5c), and the SA signaling pathway is regulated by diverse abiotic stresses, such as salt, drought, cold, and heat (Jia et al., 2023). Considering the position of this CRE in the genome (Fig. 5b), it may be an enhancer activated by abiotic stress to enhance the expression of nearby genes. *LOK49_LG08G00640* is near this CRE, this gene encodes “cytochrome P450 82A3” (Fig. 5b and Table 1). The cytochrome P450 (CYP) represents a large and important enzyme superfamily in plants, widely involved in the biosynthesis of metabolites such as fatty acids, plant hormones, and phenylpropanoids (Zhao et al., 2023). The CYP82 family genes exist specifically in dicots and are usually closely related to the responses to environmental stresses. For instance, *GmCYP82A3* is from soybean (*Glycine max* L.) CYP82 family, and the results of transgenic experiment indicated that the overexpression of this gene could improve the tolerance to various biotic and abiotic stresses (Yan et al., 2016). *LOK49_LG08G00640* may have similar functions and the gene expression could be enhanced by this CRE. According to the genotypes of SNP_9_44085405, in wild *C. oleifera* in the cold winter areas, this CRE was normal and could enhance the expression of *LOK49_LG08G00640* under freezing stress, and high expression level of this gene could improve freezing tolerance of wild *Camellia oleifera* to survive in adverse environment (Figs. 4c and 5c). On the other hand, wild *C. oleifera* in the warm winter areas did not need to expend much energy on low temperature tolerance due to the suitable environment, and mutations in this CRE can be preserved, which might lead to the abnormal function of this CRE, resulting in low expression level of *LOK49_LG08G00640*. The expression pattern of the potential transcripts (*isoform_222481* and *isoform_284377*) of *LOK49_LG08G00640* under freezing temperature supports the above hypothesis (Fig. 5a).

4.3. Selected CNVRs in hexaploid wild Camellia oleifera may contribute to freezing tolerance

Polyploidy is a major driver in evolution, and specific mutations are more likely to be preserved in polyploids under the strong selective pressure caused by extreme cold climate (Heslop-Harrison et al., 2022). Due to hybridization and whole-genome duplication events, copy number variations are particularly common in allopolyploid plants, such as allopolyploid wheat (Schiessl et al., 2019). As an allohexaploid, wild *C. oleifera* may originate from hybridization and polyploidization during the Quaternary under extreme cold stress (Qin et al., 2023). In this study, 1175 selected CNVRs were found between the wild *C. oleifera* in the cold winter areas and in the warm winter areas. Interestingly, we found that a selected CNVR, CNVR_9_43980001_44136800, overlapped with the 5 key SNPs near *LOK49_LG08G00640* closely related to freezing tolerance, especially SNP_9_44085405 (Fig. 5b), indicating interactions might occur between CNVRs and SNPs. Genomic region copies containing specific genotypes of key adaptive SNPs may be selected for or against in adaptive evolution, which may also contribute to subgenomic dominance in allopolyploids (Cheng et al., 2018). On the other hand, CNVRs may scale up the effects of key SNPs, and therefore dramatically affect the expression of key genes associated with freezing tolerance.

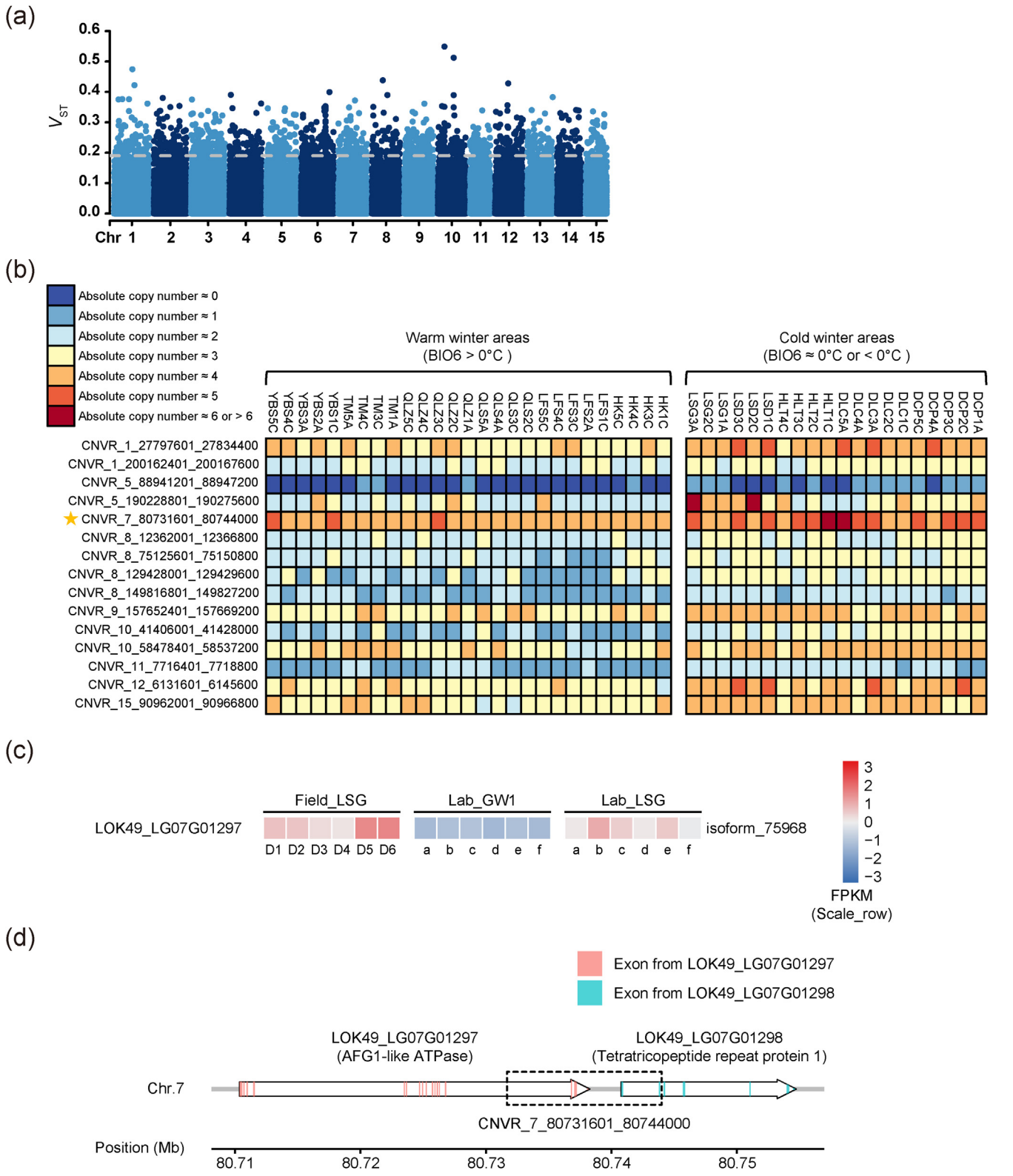


Fig. 6. Multiple analyses of CNVRs with silhouette score >0.5. (a) Manhattan plot of V_{ST} between the wild *Camellia oleifera* in the cold winter areas and the warm winter areas, the gray horizontal dashed line represents the top 1% V_{ST}. (b) Genotypic heatmap of 15 key CNVRs, the yellow five-pointed star indicates the related gene of this CNVR successfully matched with key unigene associated with freezing tolerance in the transcriptome data from Xie et al. (2023). (c) Heatmap of expression levels of unigene corresponding to the related gene *LOK49_LG07G01297* of the key CNVRs. (d) The position of the key CNVR in *LOK49_LG07G01297* in the genome.

Table 2
Basic information of 23 related genes of 15 key CNVRs with genotype differentiation exceeding 50% between wild *Camellia oleifera* in the cold winter areas and the warm winter areas. GENE: ID of CNVRs related genes in the genome, CHR: the chromosome where the genes are located in the genome, START: the starting position of genes in the genome, END: the ending position of genes in the genome, LEN: the length of genes, ANN: annotation results of genes in the genome.

GENE	CHR	START	END	LEN	ANN
LOK49_LG01G00549	1	27805891	27816906	11016	Serine carboxypeptidase-like 27
LOK49_LG01G00550	1	27822077	27824980	2904	Serine carboxypeptidase-like 27
LOK49_LG01G03455	1	200163508	200164658	1151	Hypothetical protein
LOK49_LG01G03456	1	200166642	200167145	504	Hypothetical protein
LOK49_LG06G01678	5	88943335	88967826	24492	Exocyst complex component SEC3A
LOK49_LG06G03458	5	190256946	190259219	2274	Hypothetical protein
LOK49_LG07G01297	7	80710360	80738286	27927	AFG1-like ATPase
LOK49_LG07G01298	7	80740755	80754736	13982	Tetratricopeptide repeat protein 1
LOK49_LG09G00305	8	12359461	12370309	10849	Hypothetical protein
LOK49_LG09G01101	8	75054085	75126801	72717	Ankyrin repeat-containing protein ITN1
LOK49_LG09G01102	8	75128002	75128695	694	Conserved oligomeric Golgi complex subunit 2
LOK49_LG09G01103	8	75133830	75138045	4216	Hypothetical protein
LOK49_LG09G01104	8	75139193	75139840	648	Hypothetical protein
LOK49_LG09G01845	8	129427390	129429102	1713	Hypothetical protein
LOK49_LG09G02306	8	149822086	149822736	651	Hypothetical protein
LOK49_LG08G02591	9	157648156	157653375	5220	Nuclear transport factor 2
LOK49_LG08G02592	9	157655756	157665503	9748	Hypothetical protein
LOK49_LG10G00813	10	41424921	41425545	625	Protein DNA-DAMAGE INDUCIBLE 1
LOK49_LG10G01036	10	58509733	58517875	8143	Peroxisomal membrane protein 13
LOK49_LG15G00085	11	7714855	7717682	2828	Hypothetical protein
LOK49_LG11G00126	12	6134656	6206845	72190	ADP-ribosylation factor GTPase-activating protein AGD4
LOK49_LG14G01405	15	90960095	90965352	5258	4-coumarate–CoA ligase 1
LOK49_LG14G01406	15	90965710	90966503	794	Hypothetical protein

In the 1175 selected CNVRs, 297 CNVRs were located in the exon regions (Fig. S5 and Table S8). The genotypes of the CNVRs were divided into six types according to absolute copy number, and 15 key CNVRs were found with genotype differentiation exceeding 50% between wild *C. oleifera* in the cold winter areas and the warm winter areas (Fig. 6b and Table S8). *LOK49_LG07G01297* is one of the related genes of the 15 key CNVRs, encoding "AFG1-like ATPase" (Table 2). In previous studies, AFG1-like ATPase was identified as belonging to the extended superfamily of AAA+-ATPases, and it was found to be involved in regulating calcium signal transduction in *Arabidopsis thaliana* (Bussemer et al., 2009). As one of the most important secondary messengers in plants, calcium signal plays a key role in activating the responses to abiotic stresses (Dong et al., 2022). We found that CNVR_7_80731601_80744000 covered partial exon regions of *LOK49_LG07G01297* (Fig. 6d), and the absolute copy number of CNVR_7_80731601_80744000 in wild *C. oleifera* in the cold winter areas was more than that in wild *C. oleifera* in the warm winter areas (Fig. 6b). More absolute copy number of the exon regions of *LOK49_LG07G01297* may mean higher expression level of *LOK49_LG07G01297* under freezing stress, which may contribute to more efficient cold signal transduction of wild *C. oleifera* in the cold winter areas, accelerating the activation of relevant defense mechanisms. The expression pattern of the potential transcript (*isoform_75968*) of *LOK49_LG07G01297* under freezing temperature supported the hypothesis (Fig. 6c).

4.4. *The analysis results based on the diploid* Camellia lanceoleosa *genome are reliable*

For many economically important polyploid crops such as wheat, cotton and oilseed rape, the genomes of their diploid ancestors provided valuable bases for polyploid population genomic studies (Li and Liu, 2019). Based on homologous genes, Qin et al. (2023) found that *Camellia lanceoleosa* and *C. oleifera* are closely related species, and *C. lanceoleosa* may be homologous to one of the diploid ancestors of *C. oleifera*. Therefore, since the hexaploid *C. oleifera* genome was not available during the study, we conducted the analyses based on the diploid *C. lanceoleosa* genome. Finally, we do succeed in finding the related genes of key SNPs and CNVRs, which may be closely related to freezing tolerance in wild *C. oleifera*. In addition, some of the genes were found actively expressed under freezing temperature in wild *C. oleifera* with strong freezing tolerance (Figs. 5a and 6c). Meanwhile, Zhu et al. (2024) have reported the chromosome-scale genome of hexaploid cultivated *C. oleifera*. Their results of collinearity analysis showed that the diploid *C. lanceoleosa* genome was highly homologous with the three homologous chromosome groups of hexaploid *C. oleifera* genome with high gene coverage (Zhu et al., 2024). Such results demonstrate that the diploid *C. lanceoleosa* genome can be used as reference genome for population genomic analyses in hexaploid *C. oleifera*. Moreover, according to the BLAST results with the full-length cDNA sequences, we discovered that most of the related genes of key SNPs and CNVRs based on the diploid *C. lanceoleosa* genome did exist in the hexaploid *C. oleifera* genome with high similarity (Table S9). In sum, all the results strongly support that our analysis results based on the diploid *C. lanceoleosa* genome are reliable.

In the future, based on the hexaploid *Camellia oleifera* genome, the regulatory mechanisms of key variation sites related to freezing tolerance and the functions of genes associated with these key sites can be further revealed and validated. Developing molecular markers based on these key variation sites may contribute to the efficient exploration and utilization of genetic resources of wild *C. oleifera* with strong freezing tolerance. These valuable genetic resources of wild *C. oleifera* may promote the breeding of *C. oleifera* through hybridization or as rootstocks.

5. Conclusion

In the present study, we performed genome sequencing of 47 wild *Camellia oleifera* from 7 natural distribution sites in the northern distribution edge and 4 natural distribution sites in the southern distribution region of wild *C. oleifera* in China. Based on a large number of SNPs and 19 bioclimatic variables, the analyses of population genetics and landscape genomics indicated that wild *C. oleifera* populations could be divided into two groups: in the cold

winter areas and in the warm winter areas, and they may have experienced different selection pressures because of the differences in winter temperature. On the other hand, a set of climate-associated SNPs were identified and some selected CNVRs between the two groups of wild *C. oleifera* were revealed for the first time. Some key SNPs may regulate the expression of key gene associated with freezing tolerance by affecting the function of cis-regulatory elements in the genome, and these SNPs were also found located within a CNVR, suggesting interactions may occur between SNPs and CNVRs. Some key CNVRs located in the exon regions of the genome were found closely related to the expression of genes associated with cold signal transduction. This study demonstrated the important roles of SNPs and CNVRs in the adaptive evolution to freezing stress in wild *C. oleifera*. Allopolyploids may contain rich SNPs and CNVRs, and therefore they may have great potential for adaptive evolution. This study provides a case for understanding the molecular bases of adaptive evolution to freezing stress in polyploid trees.

Data accessibility statement

The data that support the findings in this research are deposited in the short read archive (SRA) databank (https://www.ncbi.nlm.nih.gov/sra) with the accession number PRJNA1101446.

CRediT authorship contribution statement

Haoxing Xie: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kaifeng Xing:** Writing – review & editing, Writing – original draft, Software, Resources, Formal analysis. **Jun Zhou:** Writing – review & editing, Writing – original draft, Resources, Formal analysis. **Yao Zhao:** Writing – review & editing, Writing – original draft, Software, Formal analysis, Conceptualization. **Jian Zhang:** Writing – review & editing, Writing – original draft, Formal analysis. **Jun Rong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant no. 32270238 and 31870311).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pld.2024.07.009.

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664.
Altschul, S.F., Gish, W., Miller, W., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Aslam, M., Fakher, B., Ashraf, M.A., et al., 2022. Plant low-temperature stress: signaling and response. Agronomy 12, 702.
Behr, A.A., Liu, K.Z., Liu-Fang, G., et al., 2016. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics 32, 2817–2823.
Bussemer, J., Chigri, F., Vothknecht, U.C., 2009. *Arabidopsis* ATPase family gene 1-like protein 1 is a calmodulin-binding AAA+-ATPase with a dual localization in chloroplasts and mitochondria. FEBS J. 276, 3870–3880.
Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., et al., 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. 38, 5825–5829.
Chen, H., Patterson, N., Reich, D., 2010. Population differentiation as a test for selective sweeps. Genome Res. 20, 393–402.
Chen, Y.X., Chen, Y.S., Shi, C.M., et al., 2017. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience 7, 1–6.
Cheng, F., Wu, J., Cai, X., et al., 2018. Gene retention, fractionation and subgenome differences in polyploid plants. Nat. Plants 4, 258–268.
Cui, X., Wang, W., Yang, X., et al., 2016. Potential distribution of wild *Camellia oleifera* based on ecological niche modeling. Biodivers. Sci. 24, 1117–1128.
Cui, X.Y., Li, C.H., Qin, S.Y., et al., 2022. High-throughput sequencing-based microsatellite genotyping for polyploids to resolve allele dosage uncertainty and improve analyses of genetic diversity, structure and differentiation: a case study of the hexaploid *Camellia oleifera*. Mol. Ecol. Resour. 22, 199–211.
Danecek, P., Auton, A., Abecasis, G., et al., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.
Danecek, P., Bonfield, J.K., Liddle, J., et al., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008.
Dong, S.-S., He, W.-M., Ji, J.-J., et al., 2020. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinform. 22, bbaa227.
Dong, Q., Wallrad, L., Almutairi, B.O., et al., 2022. Ca^{2+} signaling in plant responses to abiotic stresses. J. Integr. Plant Biol. 64, 287–300.
Felsenstein, J., 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics Int J Willi Hennig Soc. 5, 164–166.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315.
Forester, B.R., Lasky, J.R., Wagner, H.H., et al., 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Mol. Ecol. 27, 2215–2233.
Frichot, E., Schoville, S.D., Bouchard, G., et al., 2013. Testing for associations between Loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. 30, 1687–1699.
Gao, L., Jin, L., Liu, Q., et al., 2024. Recent advances in the extraction, composition analysis and bioactivity of Camellia (*Camellia oleifera* Abel.) oil. Trends Food Sci. Technol. 143, 104211.
Gong, W., Xiao, S., Wang, L., et al., 2022. Chromosome-level genome of *Camellia lanceoleosa* provides a valuable resource for understanding genome evolution and self-incompatibility. Plant J. 110, 881–898.
Gullotta, G., Korte, A., Marquardt, S., 2022. Functional variation in the non-coding genome: molecular implications for food security. J. Exp. Bot. 74, 2338–2351.
Heslop-Harrison, J.S., Schwarzacher, T., Liu, Q., 2022. Polyploidy: its consequences and enabling role in plant diversification and evolution. Ann. Bot. 131, 1–10.
Hu, Y., Wang, X., Xu, Y., et al., 2023. Molecular mechanisms of adaptive evolution in wild animals and plants. Sci. China Life Sci. 66, 453–495.
Jia, X., Wang, L., Zhao, H., et al., 2023. The origin and evolution of salicylic acid signaling and biosynthesis in plants. Mol. Plant 16, 245–259.
Kanehisa, M., Goto, S., Kawashima, S., et al., 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32, D277–D280.
Ke, C., 2019. Main nutrient composition and health function of tea oil. Modern Food 13, 105–108.
Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15, 356.
Lescot, M., Déhais, P., Thijs, G., et al., 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 30, 325–327.
Letunic, I., Bork, P., 2021. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296.
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Genomics.
Li, L.-F., Liu, B., 2019. Recent advances of plant polyploidy and polyploid genome evolution. Sci. Sin. Vitae 49, 327–337.
Liu, J.Y., Shi, Y.T., Yang, S.H., 2018. Insights into the regulation of C-repeat binding factors in plant cold signaling. J. Integr. Plant Biol. 60, 780–795.
Liu, X., Fu, Y.-X., 2020. Stairway plot 2: demographic history inference with folded SNP frequency spectra. Genome Biol. 21, 280.
Lye, Z.N., Purugganan, M.D., 2019. Copy number variation in domestication. Trends Plant Sci. 24, 352–365.
Ming, T.L., 2000. Monograph of the Genus *Camellia*. Yunnan Science and Technology Press.
Ni, J., Yu, G., Harrison, S.P., et al., 2010. Palaeovegetation in China during the late quaternary: biome reconstructions based on a global scheme of plant functional types. Palaeogeogr. Palaeoclimatol. Palaeoecol. 289, 44–61.
Pertea, G., Pertea, M., 2020. GFF utilities: GffRead and GffCompare. F1000Research 9, 304.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.
Purcell, S., Neale, B., Todd-Brown, K., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Qin, S.Y., Chen, K., Zhang, W.J., et al., 2023. Phylogenomic insights into the reticulate evolution of *Camellia* sect. *Paracamellia* Sealy (Theaceae). J. Syst. Evol. 62, 38–54.
Redon, R., Ishikawa, S., Fitch, K.R., et al., 2006. Global variation in copy number in the human genome. Nature 444, 444–454.
Rellstab, C., Gugerli, F., Eckert, A.J., et al., 2015. A practical guide to environmental association analysis in landscape genomics. Mol. Ecol. 24, 4348–4370.
Sang, Y., Long, Z., Dan, X., et al., 2022. Genomic insights into local adaptation and future climate-induced vulnerability of a keystone forest tree in East Asia. Nat. Commun. 13, 6541.
Sattler, M.C., Carvalho, C.R., Clarindo, W.R., 2016. The polyploidy and its key role in plant breeding. Planta 243, 281–296.
Schiessl, S.-V., Katche, E., Ihien, E., et al., 2019. The role of genomic structural variation in the genetic improvement of polyploid crops. Crop J. 7, 127–140.
Schmitz, R.J., Grotewold, E., Stam, M., 2021. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell 34, 718–741.
Sork, V.L., Aitken, S.N., Dyer, R.J., et al., 2013. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genet. Genomes 9, 901–911.
Sork, V.L., Waits, L., 2010. Contributions of landscape genetics – approaches, insights, and future potential. Mol. Ecol. 19, 3489–3495.
Strimbeck, G.R., Schaberg, P.G., Fossdal, C.G., et al., 2015. Extreme low temperature tolerance in woody plants. Front. Plant Sci. 6, 884.
Tang, D., Chen, M., Huang, X., et al., 2023. SRplot: a free online platform for data visualization and graphing. PLoS One 18, e0294236.
Tarasov, A., Vilella, A.J., Cuppen, E., et al., 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034.
Terhorst, J., Kamm, J.A., Song, Y.S., 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. 49, 303–309.
Van der Auwera, G.A., Carneiro, M.O., Hartl, C., et al., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinform. 43, 11.10.11–11.10.33.
Wang, J., Hu, Z., Liao, X., et al., 2022. Whole-genome resequencing reveals signature of local adaptation and divergence in wild soybean. Evol. Appl. 15, 1820–1833.
Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164.
Wang, X., Zheng, Z., Cai, Y., et al., 2017. CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. GigaScience 6, 1–12.
Wisniewski, M., Nassuth, A., Teulieres, C., et al., 2014. Genomics of cold hardiness in woody plants. Crit. Rev. Plant Sci. 33, 92–124.
Wright, S.I., Gaut, B.S., 2004. Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. Evol. 22, 506–519.
Xiang, X., Zhou, X., Zi, H., et al., 2023. *Populus cathayana* genome and population resequencing provide insights into its evolution and adaptation. Hortic. Res. 11, uhad255.
Xie, H., Xing, K., Zhang, J., et al., 2024. Genome survey and identification of key genes associated with freezing tolerance in genomic draft of hexaploid wild *Camellia oleifera*. J. Hortic. Sci. Biotechol. 99, 326–335.
Xie, H.X., Zhang, J., Cheng, J.Y., et al., 2023. Field plus lab experiments help identify freezing tolerance and associated genes in subtropical evergreen broadleaf trees: a case study of *Camellia oleifera*. Front. Plant Sci. 14, 1113125.
Yan, Q., Cui, X., Lin, S., et al., 2016. *GmCYP82A3*, a soybean cytochrome P450 family gene involved in the Jasmonic acid and ethylene signaling pathway, enhances plant resistance to biotic and abiotic stresses. PLoS One 11, e0162253.
Yang, J., Lee, S.H., Goddard, M.E., et al., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82.
Ye, J., Zhang, Y., Wang, X., 2017. Phylogeographic history of broad-leaved forest plants in subtropical China. Acta Ecol. Sin. 37, 5894–5904.
Yin, L., Zhang, H., Tang, Z., et al., 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Dev. Reprod. Biol. 19, 619–628.
Yu, G., Chen, X., Ni, J., et al., 2000. Palaeovegetation of China: a pollen data-based synthesis for the mid-Holocene and last glacial maximum. J. Biogeogr. 27, 635–664.
Zhang, C., Dong, S.-S., Xu, J.-Y., et al., 2018. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35, 1786–1788.
Zhang, H., Zhu, J., Gong, Z., et al., 2022. Abiotic stress responses in plants. Nat. Rev. Genet. 23, 104–119.
Zhao, S., Gibbons, J.G., 2018. A population genomic characterization of copy number variation in the opportunistic fungal pathogen *Aspergillus fumigatus*. PLoS One 13, e0201611.
Zhao, X., Zhao, Y., Gou, M., et al., 2023. Tissue-preferential recruitment of electron transfer chains for cytochrome P450-catalyzed phenolic biosynthesis. Sci. Adv. 9, eade4389.
Zhu, H., Wang, F., Xu, Z., et al., 2024. The complex hexaploid oil-Camellia genome traces back its phylogenomic history and multi-omics analysis of Camellia oil biosynthesis. Plant Biotechnol. J. 22, 2890–2906. https://doi.org/10.1111/pbi.14412.