Plant Diversity 46 (2024) 169–180

Contents lists available at ScienceDirect

Plant Diversity

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

Research paper

Cryptic divergences and repeated hybridizations within the endangered "living fossil" dove tree (*Davidia involucrata*) revealed by whole genome resequencing

Yumeng Ren [a], Lushui Zhang [a], Xuchen Yang [a, c], Hao Lin [a], Yupeng Sang [a], Landi Feng [a], Jianquan Liu [a, b, **], Minghui Kang [a, b, *]

[a] Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China
[b] State Key Laboratory of Herbage Improvement and Grassland Agro-ecosystem, College of Ecology, Lanzhou University, Lanzhou 730000, China
[c] Guangdong Provincial Key Laboratory of Plant Adaptation and Molecular Design, Guangzhou Key Laboratory of Crop Gene Editing, Innovative Center of Molecular Genetics and Evolution, School of Life Sciences, Guangzhou University, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 2 November 2023
Received in revised form
5 February 2024
Accepted 8 February 2024
Available online 12 February 2024

Keywords:
Davidia involucrata
Cryptic lineage
Hybridization
Population genomics
Positive evolution

ABSTRACT

The identification and understanding of cryptic intraspecific evolutionary units (lineages) are crucial for planning effective conservation strategies aimed at preserving genetic diversity in endangered species. However, the factors driving the evolution and maintenance of these intraspecific lineages in most endangered species remain poorly understood. In this study, we conducted resequencing of 77 individuals from 22 natural populations of *Davidia involucrata*, a "living fossil" dove tree endemic to central and southwest China. Our analysis revealed the presence of three distinct local lineages within this endangered species, which emerged approximately 3.09 and 0.32 million years ago. These divergence events align well with the geographic and climatic oscillations that occurred across the distributional range. Additionally, we observed frequent hybridization events between the three lineages, resulting in the formation of hybrid populations in their adjacent as well as disjunct regions. These hybridizations likely arose from climate-driven population expansion and/or long-distance gene flow. Furthermore, we identified numerous environment-correlated gene variants across the total and many other genes that exhibited signals of positive evolution during the maintenance of two major local lineages. Our findings shed light on the highly dynamic evolution underlying the remarkably similar phenotype of this endangered species. Importantly, these results not only provide guidance for the development of conservation plans but also enhance our understanding of evolutionary past for this and other endangered species with similar histories.

1. Introduction

The question of how biodiversity evolves is a fundamental inquiry in the fields of evolutionary and conservation biology (Schluter, 2000; Seehausen, 2009). A crucial prerequisite for understanding the evolution of biodiversity in natural populations is the accurate definition of evolutionarily distinct and conservation units (Crandall et al., 2000). This is essential for preserving biological diversity, as it provides managers and policymakers with a clear understanding of the boundaries of population units for endangered species (Funk et al., 2012). Understanding the genetic diversity of endangered species holds significant importance in conservation biology. Despite their morphologically similar appearance, cryptic lineages may have evolved within a single endangered species, and each lineage deserves conservation efforts due to their distinct adaptations to different environments (Bolnick et al., 2003; Bickford et al., 2007; Palsbøll et al., 2007). Neglecting

* Corresponding author. Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, 610065, China.
** Corresponding author. Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, 610065, China.
E-mail addresses: liujq@nwipb.ac.cn (J. Liu), kangminghui0106@gmail.com (M. Kang).
Peer review under responsibility of Editorial Office of Plant Diversity.

https://doi.org/10.1016/j.pld.2024.02.004

such lineages can introduce bias into predictions regarding their responses to environmental change and biodiversity loss (Scheffers et al., 2012; Pauls et al., 2013; Adams et al., 2014; Feckler et al., 2014). Therefore, accurately determining conservation units can prevent both the underestimation of the necessary protection units for endangered species and the misallocation of resources on abundant species (Frankham et al., 2010). The recognition of cryptic lineages has become increasingly apparent due to advancements in molecular techniques (Moritz, 1994; Petit et al., 1998; Pfenninger and Schwenk, 2007; Wang et al., 2009; Jörger and Schrödl, 2013; Shang et al., 2015; Torres-Cambas et al., 2017; Feng et al., 2019). Beyond their implications for conservation, the presence of such lineages and hybridizations also challenges our understanding of population connectivity (Pante et al., 2015), speciation (Seehausen, 2009) and even ecosystem functioning (Brodersen and Seehausen, 2014).

The deep intraspecific lineages were found to have emerged as a result of both geographical and climatic changes in various species (Wang et al., 2009, 2015; Li et al., 2021; Hu et al., 2022; Sang et al., 2022; Shen et al., 2022; Wu et al., 2023). For instance, the Quaternary glaciation and the accompanying cooling climates drove intraspecific divergences in different refugia for plants and animals alike (Hewitt, 2000; Qiu et al., 2011). However, postglacial range expansions further resulted in the meeting of the diverged lineages and subsequent hybridizations (Zachos et al., 2001; Petit et al., 2003; Hewitt, 2004). Additionally, with each deep intraspecific lineage, long-term selection may favor the most suitable genotypes that adapt to local habitats, which can be detected through population genomic data (Zhang et al., 2020; Hu et al., 2022). For the ancient plants known as “fossil” relics, which originated early and left behind ancient fossils, it is likely that past geographic and climatic fluctuations played a significant role in driving both interspecific and intraspecific divergences (Mao and Liu, 2012). A notable example can be found within the fossil genus *Cercidiphyllum*, where two lineages, representing two distinct species, diverged approximately 5 million years ago (Mya) during the Pliocene epoch. This period coincided with significant geographic and climatic changes in eastern Asia (Qi et al., 2012). Moreover, it is probable that subsequent demographic fluctuations and hybridization events occurred between these lineages in response to the later climatic changes (Zhu et al., 2020). These events likely led to the introgression of both chloroplast and nuclear DNA (Qi et al., 2012). Apart from the genus *Cercidiphyllum*, many more fossil genera are found in this region, such as *Ginkgo* (Zhao et al., 2019), *Euptelea* (Cao et al., 2020), *Tetracentron* (Liu et al., 2020), and *Davidia* (Tang et al., 2017). Understanding how past geographic and climatic fluctuations have influenced the divergence and hybridization of these fossil genera over their extended presence in this region is of great interest. These evolutionary events have the potential to complicate conservation efforts and challenge our understanding of the formation of hidden lineages (Allendorf et al., 2010; Naciri and Linder, 2015). Additionally, it is crucial to identify the key genes that may have played a significant role in the local adaptation of these distinct lineages in the genomic era (Hu et al., 2022). In general, regions of the genome that have experienced positive or adaptive selection exhibit significantly greater genetic differentiation compared to other regions that have evolved neutrally. Therefore, we can identify selection signals by examining the “outliers” of the F_{ST} values, which serve as an index of genetic differentiation between populations (Zhao et al., 2019; Zhu et al., 2020). Gene-environment association analyses address the issue of environmental heterogeneity across different populations and aim to establish relationships between patterns of allele frequencies and environmental gradients (Sang et al., 2022; Shen et al., 2022). By employing these two approaches, we can gain a better understanding of the process and potential of adaptation.

The only species of the genus *Davidia*, *Davidia involucrata*, is found exclusively in southwestern and central China in eastern Asia (He et al., 2004; Yang et al., 2019). It thrives in humid, rainy, and foggy evergreen and deciduous mixed forests but is sensitive to drought stress (Zhang et al., 2000; Liu et al., 2019). Detailed reports on the geographic distribution and eco-physiological characteristics of this species suggest that it has developed local adaptations (Tang et al., 2017; Yang et al., 2020). Its unique white bracts, which resemble doves, have earned it the name “dove tree” and contribute to its high ornamental value (He et al., 2004). This species is considered as a Cenozoic relict plant endemic to China (Tang et al., 2017) and is often referred to as the “giant panda” of the plant world. While it was once widely distributed in North America and East Asia, the Quaternary glaciation event caused a significant decline in its population size (Eyde, 1997; Manchester et al., 2009). The relic rarity and limited distribution of natural populations have led to the classification of *D. involucrata* as a nationally protected first-class wild plant in China (Liu et al., 2019). The fruits of *D. involucrata* are dispersed by animals, which limits their long-distance dispersal (Zhang et al., 2000). The seeds of this species experience a dormancy period of 2–3 years, with a very low germination rate (Song and Bao, 2006). *D. involucrata* is an entomophilous species, likely pollinated by bees and other insects (Sun et al., 2008). Extensive studies have reported high genetic diversity and inter-population differentiation in this endangered species using molecular methods and genetic analysis methods (Song and Bao, 2006; Luo et al., 2011; Li et al., 2012; Chen et al., 2015; Ma et al., 2015). Two major cryptic lineages have been identified: one in central China and the other in southwest China (Chen et al., 2015; Ma et al., 2015). However, one population in southwest China had a chloroplast haplotype closely related to the eastern lineage (Chen et al., 2015). These two lineages were estimated to have diverged 4–5 Mya (Chen et al., 2015; Ma et al., 2015). Nuclear SSR markers revealed multiple inter-lineage introgressions due to long-distance gene flow (Ma et al., 2015).

In this study, we conduct whole-genome resequencing on 77 individuals of *Davidia involucrata*, including 67 newly sequenced individuals, from its known natural distribution range in China. Our objective is to examine cryptic divergence and hybridization within this relic “fossil” plant. We identify one more cryptic lineage in addition to two major lineages previously identified. We also investigate whether certain genes exhibited positive signals during local adaptation and assessed the inbreeding levels and mutation load of these lineages and populations. Through these analyses, our aim is to evaluate evolutionary units, providing valuable insights for the conservation and management of this species.

2. Materials and methods

2.1. Plant sampling and genome resequencing

In addition to 10 previously sequenced genomes (Chen et al., 2020), we added whole-genome resequencing data of 67 individuals representing 22 natural populations within the known distribution range of dove-tree in central and southwest China. For each population, individuals were sampled at least 100 m apart from each other. Silica gel-dried leaves of 67 individuals were collected for DNA extraction in this study (Table S1). All of the samples were sequenced using DNBSEQ-T7 platforms with a sequencing depth of 20× .

2.2. Data filtering and single nucleotide polymorphism (SNP) calling

Raw reads were filtered using Fastp v.0.20.0 (Chen et al., 2018) to remove adapters and low quality base. Reads with a "N" base number of 5 bp and reads containing more than 40% of mass value ≤ 20 bases were removed: cropping with a 4 bp sliding window with an average mass below 20. The quality-filtered reads were mapped to the *Davidia involucrata* reference genome (Chen et al., 2020) with BWA-MEM v.0.7.12 (Li and Durbin, 2009) and sorted with Samtools v.1.10 (Li et al., 2009). BaseRecalibrator and ApplyBQSR in GATK v.4.1.1.7 (McKenna et al., 2010) were used to calculate the distribution of systematic error and correct base quality according to the base quality value of BAM file. Finally, single nucleotide polymorphism (SNP) calling was performed using HaplotypeCaller, CombineGVCFs and GenotypeGVCFs of GATK to generate a VCF file with the parameter "all sites" (including non-variant sites). To reduce the false positive error, SNPs were further filtered with VCFtools v.0.1.13 (Danecek et al., 2011) as following steps: (1) Sites with a quality score lower than 30 and SNP sites within 5 bp near InDel were filtered out. (2) Remove SNPs with more than two alleles. (3) Sites with extremely high (less than one-third of the average depth) coverage and extremely low (more than threefold of the average depth) coverage were treated as missing sites. (4) Filter out sites with more than 20% percentage missing and minimum allele frequency (MAF) < 0.05.

2.3. Population structure and phylogenetic analysis

Phylogenetic and population structure analyses require the use of SNPs with low selection pressure and a true response to natural variation in order to obtain an accurate topology of the phylogenetic tree. We used VCFtools to convert the data to a binary file as an input file for PLINK v.1.07 (Purcell et al., 2007), linked loci was removed using the parameters: -indep-pairwise 50 5 0.2. We constructed a maximum likelihood (ML) phylogenetic tree with FastTree v.2 (Price et al., 2010) software in the use of 1,722,440 LD-pruned, evolution-information SNPs. We searched for single copy of lineal homologous gene of *Camptotheca acuminata* (Kang et al., 2021), *Nyssa yunnanensis* (Mu et al., 2020), *Nyssa sinensis* (Yang et al., 2019) and *Davidia involucrata* (Chen et al., 2020), with protein sequences, then OrthoFinder2 (Emms and Kelly, 2019) software was used to classify the gene families of the protein sequence, distinguish the orthologous and paralogous genes, and screen out the orthologous gene families with only one copy in each of the four species, which was regarded as the single-copy orthologous genes of *D. involucrata* for subsequent analysis. Python script was used to extract all SNPs within the Coding sequence (CDS) interval of these single copy orthologous genes from the SNP-merged VCF file based on the extracted single copy gene information. The GTR-GAMMA model of RAxML v.8.2.11 (Alexandros and Stamatakis, 2014) was used to establish the single copy of orthologous genes extracted from 4 species, and the bootstrap support was repeatedly estimated using 100 bootstraps. It is then visualized using iTOL (https://itol.embl.de, Letunic and Bork, 2007).

We used 1,722,440 independent SNPs among 77 samples for Principal component analysis (PCA) using PLINK v.2018 and population structure inferred using ADMIXTURE v.1.3.0 (Alexander and Lange, 2011) with K-value setting ranging from 1 to 5. For further insight into relationships among lineages, we performed identity-by-descent blocks analysis after missing genotype estimation using the algorithm from BEAGLE v.4.1 (Browning and Browning, 2007) with the following parameters: window = 100,000; overlap = 10,000; ibdtrim = 150; ibdlod = 15.

2.4. Genetic diversity

We eliminated the ancestry-mixed individuals inferred from the analysis of population structure. Nucleotide diversity (π), quantified population genetic differentiation (F_{ST}), and Tajima's D between each pair of groups were calculated using VCFtools v.0.1.17. All of these three values were calculated using 20 kb non-overlapping sliding window. The π was estimated after taking into account both polymorphic and monomorphic sites. We measured and compared patterns of linkage disequilibrium (LD) for each group using PopLDdecay v.3.4.1 (Zhang et al., 2019).

2.5. Demographic history inference and gene flow

We selected "pure" individuals in structure analysis (Q > 0.9999) with sequencing coverage (>15×) to estimate the demographic history of effective population size (N_e) over time using the pairwise sequential Markovian coalescent (PSMC) method (Li and Durbin, 2011). For each lineage, we selected 2 individuals for PSMC analyses with 100 bootstrap replicates. The mutation rate was set as 1.87×10^{-9} per base per year, and generation time was set as 20 years following Chen et al. (2020). At the same time, VCFtools v.0.1.13 software was used to eliminate deviation from haven-equilibrium (0.001) SNPs for downstream analyses. We used a perl script (vcf2maf.pl; https://github.com/wk8910/bio_tools/) to generate two-dimensional joint SFS (2D-SFS) with "pure" individuals. Further, we used FASTSIMCOAL2 (Excoffier et al., 2021) to infer the differentiation time and gene flow among the lineages of *Davidia involucrata*. To reduce the influence of natural selection, we only used SNPs in intergenic region for historical population dynamics simulations. We then designed and simulated 8 different evolutionary models based on their genetic structure (Fig. S5) and 10 different evolutionary models for different gene flow scenarios (Fig. S6) in FASTSIMCOAL2. For each model, we performed 50 independent runs with 100,000 coalescence simulations per likelihood estimation and 40 cycles of the likelihood maximization algorithm to search the global ML parameter estimates. The best model was identified through the Akaike's information criterion (AIC). Identical to PSMC, we set a mutation rate of 3.74×10^{-8} per base per generation and a generation time of 20 years.

2.6. Identification of environment-associated SNPs

Two methods were used to identify genotype-environment association (GEA) loci across the whole-genome. We kept common SNPs with MAF >10%, including a total of 10,608,122 for analyses. Firstly, a univariate latent-factor linear mixed model (LFMM) in the R package LEA v.3.3.2 (Frichot and François, 2015) were implemented to search for associations between allele frequencies and the 19 BIOCLIM environmental variables (Fick and Hijmans, 2017). Three latent factors to account for population structure in the genotype data based on the number of ancestry clusters inferred with ADMIXTURE v.1.3.0 in LFMM analyses. For each environmental variable, five independent Markov chain Monte Carlo (MCMC) runs were conducted by using 5000 iterations as burn-in followed by 10,000 iterations. We used false discovery rate FDR correction of 5% for the significance cutoff. Secondly, we performed a multivariant redundancy analysis (RDA) to search for candidate loci with a low false-positive rate. We selected four uncorrelated environment variables (BIO2, BIO4, BIO17, BIO18) with a correlation < 0.7 before running RDA analyses using the R package *vegan* v.2.6–2 (Oksanen et al., 2017). The overlapped genetic loci by both approaches were regarded as "core adaptive loci" for local adaptation.

For functional annotation, *Davidia involucrata* predicted protein-coding genes were aligned to multiple public databases including

Swiss-Prot, NR, using NCBI BLAST + v.2.2.31 with an E-value of $1e^{-5}$ as the cutoff. The highest bit score was used to select the best homologous genes for comparison, so as to obtain the specific function of genes. InterProScan software was used to compare structural databases such as InterPro database and Pfam to predict the existing domains in the genome (Zdobnov and Apweiler, 2001). In addition, we then enriched the GO function of the core adaptive sites with R package *topGO* (Alexa and Rahnenfuhrer, 2022).

We used neutral (the 1,722,440 LD-pruned SNPs used for population structure analyses) and adaptive (the 22,630 core adaptive loci) variants to calculate correlation between genetic distance (F_{ST}/(1-F_{ST})) and geography (IBD; Mantel test) and environmental (IBE; partial Mantel test, after excluding geography influence), respectively. Geography and environmental distance (Euclidean distance) accounted by latitude and longitude of the samplings with significance determined using 999 permutations in the R package *vegan*.

2.7. Genes with positive selection signals

To identify genomic regions potentially under selection in *Davidia involucrata*, we calculated values of π ratio, F_{ST}, XP-CLR (Chen et al., 2010) between West lineages and East lineages (comprising genetic "pure" individuals in ADMIXTURE analyses) using a genome-wide sliding windows strategy (20 kb sliding windows and 0 kb steps) in VCFtools v.0.1.13. After calculating all tests, the windows with more than 100 SNPs and *P* values less than 0.025 (right side-Z test) were considered as significant outliers. The windows meet three conditions simultaneously were considered under significant selective pressure. To further investigate the annotation of core adaptive variants, we used the core adaptive variants identified in *D. involucrata* to run blastp v.2.10.0+ (Camacho et al., 2009) against the protein libraries of *Arabidopsis thaliana*. The best alignment for each gene was kept and considered to be homologous. Then, we searched for annotation of the homologs in *A. thaliana* in The Arabidopsis Information Resource (TAIR) database (https://www.arabidopsis.org) to obtain the annotations of the core adaptive variants.

2.8. Mutation load analyses

Whole genome heterozygosity (*He*) is the proportion of heterozygous sites in the whole genome that exclude gaps. For each sample, *He* was computed based on an in-house python script. The run of homozygosity (ROHs) was calculated using PLINK. We calculated the proportion of the genome (0–1) that is in runs of homozygosity (ROHs) (F_{ROH}) for each individual. To test whether these isolated populations experience inbreeding depression, we estimated the relative excess of derived loss-of-function (LoF) and missense variants in 22 populations. We employed methods previously used for gorillas (Xue et al., 2015) to measure the mutation load. SnpEFF v.4.3 (Cingolani et al., 2012) was used to create the gene function database based on the gene annotation file of the *D. involucrata* genome. Then, the high-quality SNPs were classified into different functional classes (synonymous, missense, loss of function) by SnpEFF with default settings. Finally, we calculated the total number of mutations of each class by counting each heterozygous genotype once and each homozygous alternative genotype twice.

We inferred the ancestral and derived allele at each location based on comparison to the genome of their close relatives *Camptotheca acuminata* and *Nyssa sinensis*. The probability of the derived versus ancestral allelic state were inferred using est-sfs v.2.03 (Keightley and Jackson, 2018). The ancestral allele was set as the non-deleterious allele, and the derived allele as the potential deleterious allele. We calculated the ratio of missense and loss-of-function to synonymous mutations at the homozygous and heterozygous sites.

3. Results

3.1. Population structure and genetic diversity

A total of 1.2 Tb of whole genome sequencing data were generated from 77 samples of 22 populations, with an average sequencing depth of 19.57× (Fig. 1A and Table S1). Using the chromosome-level *Davidia involucrata* reference genome (Chen et al., 2020), the average mapping rate of the raw reads was 97.12%, and the average genome coverage rate was 93.78% (Table S1). After filtering, 16,363,317 SNPs (MAF > 0.05) were retained for subsequent analyses.

Population genetic structure analysis by ADMIXTURE revealed that when K = 2, all individuals were divided into two distinct clusters. The western cluster comprised individuals from Gansu, Sichuan, Yunnan, and western Guizhou of southwest China, while the eastern cluster included individuals from eastern Guizhou, Hubei, Hunan, Shaanxi of central China. The individuals from the southern region of southwest China showed ancestral admixture (Fig. 1B). Approximately, 89.39% of the genetic composition of this third lineage was shared with the western lineage, while the remaining 10.61% was shared with the eastern lineage. Cross-validation error analysis suggested that the optimal classification was three genetic groups (K = 3) (Table S2). When K = 3, individuals of this southern lineage from western Guizhou were further separated from the remaining populations of the western lineage (Fig. 1C). We identified a wide hybrid zone (Hybrids1) between the western and southern lineages, as well as another hybrid zone (Hybrids2) between the southern and eastern lineage, both of which were located in the Sichuan Basin. Principal component analysis (PCA) exhibited a similar differentiation pattern among the three groups (Fig. 1C). The first principal component (PC1; explained variance = 11.53%) separated the western lineage from the eastern lineage, while the second principal component (PC2; explained variance = 5.46%) further divided them into five groups: the western lineage, the Hybrids1 group, the southern lineage, the Hybrids2 group, and the eastern lineage.

To construct phylogenetic tree, the maximum likelihood method was employed using 1,722,440 SNPs, with a *Nyssa sinensis* individual serving as an outgroup (Fig. S1A). A total of 5267 single-copy orthologous genes were identified based on the protein sequences of *N. yunnanensis*, *N. sinensis*, *Camptotheca acuminata* and *Davidia involucrata*. Similarly, maximum likelihood method was used to construct phylogenetic trees for single-copy orthologous genes (Fig. S1B). The phylogenetic trees constructed by SNPs and single-copy genes yielded consistent results (Fig. S1).

To assess patterns of genetic differentiation and genetic diversity across the genome, genetically admixed individuals were excluded, and parameters such as genetic diversity (π), LD patterns, and Tajima's D were calculated. The π values of each lineage ranged from 2.23×10^{-3} to 2.84×10^{-3}, with the eastern lineage exhibiting the highest diversity and the southern lineage showing the lowest diversity (Fig. S2). Pairwise genome-wide averages genetic divergence (F_{ST}) values among the three groups ranged from 0.255 to 0.422 (Table S4). The F_{ST} values between the eastern lineage and western or southern lineage were notably higher than that between the western and southern lineages (Table S4), indicating consistent genetic cluster patterns observed in ADMIXTURE, PCA, and NJ tree analyses. When assessing pairwise F_{ST} among populations, a high level of differentiation was observed among *D. involucrata* groups: the western lineage VS the eastern lineage (mean F_{ST} = 0.421), the southern lineage VS the eastern lineage

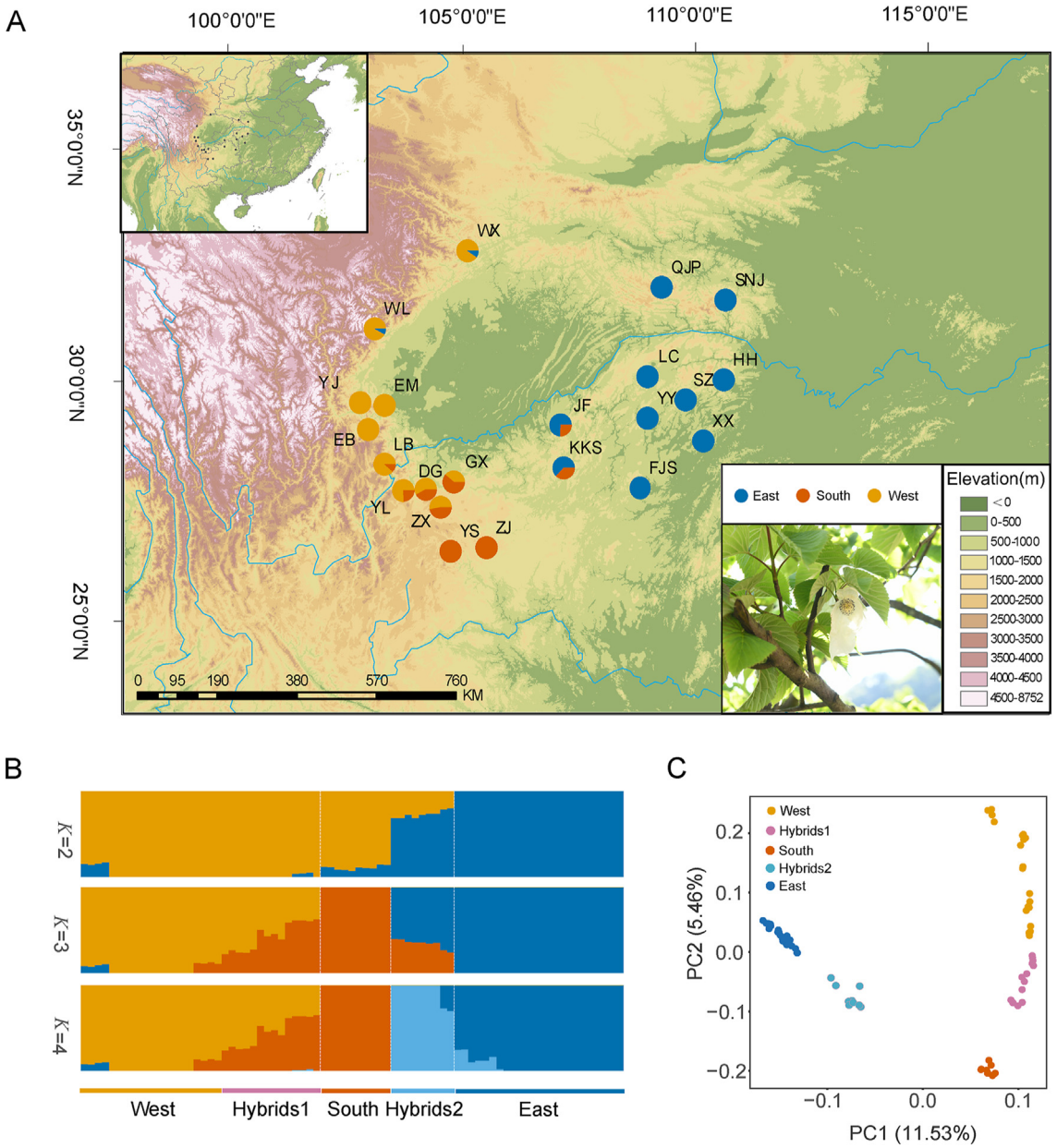

Fig. 1. Population genomic analyses of *Davidia involucrata*. (A) Sample locations of sampled *D. involucrata* populations. Charts at sampling locations indicate the distribution of genetic groups identified by ADMIXTURE analysis when K = 3. The map was retrieved from Google Earth (https://www.google.com/earth/). Elevation data for the map were derived from SRTM elevation data through the WorldClim data website (https://www.worldclim.org/). (B) Population structure bar plots. The scenarios of K = 2–4 was shown, and K = 3 is the best value according to cross-validation analysis. (C) Principal component analysis (PCA) plots showing the first two principal components.

(mean F_{ST} = 0.422). However, the mean pairwise F_{ST} between the western lineage and the southern lineage was significantly lower (0.255). The mean pairwise nucleotide differences in inter-lineage comparisons (*dxy*) ranged from 0.0112 to 0.0137 (Table S4), showing a same pattern. Tajima's D values varied from 0.5414 to 1.1786 (Fig. S3). Positive Tajima's D value may indicate a common historical demographic bottleneck in *D. involucrata*. The level of genetic diversity and LD decay revealed different demographic histories among the three lineages. The southern lineage exhibited more link imbalance, a slower LD decay rate (Fig. 2A), lower nucleotide diversity, and may have experienced a long-term population bottleneck or recent demographic expansions resulting in a decreased genetic diversity.

3.2. Demographic histories and gene flow

We used the pairwise sequentially Markovian coalescent (PSMC; Li and Durbin, 2011) to reconstruct the demographic history of each lineage. All three lineages exhibited similar demographic histories, characterized by a high effective population size (~1.2 Mya) followed by a bottleneck around 0.8 Mya (Fig. 2C). From the Pliocene to the early Pleistocene, all lineages experienced population expansion, reaching their maximum effective population sizes (Fig. 2C). The eastern lineage underwent three expansions and two contractions. The first expansion occurred in the late Miocene and early Pliocene, with the effective population size peaking at 38,000 around 2 Mya, followed by a decline (the first contraction), which

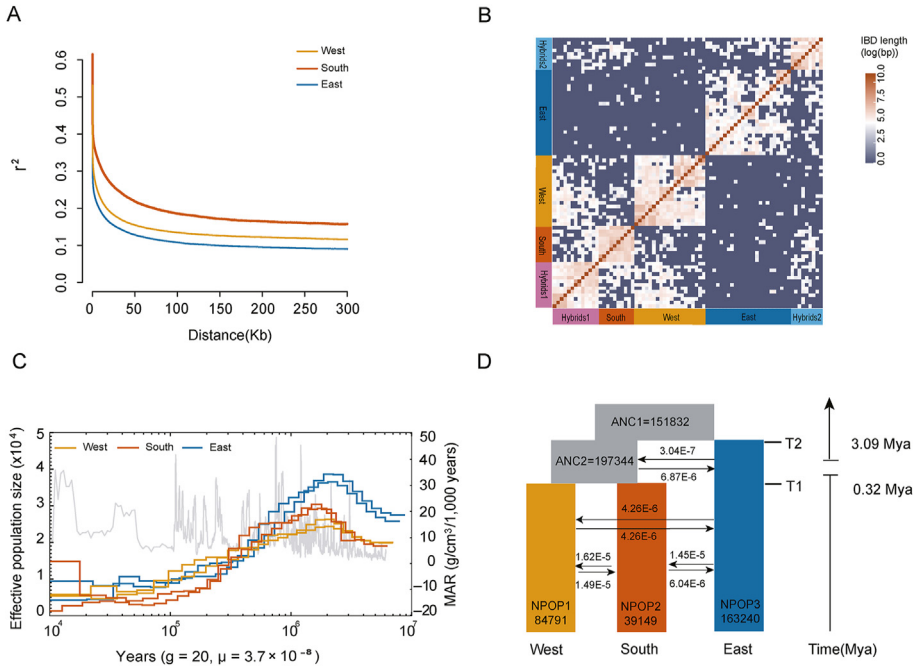

Fig. 2. Demographic history of *Davidia involucrata*. (A) Patterns of linkage disequilibrium for the three lineages. (B) Estimated haplotype sharing between individuals. Heatmap colors represent the total length of identity-by-descent blocks for each pairwise comparison. (C) Inferred demographic history of the West (yellow lines), South (orange lines), and East (blue lines) populations inferred by the pairwise sequentially Markovian coalescent (PSMC) method over the past 10 million years. (D) The best fitting model (model10) diagram and simulated parameters with maximum likelihood values are obtained using FASTSIMCOAL2 software, and estimated effective population size and differentiation time are given. The numbers next to the arrows represent mobility per generation between populations.

coincided with the increase in the mass accumulation rate (MAR) of loess in China. Subsequently, the population expanded from 0.07 Mya to 22 thousand years ago (Kya) (the second expansion) and experienced a contraction during 10–20 Kya (the second contraction). After the population expanded again around 10 Kya (the third expansion), the effective population size remained stable at 10,000. The western lineage experienced two expansions and one contraction, while the southern lineage experienced two expansions and two contractions. From 10 Mya to 1.1 Mya, the effective population size of the western lineages increased (the first expansion), reaching a maximum around 1.1 Mya (25,000 and 30,000, respectively), and then continuously declined (the first contraction). The effective population size of the western lineage declined continuously (the first contraction) from 1.1 Mya to 20 Kya, remained stable between 0.1 Mya and 0.3 Mya, increased (second expansion) thereafter, and remained constant at around 7000 during 20–10 Kya. The southern lineage reached its maximum effective population size of 30,000 around 1.05 Mya, followed by a decline from 1.05 Mya to 30 Kya (the first contraction), and then an increase during 30–10 Kya (the second expansion).

We employed a coalescent simulation-based method using fastsimcoal2 to estimate the timing of divergence and demographic histories of the three groups. We constructed ten models to represent the divergence of the three groups, and the best-fit model was determined based on the lowest AIC value and highest likelihood (Tables S7 and S8). The best-supported model indicated that the eastern lineage diverged from the common ancestor of the western and southern lineages during the Late Pliocene, approximately 3.09 Mya, while the divergence between the western and southern lineages occurred during the Middle Pleistocene, around 0.32 Mya (Fig. 2D). The simulations also revealed different rates of gene flow among the three lineages. Low levels of ancient gene flow were estimated, showing significant asymmetric gene flow between the common ancestor of the western and southern lineages and the eastern lineage ($M_{ANC2\leftarrow E} = 3.04 \times 10^{-7}$, $M_{E\leftarrow ANC2} = 6.87 \times 10^{-6}$), with greater gene flow from the ANC2 to the eastern lineage compared to the reverse direction. The primary direction of gene flow was from the common ancestor (ANC2) of the western and southern lineages to the eastern lineage. Furthermore, after divergence of the western and southern lineages, estimates of gene flow between the two lineages were low and symmetrical ($M_{S\leftarrow W} = 1.49 \times 10^{-5}$, $M_{W\leftarrow S} = 1.62 \times 10^{-5}$), and the same was observed for western and eastern lineages ($M_{E\leftarrow W} = 4.26 \times 10^{-6}$, $M_{W\leftarrow E} = 4.26 \times 10^{-6}$). The gene flow between the latter two lineages was lower ($M_{E\leftarrow S} = 6.04 \times 10^{-6}$, $M_{S\leftarrow E} = 1.45 \times 10^{-5}$), with a greater gene flow from the eastern to the southern lineage compared to the reverse direction.

3.3. Identification of gene variants associated with local adaptation

We employed two genotype–environment association (GEA) methods, LFMM and RDA, to identify gene variants associated with environmental factors across the total distribution range of *Davidia involucrata*. Using LFMM, we tested the GEA of 19 environmental variables (Table S9) and identified 311,096 SNPs significantly associated with one or more environmental variables, based on a q-value cutoff of 0.05 (Fig. 3A). The RDA method, which detects GEA SNPs associated with multivariate environments, was performed using four uncorrelated variables (BIO2, BIO4, BIO17, and BIO18) selected to avoid issues related to multicollinearity (Fig. S7). A total of 439,424 SNPs were identified by RDA, with 22,630 SNPs overlapping with LFMM (Fig. 3A). These shared SNPs were considered "core adaptive variants" for local adaptation and were associated with 2881 genes (gene region and upstream/downstream 2 Kb) (Fig. 3A).

To assess the patterns of isolation by distance (IBD) and isolation by environment (IBE) in neutral and potentially adaptive SNPs, we

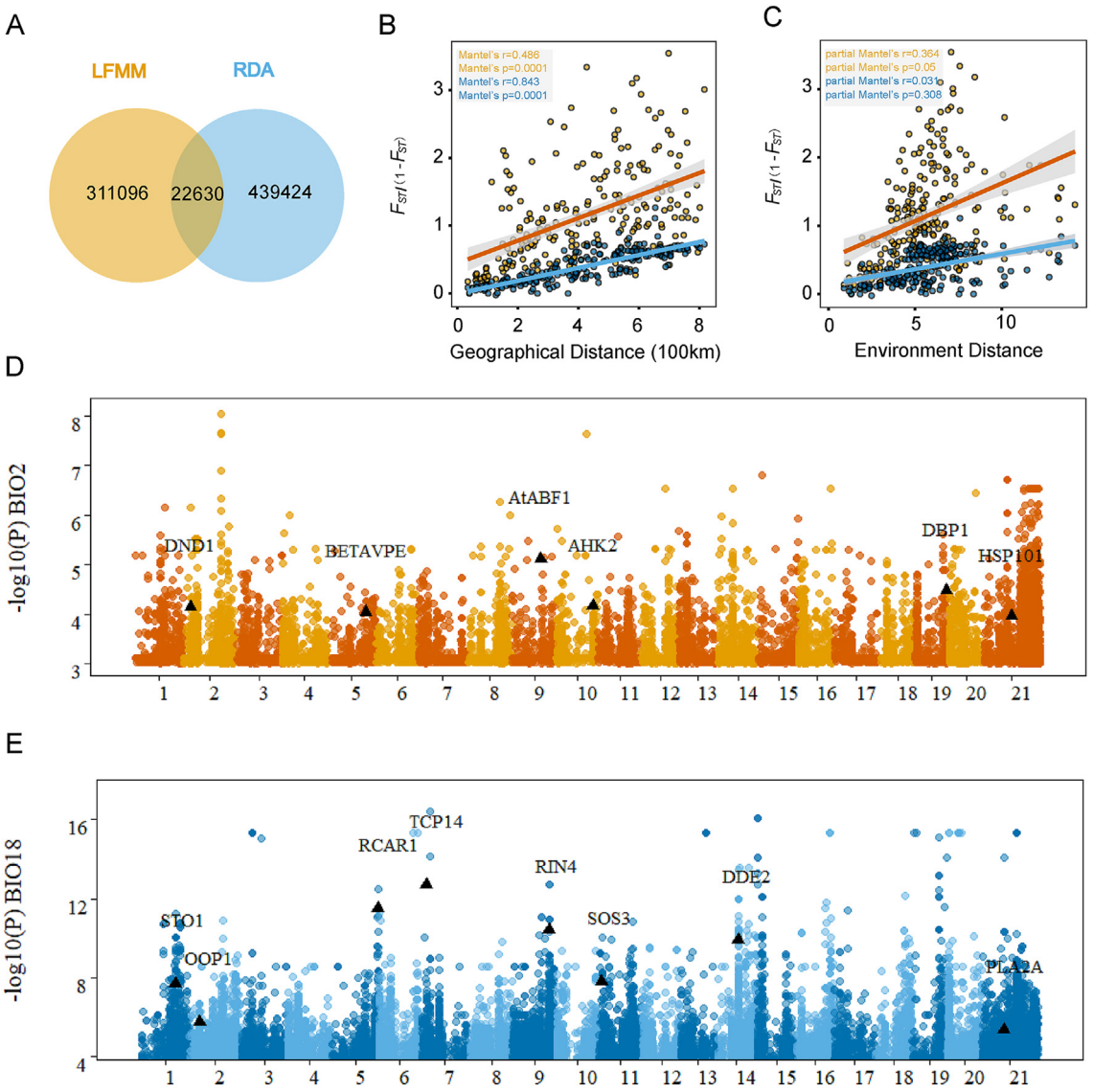

Fig. 3. Genome-wide screening of the loci associated with local environmental adaptation. (A) Genotype-environment association (GEA) genetic loci explored by latent factor linear mixed model (LFMM) and redundancy analysis (RDA). (B) Isolation-by-distance analyses (Mantel test, two-sided) for populations (n = 77) based on neutral (dark blue dots and light blue line) and adaptive variants (yellow dots and orange line) separately. The shadow of linear regression denotes the 95% confidence interval. (C) Isolation-by-environment analyses (partial Mantel test, two-sided, controlling for the effect of geographic distance) for populations (n = 77) based on neutral (dark blue dots and light blue line) and adaptive variants (yellow dots and orange line) separately. The shadow of linear regression denotes the 95% confidence interval. (D, E) Manhattan plot for variants associated with the Mean Diurnal Range (BIO2) (red) and the Precipitation of the Warmest Quarter (BIO18) (blue). Selected candidate genes are labeled in the plot at their respective genomic positions.

conducted Mantel and partial Mantel tests (Fig. 3B and C). The results showed a significant pattern of IBD in both neutral and adaptive variants (Fig. 3B). However, the partial Mantel test revealed a significant IBE only in adaptive SNPs, indicating that the genetic variation of adaptive variants was primarily influenced by the environment (Fig. 3C). Furthermore, we performed gene ontology (GO) enrichment analysis of the core adaptive genes, which revealed enrichment in growth and metabolic pathways, suggesting their role in environmental adaptation in *Davidia involucrata* (Fig. S8 and Table S10).

To provide examples of significant and well-known *Arabidopsis thaliana* homologous genes identified in associated with representative climate variables, we selected BIO2 (temperature) and BIO18 (precipitation) (Fig. 3D and E and Table S11). For example, *HSP101* (Dinv16319), associated with BIO2, is a cytosolic heat shock protein required for acclimation to high temperature and essential for plant survival under heat stress (Babbar et al., 2023). *AHK2* (Dinv21174) regulate plant organ size, flowering time and plant longevity (Bartrina et al., 2017). Additionally, there are other genes involved in immunity (e.g. *DND1* and *CAD1*), pollen development and fertility (*CER22* and *DRP1C*), embryogenesis (*EMB2775*), ABA responses (*AtABF1* and *CYP707A4*). *STO1* (Dinv13307), associated with BIO18, encodes 9-cis-epoxycarotenoid dioxygenase, encodes a key enzyme in the biosynthesis of abscisic acid. It catalyzes the first step of ABA biosynthesis from carotenoids and in doing so, enabling plant response to water stress (Al-Younis et al., 2021). Other precipitation-associated genes, such as *PUB23* (Dinv43996) and *ATHB29* (Dinv16647), are also involved in the response to water stress. Additionally, there are genes functionally involved in abscisic acid signaling regulation (e.g. *ABI3* and *RCAR1*), immunity (e.g. *RIN4* and *WRKY18*), and growth and development (e.g. *IAA22* and *PUR4*).

3.4. Positively selected genes within the eastern or western lineage

Due to the low population and individual numbers in the southern lineage, we focused our analyses on identifying positively selected genes in the eastern and western lineages. We analyzed the population genomic data of the "pure" individuals, excluding the hybrid populations, using metrics such as the π ratio, interspecific differentiation (F_{ST}), and XP-CLR. In the western lineage, we discovered a total of 28 protein-coding genes homologous to *Arabidopsis thaliana*, which were located in 15 outlier windows (20 kb) exhibiting selection signatures (Fig. 4A and Table S12). Among these genes, several are associated with stress response, reproduction, development, and metal transport (Figs. 4C and S9). For instance, *HSP60-2* (Dinv12114) is involved in copper ion and ATP binding and plays a role in the inflammatory response and salt stress in *A. thaliana* (Velinov et al., 2020). Another gene, *MEE70* (Dinv19349), encodes a WD-40 repeat-containing protein that participates in chromatin assembly as part of the CAF1 and FIE complex. Mutants of *MEE70* exhibit parthenogenetic development, including the proliferation of unfertilized endosperm and embryos (Chen et al., 2023). *PGM2* (Dinv28052) encodes a cytosolic phosphoglucomutase (PGM), and the loss of both *PGM2* and *PGM3* significantly impairs male and female gametophyte development (Malinova et al., 2014). *VAN* (Dinv37124) encodes a homeodomain transcription factor with sequence similarity to the *Arabidopsis* ovule development regulator gene *Bell1*. Mutants of *VAN* exhibit additional lateral organs and phyllotaxy defects (Bencivenga et al., 2016). Lastly, *ATAMT1* (Dinv11772) encodes a plasma membrane-localized ammonium transporter and contains a cytosolic trans-activation domain essential for ammonium uptake (Fig. S9).

In the eastern lineage, we identified a total of 36 protein-coding genes located in 20 outlier windows that have signals of positive selection (Fig. 4B and Table S12). Several of these genes are associated with stress resistance, disease resistance, reproduction, and development (Figs. 4D and S10). One of the genes identified is *VCT1* (Dinv35561), which is involved in cell wall carbohydrate biosynthesis, protein glycosylation, and ascorbate (vitamin C) biosynthesis. *VCT1* may play a role in controlling the transcription of defense-related and senescence-associated genes and influence ammonium sensitivity (Zhang et al., 2022). *SGS3* (Dinv07279) is required for posttranscriptional gene silencing and natural virus resistance (Mourrain et al., 2000). *ACS* (Dinv07280) appears to function as a monomer and may have an important role in preventing the toxic accumulation of fermentation products, including acetaldehyde, acetate, and ethanol (Mou et al., 2020). *ChlAKR* (Dinv35557) is believed to play a role in detoxifying reactive carbonyl compounds that could impair the photosynthetic process (Treffon et al., 2022). *AtLEA4-5* (Dinv14293) is involved in protecting enzyme activities from the adverse effects induced by freeze–thaw cycles in vitro (Cuevas-Velazquez et al., 2021). *EDA29* (Dinv31762) regulates phytochrome and responses to continuous far-red light stimulus through the high-irradiance response system (Staneloni et al., 2009).

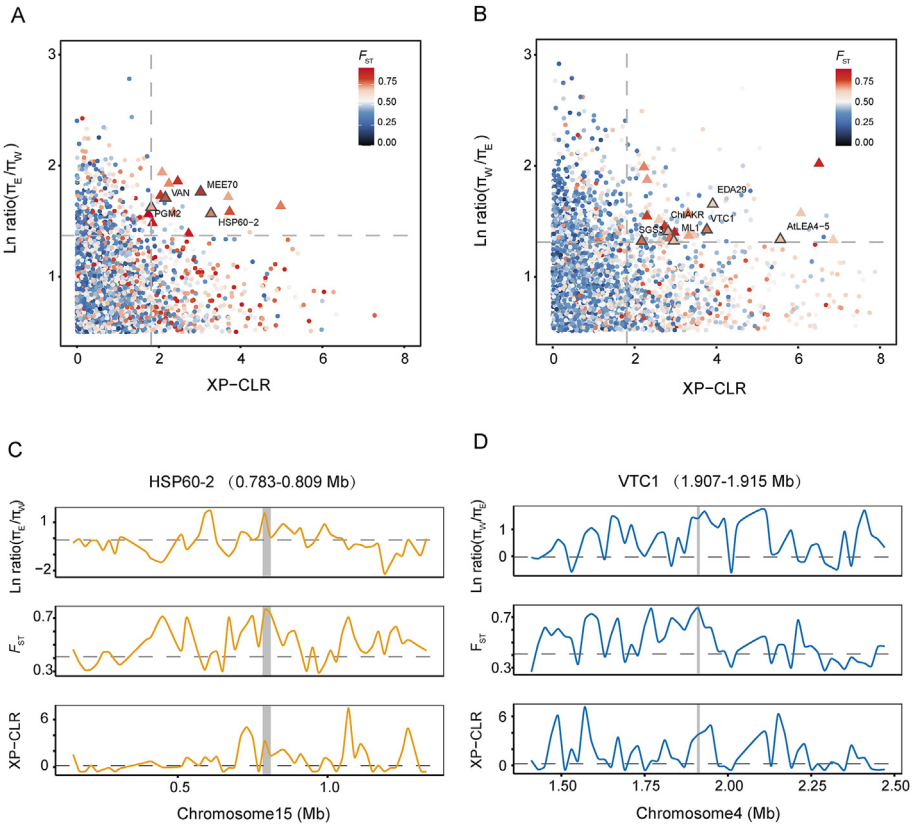

Fig. 4. Genomic region with selection signals. (A) Distribution of the XP-CLR(x-axis), ln ratio π_E/π_W (y-axis) and value of pairwise fixation index (F_{ST}) (color) between West and East lineages. The dashed vertical and horizontal lines indicate the significance threshold (corresponding to Z-test, $P < 0.025$, where XP-CLR >1.8122, ln ratio $\pi_E/\pi_W > 1.369859$, and $F_{ST} > 0.6785626$) used for extracting outliers (triangle symbol). (B) Distribution of the XP-CLR (x axis), ln ratio π_W/π_E (y axis) and value of pairwise fixation indices (F_{ST}) (color) between West and East lineages. The dashed vertical and horizontal lines indicate the significance threshold (corresponding to Z-test, $P < 0.005$, where XP-CLR >1.8122, ln ratio $\pi_W/\pi_E > 1.299672$, and $F_{ST} > 0.6785626$) used for extracting outliers (triangle symbol). (C) Selective sweep on chromosome 15 (0.783–0.809 Mb). (D) Selective sweep on chromosome 4 (1.907–1.915 Mb). Horizontal dashed lines represent the whole genome mean for the corresponding parameters.

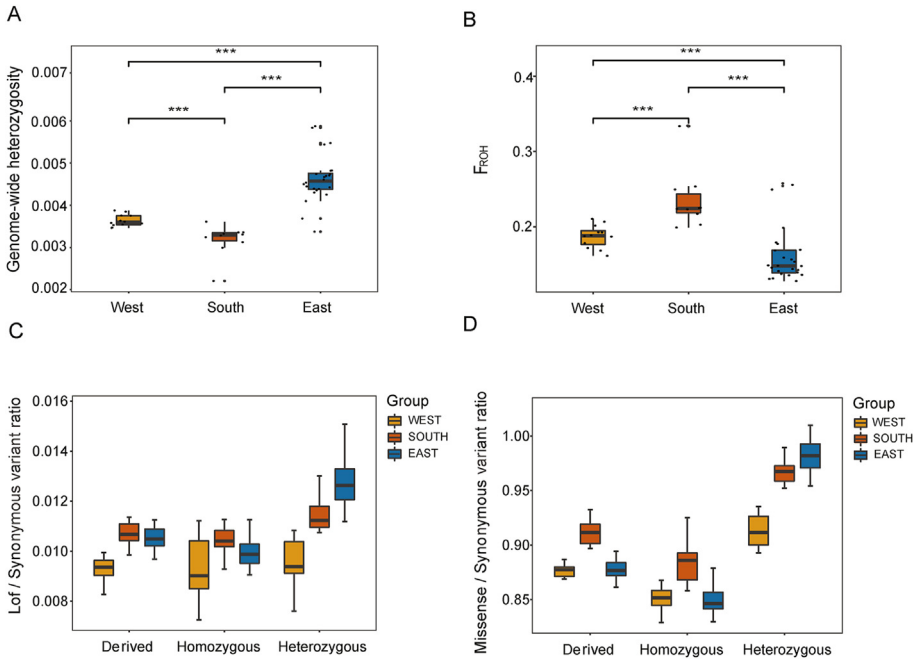

Fig. 5. Diversity, linkage disequilibrium, and mutation load metrics for *Davidia involucrata*. (A) Box plot of heterozygosity of the whole genome. (B) Box plot of F_{ROH} (sum of ROH >10 kb/genome effective length) for each individual in the two species. The line in the center of the box represents the median values, the edges of the box represent the first and third quartiles, and the whiskers above and below the box show the range of values. (C) The ratio of loss-of-function (LoF) variations in homozygous and heterozygous related to synonymous variations. (D) The ratio of missense variations in homozygous and heterozygous related to synonymous variations.

3.5. *Heterozygosity, inbreeding level and mutation load of* Davidia involucrata

We conducted an analysis of whole-genome heterozygosity, inbreeding level, and mutation load in the three defined groups. The heterozygosity values ranged from 2.20×10^{-3} to 5.87×10^{-3} (Fig. 5A). The eastern lineage exhibited the highest heterozygosity (mean $He = 4.63 \times 10^{-3}$) and the largest population size, while the southern lineage showed the lowest heterozygosity (mean $He = 3.18 \times 10^{-3}$) and the smallest population size. The western lineage displayed a moderate level of heterozygosity (mean $He = 3.63 \times 10^{-3}$) and population size.

The decrease in heterozygosity observed in all *Davidia involucrata* individuals may be associated with high inbreeding coefficients. Genomic inbreeding coefficients (F) ranged from 0.12 to 0.33 (Fig. 5B). Individuals with a higher degree of inbreeding generally had lower heterozygosity and higher levels of continuous pure sum fragments. The southern lineage exhibited the highest level of genomic inbreeding (mean F = 0.2354), followed by the western lineage (mean F = 0.1873) and the eastern lineage (mean F = 0.1594) (Fig. 5B). These findings indicate that the southern lineage had a higher overall inbreeding level compared to the western lineage, and the western lineage had a higher inbreeding level compared to the eastern lineage.

To assess the mutation load of each lineage, we calculated the ratio of deleterious derived mutations to synonymous derived mutations (Del/Syn) and the ratio of loss-of-function derived mutations to synonymous derived mutations (LoF/Syn). The homozygous Missense/Syn and LoF/Syn values of derived alleles were highest in the South lineage (Fig. 5C and D), suggesting that the southern lineage accumulated a greater mutation load compared to the eastern and western lineages. We also calculated the Missense/Syn and LoF/Syn values for derived alleles in both homozygous and heterozygous genotypes across different lineages. The Missense/Syn and LoF/Syn values were higher in the heterozygous form compared to the homozygous state. This may be due to the fact that homozygous deleterious mutations are more likely to cause individual death, resulting in higher fitness for heterozygous mutations compared to homozygous mutations. Consequently, most harmful mutations are present in the population in heterozygous genotypes. The proportion of alleles derived from homozygous genotypes and alleles derived from heterozygous genotypes was highest in the southern lineage (LoF/Syn). The eastern lineage exhibited the highest proportion of alleles derived from heterozygous genotypes, while the proportion of alleles derived from homozygous genotypes in the eastern lineage was similar to the western lineage and lower than that of the southern lineage. The western lineage had a low proportion of alleles derived from both homozygous and heterozygous genotypes compared to the other lineages. This pattern of mutation load suggests that the southern lineage experienced a more prolonged population bottleneck and thus carried more alleles derived from homozygous genotypes, indicating a higher risk of sustained decline in effective population size.

4. Discussion

In this study, we identified three cryptic local lineages within the endangered “fossil” plant *Davidia involucrata* based on whole-genome data from representative populations across its entire distribution region. Additionally, we observed hybrid populations that resulted from gene flow between the three lineages. We discovered numerous gene variants correlated with the environment and also identified many genes showing signals of positive evolution in the local lineages. These findings indicate the highly dynamic evolution occurring within this species despite its similar phenotype. Our results are crucial for the development of new conservation strategies for this endangered species.

4.1. Cryptic divergence and hybridization during the long evolutionary history

The current phenotype of *Davidia involucrata* closely resembles its Cenozoic fossil counterpart (Eyde, 1997; Manchester et al., 2009; Tang et al., 2017). This lineage is estimated to have diverged from the monotypic genus *Camptotheca* approximately 60 Mya (Chen et al., 2020). If no other species underwent further evolution and differentiation for the genus *Davidia*, *D. involucrata* has survived since its split from *Camptotheca* for a considerable period. It was once distributed in North America, with populations or distinct lineages there extinguished (Eyde, 1997; Manchester et al., 2009). The relic populations in central and southwest China may now represent only a fraction of the once widely distributed populations and lineages. Our analysis, based on whole genomic data, indicates that the earliest divergence between the eastern lineage and the common ancestor of the western and southern lineages occurred around 3 Mya (Fig. 2). This falls within the Pliocene epoch, although it is younger than the estimates of 4–5 Mya from two earlier studies that relied on several DNA markers only (Chen et al., 2015; Ma et al., 2015). Furthermore, our analysis suggests that the divergence between the western and southern lineages took place around 0.3 Mya (Fig. 2). These findings indicate that all of the other lineages or populations originated before the Pliocene may have become extinct, as suggested by fossil records (Eyde, 1997; Manchester et al., 2009). The current three lineages originated from further divergence when *D. involucrata* retreated to central and southwest China. The Pliocene to Quaternary period witnessed significant geographical and climatic fluctuations (An et al., 2001), which likely contributed to the formation of the three observed lineages. Interestingly, intraspecific divergences have been discovered during this period in other "fossil" plants, such as *Cercidiphyllum japonicum* (Zhu et al., 2020), *Euptelea pleiospermum* (Cao et al., 2020), and *Tetracentron sinense* (Liu et al., 2020). Therefore, it is highly probable that many plants in this region experienced similar geographical and climatic influences that facilitated their intraspecific local divergences.

We further discovered that these divergences between lineages involved the selection of allelic variations in many genes related to environmental adaptation. For instance, we identified allelic variations in 2881 genes that were correlated with environmental factors (Fig. 3A). These genes are enriched in growth and metabolic pathways (Fig. S8 and Table S10). In the eastern lineage, we observed signals of positive selection in many genes associated with stress resistance, disease resistance, reproduction, and development (Figs. 4B–D and S10; Table S12). In the western lineage, we identified several genes with selection signatures (Figs. 4A–C and S9; Table S12) that are linked to stress response, reproduction, development, and metal transport (Fig. 3C). These genetic variations likely provided the basis for the local adaptation and the following divergence of *Davidia involucrata*. Despite these localized selections, we still observed multiple hybridization events between the three lineages. Between the eastern and southern lineages, we identified two distinct hybrid populations in the areas between their respective distributions. Additionally, two populations at the northern distribution range of the western lineage, with a significant geographic gap, exhibited clear introgressions from the eastern lineage (Fig. 1). Four populations located in the intermediate distribution range between the western and southern lineages clearly originated from hybridization between them. This hybrid zone extended over a relatively wide range. Our demographic analyses of the three lineages (Fig. 2) suggested that gene flow between them has persisted from their divergence to the present. These frequent hybridizations contrast with previous studies (Chen et al., 2015; Ma et al., 2015) that were based on a limited number of DNA fragments and only detected minimal gene flow between the eastern and western lineages. Moreover, this finding is inconsistent with the limited gene flow inferred from both seed dispersal (Zhang et al., 2000) and pollinators (Sun et al., 2008). The existence of unknown seed and pollination dispersal animals responsible for long-distance dispersals should be further investigated.

It is important to note that the third southern lineage demonstrated genetic admixture from both the western and eastern lineages when K = 2 (Fig. 1). All sampled individuals of two populations showed the stable genomic compositions from the other two lineages, indicating that they may have evolved as one independently evolving lineage for many generations after the initial hybridization as modelled by our coalescent analyses (Fig. 2). This differs from those partly introgressions only a few individuals and unstable genetic admixtures between individuals in the hybrid populations. However, our coalescent analyses of the alternative origin models still support its bifurcating divergence from the western lineage, with substantial introgression from the eastern lineage. This is likely due to approximately 89% of genomic elements originating from the western lineage. On the other hand, the hybrid lineages with a composition of 75% from one parent and 25% from the other parent unequivocally indicate their origin from hybridization rather than a bifurcating divergence (Wang et al., 2021). This situation warrants further investigation, particularly regarding the extent to which a small genomic contribution from the lineage influenced the establishment and maintenance of the southern lineage through reproductive isolation from hybridization (Wang et al., 2021). One caveat in our study, only two populations were sampled this southern lineage. More populations should be added in the future analyses.

4.2. Mutation load and conservation implication

Among the three recovered lineages, the southern lineage exhibited the lowest diversity, possibly due to the small number of individuals sampled and their restricted distribution range. We discovered evidence of inbreeding in each lineage, as indicated by the presence of continuous homozygous segments ranging from 0.12 to 0.33 (Fig. 5B). This inbreeding can result in the accumulation of deleterious mutations, known as mutation load, which is primarily influenced by heterozygous recessive deleterious mutations. The southern lineage demonstrated a relatively higher mutation load, while the eastern lineage displayed a comparatively lower level of mutation load. The accumulation of mutation load is closely associated with the population dynamics history, including factors such as effective population size, population expansion and contraction, and the duration of population bottlenecks (Robinson et al., 2023). The southern lineage may have experienced rapid population expansion or bottlenecks, with the effective population size remaining around 1000 individuals (Fig. 2A). This would result in the loss of rare alleles, and subsequent genetic drift and inbreeding events would increase the rate of fixation of homozygotes, converting heterozygous recessive deleterious mutations into homozygotes (Lynch et al., 1995). On the other hand, both the eastern and western lineages experienced a significant decrease in effective population size, but they later rebounded and maintained relatively large effective population sizes (Fig. 2). However, it is important to note that the genetic diversity and all examined mutation load are significantly higher or lower than those observed in relic trees based on genomic data from eastern Asia (Chen et al., 2019; Cao et al., 2020; Liu et al., 2020; Zhu et al., 2020).

The species level diversity $\theta\pi$ values and θ_W values of *Davidia involucrata* were obtained based on resequencing data ($\theta\pi = 4.67 \times 10^{-3}$; $\theta_W = 1.16 \times 10^{-3}$; Table S3). Through comparison,

it was found that the genetic diversity level of *D. involucrata* was higher than that of other "fossil" plants: *Ginkgo biloba* ($\theta\pi = 2.11 \times 10^{-3}$, $\theta w = 2.36 \times 10^{-3}$) (Zhao et al., 2019); Chinese lineage of *Cercidiphyllum japonicum* ($\theta\pi = 1.05 \times 10^{-3}$) (Zhu et al., 2020); *Liriodendron chinense* CW lineage ($\theta\pi = 6.89 \times 10^{-4}$), CE lineage ($\theta\pi = 5.39 \times 10^{-4}$) (Chen et al., 2019). Our population genomic analyses of *D. involucrata* indicate relatively high genetic diversity. This suggests that despite being an endangered species, it does not face significant extinction pressure. One of the main reasons for this may be its outcrossing breeding system and the observed long-distance gene flow. Furthermore, gene flow and hybridization may contribute to the fitness of this endangered relic species. However, in addition to the two previously identified cryptic evolutionary lineages (Chen et al., 2015, 2020; Ma et al., 2015), we have identified a third lineage in southern Sichuan and northern Yunnan. All three lineages should be effectively conserved as they represent distinct units, and attention should also be given to the hybrid populations between them. These hybrid populations have the potential to evolve into new evolutionary units, similar to what has been observed for the southern lineage. Additionally, priority should be given to *in situ* improvements in population size of each lineage and each population. It is recommended to protect the species in its natural habitat to maintain a relatively large effective population size and gene flow among the different lineages and populations. The conservation of the southern lineage should be prioritized. For *ex situ* conservation, germplasm resources from all three lineages and hybrid populations should be collected to preserve the maximum genetic diversity of this fossil plant.

Data availability

The whole-genome resequencing data generated in this study have been deposited in the National Genomics Data Center (NGDC, https://ngdc.cncb.ac.cn) under accession number PRJCA020962.

CRediT authorship contribution statement

Yumeng Ren: Writing – review & editing, Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Lushui Zhang:** Investigation. **Xuchen Yang:** Formal analysis. **Hao Lin:** Formal analysis. **Yupeng Sang:** Formal analysis. **Landi Feng:** Formal analysis. **Jianquan Liu:** Writing – review & editing, Methodology, Funding acquisition. **Minghui Kang:** Writing – review & editing, Writing – original draft, Project administration.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

This study was supported by the Second Tibetan Plateau Scientific Expedition and Research program (No. 2019QZKK0502), Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB31010300), Fundamental Research Funds for the Central Universities, and International Collaboration 111 Program (BP0719040).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pld.2024.02.004.

References

Adams, M., Raadik, T.A., Burridge, C.P., et al., 2014. Global biodiversity assessment and hyper-cryptic species complexes: more than one species of elephant in the room? Syst. Biol. 63, 518–533.
Al-Younis, I., Moosa, B., Kwiatkowski, M., et al., 2021. Functional crypto-adenylate cyclases operate in complex plant proteins. Front. Plant Sci. 12, 711749.
Alexa, A., Rahnenfuhrer, J., 2022. topGO: Enrichment Analysis for Gene Ontology. R Package Version 2.48.0.
Alexander, D.H., Lange, K., 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12, 1–6.
Allendorf, F.W., Hohenlohe, P.A., Luikart, G., 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11, 697–709.
An, Z., Kutzbach, J.E., Prell, W.L., et al., 2001. Evolution of Asian monsoons and phased uplift of the Himalaya-Tibetan plateau since Late Miocene times. Nature 411, 62–66.
Babbar, R., Tiwari, L.D., Mishra, R.C., et al., 2023. *Arabidopsis* plants overexpressing additional copies of heat shock protein Hsp 101 showed high heat tolerance and endo-gene silencing. Plant Sci. 330, 111639.
Bartrina, I., Jensen, H., Novák, O., et al., 2017. Gain-of-function mutants of the cytokinin receptors AHK2 and AHK3 regulate plant organ size, flowering time and plant longevity. Plant Physiol. 173, 1783–1797.
Bencivenga, S., Serrano-Mislata, A., Bush, M., et al., 2016. Control of oriented tissue growth through repression of organ boundary genes promotes stem morphogenesis. Dev. Cell 39, 198–208.
Bickford, D., Lohman, D.J., Sodhi, N.S., et al., 2007. Cryptic species as a window on diversity and conservation. Trends Ecol. Evol. 22, 148–155.
Bolnick, D.I., Svanbäck, R., Fordyce, J.A., et al., 2003. The ecology of individuals: incidence and implications of individual specialization. Am. Nat. 161, 1–28.
Brodersen, J., Seehausen, O., 2014. Why evolutionary biologists should get seriously involved in ecological monitoring and applied biodiversity assessment programs. Evol. Appl. 7, 968–983.
Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097.
Camacho, C., Coulouris, G., Avagyan, V., et al., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 1–9.
Cao, Y.N., Zhu, S.S., Chen, J., et al., 2020. Genomic insights into historical population dynamics, local adaptation, and climate change vulnerability of the East Asian Tertiary relict *Euptelea* (Eupteleaceae). Evol. Appl. 13, 2038–2055.
Chen, H., Patterson, N., Reich, D., 2010. Population differentiation as a test for selective sweeps. Genome Res. 20, 393–402.
Chen, J.M., Zhao, S.Y., Liao, Y.Y., et al., 2015. Chloroplast DNA phylogeographic analysis reveals significant spatial genetic structure of the relictual tree *Davidia involucrata* (Davidiaceae). Conserv. Genet. 16, 583–593.
Chen, S., Zhou, Y., Chen, Y., et al., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.
Chen, J., Hao, Z., Guang, X., et al., 2019. *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. Nat. Plants 5, 18–25.
Chen, Y., Ma, T., Zhang, L., et al., 2020. Genomic analyses of a "living fossil": the endangered dove-tree. Mol. Ecol. Resour. 20, 756–769.
Chen, X., MacGregor, D.R., Stefanato, F.L., et al., 2023. A VEL3 histone deacetylase complex establishes a maternal epigenetic state controlling progeny seed dormancy. Nat. Commun. 14, 2220.
Cingolani, P., Platts, A., Wang, L.L., et al., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6, 80–92.
Crandall, K.A., Bininda-Emonds, O.R., Mace, G.M., et al., 2000. Considering evolutionary processes in conservation biology. Trends Ecol. Evol. 15, 290–295.
Cuevas-Velazquez, C.L., Vellosillo, T., Guadalupe, K., et al., 2021. Intrinsically disordered protein biosensor tracks the physical-chemical effects of osmotic stress on cells. Nat. Commun. 12, 5438.
Danecek, P., Auton, A., Abecasis, G., et al., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.
Emms, D., Kelly, S., 2019. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. Genome Biol. 20, 238.
Excoffier, L., Marchi, N., Marques, D.A., et al., 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics 37, 4882–4885.
Eyde, R.H., 1997. Fossil record and ecology of *Nyssa* (Cornaceae). Bot. Rev. 63, 97–123.
Feckler, A., Zubrod, J.P., Thielsch, A., et al., 2014. Cryptic species diversity: an overlooked factor in environmental management? J. Appl. Ecol. 51, 958–967.
Feng, L., Xu, Z.Y., Wang, L., 2019. Genetic diversity and demographic analysis of an endangered tree species *Diplopanax stachyanthus* in subtropical China: implications for conservation and management. Conserv. Genet. 20, 315–327.
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.
Frankham, R., Ballou, J.D., Briscoe, D.A., 2010. Introduction to Conservation Genetics, Second ed. Cambridge University Press, UK.
Frichot, E., François, O., 2015. LEA: an R package for landscape and ecological association studies. Methods Ecol. Evol. 6, 925–929.
Funk, W.C., McKay, J.K., Hohenlohe, P.A., et al., 2012. Harnessing genomics for delineating conservation units. Trends Ecol. Evol. 27, 489–496.

He, Z.C., Li, J.Q., Wang, H.C., 2004. Karyomorphology of *Davidia involucrata* and *Camptotheca acuminata*, with special reference to their systematic positions. Bot. J. Linn. Soc. 144, 193–198.
Hewitt, G., 2000. The genetic legacy of the Quaternary ice ages. Nature 405, 907–913.
Hewitt, G.M., 2004. Genetic consequences of climatic oscillations in the Quaternary. Phil. Trans. R. Soc. Lond. B-Biol. Sci. 359, 183–195.
Hu, H., Yang, Y., Li, A., et al., 2022. Genomic divergence of *Stellera chamaejasme* through local selection across the Qinghai-Tibet Plateau and northern China. Mol. Ecol. 31, 4782–4796.
Jörger, K.M., Schrödl, M., 2013. How to describe a cryptic species? Practical challenges of molecular taxonomy. Front. Zool. 10, 1–27.
Kang, M., Fu, R., Zhang, P., et al., 2021. A chromosome-level *Camptotheca acuminata* genome assembly provides insights into the evolutionary origin of camptothecin biosynthesis. Nat. Commun. 12, 3531.
Keightley, P.D., Jackson, B.C., 2018. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. Genetics 209, 897–906.
Letunic, I., Bork, P., 2007. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.
Li, H., Handsaker, B., Wysoker, A., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Li, X., Li, Z., He, C., et al., 2012. Genetic diversity of the endangered *Davidia involucrata* by AFLP analysis. Acta Hortic. Sin. 39, 992–998.
Li, J.L., Zhong, L.L., Wang, J., et al., 2021. Genomic insights into speciation history and local adaptation of an alpine aspen in the Qinghai-Tibet Plateau and adjacent highlands. J. Syst. Evol. 59, 1220–1231.
Liu, Q., Vetukuri, R.R., Xu, W., et al., 2019. Transcriptomic responses of dove tree (*Davidia involucrata* Baill.) to heat stress at the seedling stage. Forests 10, 656.
Liu, P.L., Zhang, X., Mao, J.F., et al., 2020. The *Tetracentron* genome provides insight into the early evolution of eudicots and the formation of vessel elements. Genome Biol. 21, 1–30.
Luo, S., He, Y., Ning, G., et al., 2011. Genetic diversity and genetic structure of different populations of the endangered species *Davidia involucrata* in China detected by inter-simple sequence repeat analysis. Trees (Berl.) 25, 1063–1071.
Lynch, M., Conery, J., Burger, R., 1995. Mutation accumulation and the extinction of small populations. Am. Nat. 146, 489–518.
Ma, Q., Du, Y.J., Chen, N., et al., 2015. Phylogeography of *Davidia involucrata* (Davidiaceae) inferred from cpDNA haplotypes and nSSR data. Syst. Bot. 40, 796–810.
Malinova, I., Kunz, H.H., Alseekh, S., et al., 2014. Reduction of the cytosolic phosphoglucomutase in *Arabidopsis* reveals impact on plant growth, seed and root development, and carbohydrate partitioning. PLoS One 9, e112468.
Manchester, S.R., Chen, Z.D., Lu, A.M., et al., 2009. Eastern Asian endemic seed plant genera and their paleogeographic history throughout the Northern Hemisphere. J. Syst. Evol. 47, 1–42.
Mao, K., Liu, J., 2012. Current 'relicts' more dynamic in history than previously thought. New Phytol. 196, 329–331.
McKenna, A., Hanna, M., Banks, E., et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.
Moritz, C., 1994. Defining 'evolutionarily significant units' for conservation. Trends Ecol. Evol. 9, 373–375.
Mou, W., Kao, Y.T., Michard, E., et al., 2020. Ethylene-independent signaling by the ethylene precursor ACC in Arabidopsis ovular pollen tube attraction. Nat. Commun. 11, 4082.
Mourrain, P., Béclin, C., Elmayan, T., et al., 2000. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. Cell 101, 533–542.
Mu, W., Wei, J., Yang, T., et al., 2020. The draft genome assembly of the critically endangered *Nyssa yunnanensis*, a plant species with extremely small populations endemic to Yunnan Province, China. Gigabyte 2020, 1–11.
Naciri, Y., Linder, H.P., 2015. Species delimitation and relationships: the dance of the seven veils. Taxon 64, 3–16.
Oksanen, J., Blanchet, F., Kindt, R., et al., 2017. Vegan: community ecology package. https://cran.r-project.org/web/packages/vegan/index.html.
Palsbøll, P.J., Berube, M., Allendorf, F.W., 2007. Identification of management units using population genetic data. Trends Ecol. Evol. 22, 11–16.
Pante, E., Puillandre, N., Viricel, A., et al., 2015. Species are hypotheses: avoid connectivity assessments based on pillars of sand. Mol. Ecol. 24, 525–544.
Pauls, S.U., Nowak, C., Bálint, M., et al., 2013. The impact of global climate change on genetic diversity within populations and species. Mol. Ecol. 22, 925–946.
Petit, R.J., El Mousadik, A., Pons, O., 1998. Identifying populations for conservation on the basis of genetic markers. Conserv. Biol. 12, 844–855.
Petit, R.J., Aguinagalde, I., de Beaulieu, J.L., et al., 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300, 1563–1565.
Pfenninger, M., Schwenk, K., 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evol. Biol. 7, 1–6.
Price, M.N., Dehal, et al., 2010. FastTree 2-Approximately Maximum-Likelihood trees for large alignments. PLoS One 5, e9490.
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.
Qi, X.S., Chen, C., Comes, H.P., et al., 2012. Molecular data and ecological niche modelling reveal a highly dynamic evolutionary history of the East Asian Tertiary relict *Cercidiphyllum* (Cercidiphyllaceae). New Phytol. 196, 617–630.
Qiu, Y.X., Fu, C.X., Comes, H.P., 2011. Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. Mol. Phylogenet. Evol. 59, 225–244.
Robinson, J., Kyriazis, C.C., Yuan, S.C., et al., 2023. Deleterious variation in natural populations and implications for conservation genetics. Annu. Rev. Anim. Biosci. 11, 93–114.
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.
Scheffers, B.R., Joppa, L.N., Pimm, S.L., et al., 2012. What we know and don't know about Earth's missing biodiversity. Trends Ecol. Evol. 27, 501–510.
Schluter, D., 2000. The Ecology of Adaptive Radiation. Oxford Univ. Press Inc., New York, pp. 1–288.
Seehausen, O., 2009. Speciation affects ecosystems. Nature 458, 1122–1123.
Shang, H.Y., Li, Z.H., Dong, M., et al., 2015. Evolutionary origin and demographic history of an ancient conifer (*Juniperus microsperma*) in the Qinghai-Tibetan Plateau. Sci. Rep. 5, 10216.
Shen, Y., Xia, H., Tu, Z., et al., 2022. Genetic divergence and local adaptation of *Liriodendron* driven by heterogeneous environments. Mol. Ecol. 31, 916–933.
Song, C., Bao, M., 2006. Genetic diversity of RAPD mark for natural *Davidia involucrata* populations. Front. For. China 1, 95–99.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Staneloni, R.J., Rodriguez-Batiller, M.J., Legisa, D., et al., 2009. Bell-like homeodomain selectively regulates the high-irradiance response of phytochrome A. Proc. Natl. Acad. Sci. U.S.A. 106, 13624–13629.
Sun, J.F., Gong, Y.B., Renner, S.S., et al., 2008. Multifunctional bracts in the dove tree *Davidia involucrata* (Nyssaceae: Cornales): rain protection and pollinator attraction. Am. Nat. 171, 119–124.
Tang, C.Q., Dong, Y.F., Herrando-Moraira, S., et al., 2017. Potential effects of climate change on geographic distribution of the Tertiary relict tree species *Davidia involucrata* in China. Sci. Rep. 7, 43822.
Torres-Cambas, Y., Ferreira, S., Cordero-Rivera, A., et al., 2017. Identification of evolutionarily significant units in the Cuban endemic damselfly *Hypolestes trinitatis* (Odonata: Hypolestidae). Conserv. Genet. 18, 1229–1234.
Treffon, P., Rossi, J., Gabellini, G., et al., 2022. Proteome profiling of a S-Nitrosoglutathione reductase (GSNOR) null mutant reveals that aldo-keto reductases form a new class of enzymes involved in nitric oxide homeostasis. Faseb. J. 36, 1.
Velinov, V., Vaseva, I., Zehirov, G., et al., 2020. Overexpression of the NMig1 gene encoding a NudC domain protein enhances root growth and abiotic stress tolerance in *Arabidopsis thaliana*. Front. Plant Sci. 11, 815.
Wang, L., Abbott, R.J., Zheng, W., et al., 2009. History and evolution of alpine plants endemic to the Qinghai-Tibetan Plateau: *Aconitum gymnandrum* (Ranunculaceae). Mol. Ecol. 18, 709–721.
Wang, Z.W., Chen, S.T., Nie, Z.L., et al., 2015. Climatic factors drive population divergence and demography: insights based on the phylogeography of a riparian plant species endemic to the Hengduan Mountains and adjacent regions. PLoS One 10, e0145014.
Wang, Z., Jiang, Y., Bi, H., et al., 2021. Hybrid speciation via inheritance of alternate alleles of parental isolating genes. Mol. Plant 14, 208–222.
Wu, Y., Yang, J., Yang, Y., et al., 2023. The genome sequence and demographic history of *Przewalskia tangutica* (Solanaceae), an endangered alpine plant on the Qinghai-Tibet Plateau. DNA Res. 30, dsad005.
Xue, Y., Prado-Martinez, J., Sudmant, P.H., et al., 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348, 242–245.
Yang, X., Kang, M., Yang, Y., et al., 2019. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. Sci. Data 6, 282.
Yang, Y., Zhang, L., Huang, X., et al., 2020. Response of photosynthesis to different concentrations of heavy metals in *Davidia involucrata*. PLoS One 15, e0228563.
Zachos, J., Pagani, M., Sloan, L., et al., 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292, 686–693.
Zdobnov, E.M., Apweiler, R., 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848.
Zhang, Q., Guo, Q., Xu, D., et al., 2000. Influence of climate changes on geographical distribution of *Davidia involucrata*, a precious and endangered species native to China. Sci. Silvae Sin. 36, 47–52.
Zhang, C., Dong, S.S., Xu, J.Y., et al., 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35, 1786–1788.
Zhang, X., Sun, Y., Landis, J.B., et al., 2020. Genomic insights into adaptation to heterogeneous environments for the ancient relictual *Circaeaster agrestis* (Circaeasteraceae, Ranunculales). New Phytol. 228, 285–301.
Zhang, C., Zhao, S., Li, Y.S., et al., 2022. Crystal structures of *Arabidopsis thaliana* GDP-D-Mannose pyrophosphorylase VITAMIN c DEFECTIVE 1. Front. Plant Sci. 13, 899738.
Zhao, Y.P., Fan, G., Yin, P.P., et al., 2019. Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. Nat. Commun. 10, 4201.
Zhu, S., Chen, J., Zhao, J., et al., 2020. Genomic insights on the contribution of balancing selection and local adaptation to the long-term survival of a widespread living fossil tree, *Cercidiphyllum japonicum*. New Phytol. 228, 1674–1689.