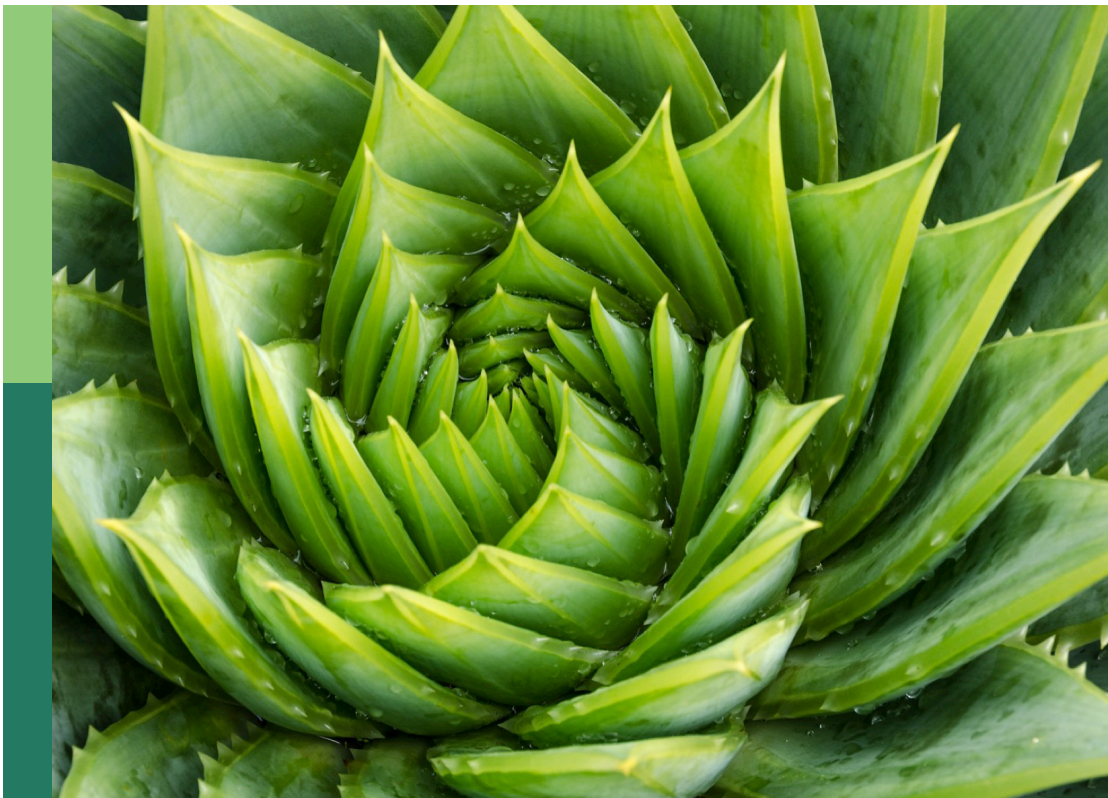
frontiers | Research Topics

Forest tree conservation genomics

Edited by
Fang Du, Rong Wang, Saneyoshi Ueno and Guillaume De Lafontaine

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-2770-2
DOI 10.3389/978-2-8325-2770-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Forest tree conservation genomics

Topic editors

Fang Du — Beijing Forestry University, China

Rong Wang — East China Normal University, China

Saneyoshi Ueno — Forestry and Forest Products Research Institute, Japan

Guillaume De Lafontaine — Université du Québec à Rimouski, Canada

Citation

Du, F., Wang, R., Ueno, S., De Lafontaine, G., eds. (2023). *Forest tree conservation genomics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2770-2

Table of contents

frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 February 2022
doi: 10.3389/fpls.2022.814718

Transposable Elements: Distribution, Polymorphism, and Climate Adaptation in *Populus*

*Yiyang Zhao[1,2], Xian Li[1,2], Jianbo Xie[1,2], Weijie Xu[1,2], Sisi Chen[1,2], Xiang Zhang[1,2], Sijia Liu[1,2], Jiadong Wu[1,2], Yousry A. El-Kassaby[3] and Deqiang Zhang[1,2]**

[1] National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [2] Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [3] Department of Forest and Conservation Sciences, Forest Sciences Centre, Faculty of Forestry, The University of British Columbia, Vancouver, BC, Canada

Transposable elements (TEs) are a class of mobile genetic elements that make effects on shaping rapid phenotypic traits of adaptive significance. TE insertions are usually related to transcription changes of nearby genes, and thus may be subjected to purifying selection. Based on the available genome resources of *Populus*, we found that the composition of Helitron DNA family were highly variable and could directly influence the transcription of nearby gene expression, which are involving in stress-responsive, programmed cell death, and apoptosis pathway. Next, we indicated TEs are highly enriched in *Populus trichocarpa* compared with three other congeneric poplar species, especially located at untranslated regions (3′UTRs and 5′UTRs) and Helitron transposons, particularly 24-nt siRNA-targeted, are significantly associated with reduced gene expression. Additionally, we scanned a representative resequenced *Populus tomentosa* population, and identified 9,680 polymorphic TEs loci. More importantly, we identified a Helitron transposon located at the 3′UTR, which could reduce *WRKY18* expression level. Our results highlight the importance of TE insertion events, which could regulate gene expression and drive adaptive phenotypic variation in *Populus*.

Keywords: transposable elements, Helitron transposons, 24 nt siRNA, forest genetic resources conservation, adaptive evolution

OPEN ACCESS

Edited by:
Guillaume De Lafontaine, Université du Québec à Rimouski, Canada

Reviewed by:
Jinhui Chen, Hainan University, China
Shengqing Shi, Chinese Academy of Forestry, China

***Correspondence:**
Deqiang Zhang
DeqiangZhang@bjfu.edu.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *14 November 2021*
Accepted: *10 January 2022*
Published: *01 February 2022*

Citation:
Zhao Y, Li X, Xie J, Xu W, Chen S, Zhang X, Liu S, Wu J, El-Kassaby YA and Zhang D (2022) Transposable Elements: Distribution, Polymorphism, and Climate Adaptation in Populus. Front. Plant Sci. 13:814718. doi: 10.3389/fpls.2022.814718

INTRODUCTION

Transposable elements (TEs), which are known as a source of genetic variation (McClintock, 1984), and its composition is extremely different in diverse species (Le Rouzic et al., 2007; Barron et al., 2014; Le et al., 2015; Li et al., 2018). TEs are mobile genetic sequences, with known two major categories: Class I retrotransposons and Class II DNA transposons (Wicker et al., 2007). Intriguingly, different transposons often form the major branches in diverse species. For example, there were less than a dozen long terminal repeat (LTR) retro-families in barley (Thiel et al., 2009), and the outbreak of retrotransposon activity shaped the current state of its reference genome in rice (El Baidouri and Panaud, 2013). Abundant information about transposon activity can be obtained by comparing closely related species (Brookfield, 2005; Agren et al., 2014; Barron et al., 2014), which will contribute to an in-depth understanding of the frequency and impact of TE activities in biological evolution. TEs play an important role in the formation of genomic structure and phenotypic variation in different organisms (Feschotte et al., 2002; Ellinghaus et al., 2008;

Feschotte, 2008). Moreover, sequence and genomic distribution variability produced by TEs had become sources for regulatory elements (Feschotte, 2008). Therefore, the TE insertions can regulate the gene expression level through *cis*- or *trans*-regulatory elements located in TE sequences, or through epigenetic modification caused by TE insertions or deletions (Chung et al., 2007; Gonzalez et al., 2009; Naito et al., 2009; Hollister and Gaut, 2011; Bousios et al., 2016; Stuart et al., 2016). In addition, TEs carry their own promoters and induce the expression of nearby genes by reading or transcription (Feschotte, 2008; Naito et al., 2009; Pereira et al., 2009). In maize, a transposon located 58.7–69.5-kb upstream of domesticated gene *teosinte branched1* (*tb1*) was an enhancer of gene expression, which explained the reason why the dominance of maize root tip was higher than its progenitor gene (Studer et al., 2011). In oil palm, the methylation deletion of *karma* transposon in the *DEFICIENS* intron leads to the origin of covered somatic clonal variation (Kok et al., 2015; Ong-Abdullah et al., 2015).

Specially, TEs may serve as a medium for rapid adaptation, because they can quickly create genetic diversity (Oliver et al., 2013; Schrader et al., 2014). Although the evolutionary forces controlling the accumulation or removal of TE between generations are not completely clear, it is known that the activity of TE in plants is inhibited by epigenetic pathways (Slotkin and Martienssen, 2007; Bousios et al., 2016; Stuart et al., 2016). These pathways require small interfering RNA (siRNAs) to target specific TE insertion through sequence consistency (Almeida and Allshire, 2005; Hollister and Gaut, 2011; Felippes et al., 2012). siRNAs with a length of 20–30 nucleotides are cleaved by different DCL proteins from dsRNAs, and finally defined by 21–24 nucleotides size (Brant and Budak, 2018). The 24-nt siRNAs target DNA methylation such as *met1* and *ddm1* mutants that decreased TE methylation levels, with concomitantly increased the expression and activity of some TEs (Lippman et al., 2004; Jia et al., 2009; Tsukahara et al., 2009). The first discovery of TE is due to their impact on phenotype and the extent to which they are ubiquitous in the genome. Until the emergence of whole genome sequencing, our vision has expanded from the selection of several transposons to the whole genome of many species that can be used to release genome sequences. Whole genome sequencing reveals the involvement of TE at the genomic structure level, and these results provide insights into the functional degree of these TE sequences. Although TEs play an important role in the formation of genomic structure and phenotypic variation, TEs in a positive selection state in natural plant populations are largely unknown. Forests cover most of the world's land surface and play a vital role in the evolution of biodiversity and function of forest ecosystem (Su et al., 2018). Poplar is one of the main woody plant model systems. Therefore, it is very important to study which genes are regulated by TE insertions to promote its rapid adaptation to climate change (Pecinka et al., 2010; Cavrak et al., 2014; Pietzenuk et al., 2016). The natural phenotypic variation of *Populus tomentosa* is distributed in the vast geographical area of northern China (30°N–40°N, 105°E–125°E), indicating poplar adaptive evolution (Du et al., 2018). In recent studies, DNA sequencing technology promotes the assessment of genetic diversity and provides a useful basis for natural populations, indicating forest genetic resources conservation is imperative (Su et al., 2018). Here we investigated TE abundance, 24-nt-siRNA targeting, and their important roles on gene expression level in *Populus trichocarpa*. Then, we identified polymorphic TEs from a representative *P. tomentosa* population, and illustrated TEs differential expansion in different climatic regions. We identified one Helitron type TE insertion which is a likely candidate to playing a role in adaptive evolution, and its insertion in 3′UTR could repress gene expression. Overall, our study highlights the potential impact of TEs on the adaptive evolution of natural population in poplar.

MATERIALS AND METHODS

More detailed information on the materials and methods used in this study were provided in **Supplementary Method 1**.

Identification and Annotation of Polymorphic Transposable Element Sites

The *de novo* annotation of TEs using RepeatModeler v2.0.1 with the parameters "ncbi" and "LTRStruct" to generate TE family consensus sequences. We then excluded unknown TE consensus sequences by TEclass v2.1.3 with default parameters (Abrusan et al., 2009). Remaining consensus sequences were assigned to defined TE superfamilies for next steps. The identified repeats in *Populus trichocarpa* (Tuskan et al., 2006), *Populus tomentosa* (CRA000903), *Populus alba* × *Populus glandulosa* (84K; Qiu et al., 2019), and *Populus alba* (Liu et al., 2019) genomes were appended to RepBase (RM Database; Version: 2017-01-27) and Tandem Repeat Finder (TRF; v4.09), resulting to be annotated by RepeatMasker v4.0.6[1] with using ncbi blast version 2.10.0+ as blast engine (Camacho et al., 2009). To reduce false positives, we retained alignments of sequence identity over 80%, as described previously (Hollister and Gaut, 2011; Niu et al., 2019). Polymorphic TEs were identified using TEPID (Stuart et al., 2016) based on the raw reads of 435 resequencing genomes of *P. tomentosa*, resulting to be divided into three subpopulations.

Clustering Analysis of *Populus tomentosa* Accessions

In 2011, 435 unrelated individuals of *P. tomentosa* were asexually propagated in Guan County, Shandong Province, China (36°230N, 115°470E) (Du et al., 2018), representing almost the entire species natural distribution (30–40°N, 105–125°E). These individuals are divided into distribution areas in the South (S), Northwest (NW), and Northeast (NE; Huang, 1992). Due to the randomness and repeatability of data, we used Python script to randomly draw individuals among different climatic regions. Finally, raw paired-end reads of 87 accessions nearly representing *P. tomentosa* entire natural range were used in this study, including 23, 35, and 29 from NE, NW, and S, respectively (**Supplementary Table 2**). The raw reads were processed to clean reads and then mapped to the *P. tomentosa* reference genome using BWA (version: 0.1.17) with default parameters

[1]http://www.repeatmasker.org

(Li and Durbin, 2009). Single nucleotide polymorphisms (SNPs) were invoked using the Genome Analysis Toolkit (GATK version: 4.0) (DePristo et al., 2011) with a mass value of 25 as the threshold. Then use EIGENSOFT (version: 7.2.1) for principal component analysis (PCA) with minor allele frequency (MAF) $\geq$ 0.05 (Price et al., 2006).

Detection of Adaptive Transposable Element Insertions

The average number of pairwise differences at each locus between any two sequences, π (Nei, 1987), was used for nucleotide diversity calculations. In addition, after screening the deletion values, Tajima's D of the neutral test was estimated using genotype data (Tajima, 1989). Based on genetic polymorphism data, population differentiation Statistics (F_{st}) were performed for each paired region differentiation with a 5,000-bp step size (Danecek et al., 2011). More details on the statistics of F_{st}, π, and Tajima's D calculations are provided in **Supplementary Method 1**, and 5,000-bp were always used as step-length. We used the first 5% of the empirical distribution of F_{st} and π values in the polymorphic region as candidates to represent the characteristics of significant difference selection between each polymorphic region in the subgroup (popNE vs. popNW, popNE vs. popS, and popNW vs. popS). Selection sweeps were evaluated using a 500-bp non-overlapping windows for reduction of diversity (ROD) statistical adjustment method, and an extensive 10-kb flanking region of each polymorphic TE region (Xu et al., 2011). A significant low π value (low ratio value) is positive selection index of TE insertion allele which were considered as putatively candidates. Then, Tajima's D was estimated in 20-kb flanking regions for the candidate adaptive TEs. All polymorphic TEs in the selective sweep regions were identified as adaptive TE insertion under selection. Details were described in **Supplementary Method 1**.

Identification of Transcription Factor Binding Sites and Enriched Motifs

Transcription factor binding sites (TFBS) were predicted by PlantPAN 2.0 (Chow et al., 2016) and PlantTFDB 4.0 (Jin et al., 2017). AME integrated in the MEME suite was used for motif enrichment analysis (Bailey et al., 2015). For enrichment analysis, Genomic regions were randomly selected as the control background for enrichment analysis, and only motifs with $P < 0.05$ (Fisher exact test) were regarded as significant. Each *cis*-element was evaluated for positional bias using two common approaches: the clustering factor (*CF-score*) (FitzGerald et al., 2004), and a recently proposed index, the *Z-score* (FitzGerald et al., 2004; Ma et al., 2013).

Identification of 24-nt siRNA Loci

The 24-nt siRNA loci were re-mapped and compared using Pln24NT database and method described by Liu et al. (2017). Non-redundant 24-nt siRNA sequences using Bowtie v1.2 (-v 1 -a -m 50 –best –strata) mapping to the reference genome (Langmead, 2010). We used ShortStack v3.8.4 to select siRNA clusters with a minimum coverage of 10 reads (Axtell, 2013). For 24-nt siRNA clusters, if they exist in a 150-bp window, they are merged to generate the final set of 24-nt siRNA loci (El Baidouri et al., 2015). Furthermore, we mapped siRNA loci to genomic locations and labeled at least one 24-nt siRNA matched single TE as siRNA+; TEs without matching siRNA loci were labeled as siRNA−.

Gene Expression of *Populus trichocarpa* Under Heat Stress

Total RNAs were extracted from 1-year-old *P. trichocarpa* at 0, 4, 8, 12, 24, 36, and 48 h after exposure to 40°C heat treatment with three technical replicates per sample. The method for estimating the genes expression levels by fragments per kilobase of exon per million (FPKM) reads mapped is described in Trapnell et al. (2012). In total, the expression levels of 17,003 genes were measured across different treatment time points. The expressed datasets used in this study were published by our previous study with accession number CRA001776 available at the BIGD Genome Sequence Archive[2].

Real-Time Reverse Transcription Polymerase Chain Reaction Analyses

When cDNA was synthesized using superscript II reverse transcriptase (Invitrogen), about 2.0 mg of total RNA was used. 1.0 ml of cDNA, diluted twice fold, and analyzed with SYBR premix ex Taq (Takara). Delta-Delta CT quantitative method was used to assess the difference between repetitions. The ACTIN gene (*Actin1*, Accession number: EF145577) and 18S were used as internal controls. PCR conditions included an initial denaturation step at 95°C for 3 min, then 40 cycles at 95°C for 30 s, 1 min at 60°C, 30 s at 72°C, and finally an extension of 5 min at 72°C. Primers used for Real-time reverse transcription polymerase chain reaction (RT-qPCR) in our study are listed in **Supplementary Table 7**.

Transient Luciferase Activity Assays

We amplified full-length *WRKY18* DNA, full-length *WRKY18* without the 3′UTR, and the full-length *WRKY18* without the Helitron from *P. tomentosa* clone "1316" using the primers listed in **Supplementary Table 7**. All amplified fragments were cloned into pCAMBIA1301H, where the *WRKY18* genes were driven by its promoter region (2-kb fragments of upstream of the translational start site). The *WRKY18* 3′UTR deletion mutants (Del1, Del2) were generated from the full-length *WRKY18* 3′UTR with the primers. Details were described in **Supplementary Method 1**. *Arabidopsis* protoplast preparation and transient expression assay were performed as previously described (Yoo et al., 2007). The luciferase (LUC) assay and the method of calculating the relative expression levels were detailed describe in **Supplementary Method 1**.

RESULTS

Distribution of Transposable Elements Among Four Poplar Whole Genomes

To study the genetic variation and evolutionary relationship of TEs in *Populus*, we investigated four poplar species,

[2]https://bigd.big.ac.cn

resulting in 162,517∼212,595 TE insertions (average covering a total of 167.98 Mb; **Supplementary Table 1**). Additionally, transposons were observed to be enriched in the centromere region, and the distribution of TEs and genes were negatively correlated on chromosome arms across the four studied species (**Supplementary Figures 1A,B**). No significant differences of TE and gene density were detected between these species, with a median TE density of 22-kb per 100-kb window across the chromosome [Mann–Whitney *U* test (MWU), $P = 0.064$], and gene density averaged 0.33 (MWU, $P = 0.087$). Meanwhile, the median TE density of *P. trichocarpa* chromosome arms was 24-kb per 100-kb window (0.24), illustrating that more *P. trichocarpa* genes are closer to TEs than that of the other three genomes (MWU, $P < 0.01$; **Supplementary Figure 2E**). Notably, we detect more genes harboring TE insertions in *P. trichocarpa* than the other three genomes (11.06%; 8.48–10.81%), demonstrating that TE insertions might contribute to the diversification of gene expression in *Populus*. Detailed analyses revealed that 3.92% (or 1,392) of *P. trichocarpa* genes have a TE insertion located within 5′ and 3′ of the coding region, with a larger proportion than the other species (0.45–0.62%) (**Supplementary Figures 2A,B**; Fisher's exact test; $P < 0.01$), suggesting that TEs insertions had a higher potential of affecting genes in *P. trichocarpa*. In addition, three super-families of elements, Gypsy, Copia and Helitron, showed a particularly striking accumulation in the genic regions of *P. trichocarpa* (**Supplementary Figures 2A–D**). Especially, Helitron elements had higher frequency in introns and exons (**Supplementary Figure 2C**), suggesting that they may be under strong purifying selection.

To study the genetic effect of genomic TEs, we first identified 25,321 conserved TEs among the four studied species (**Supplementary Data 2**). Among them, 84.03–95.72% (22,218 on average) TEs were highly enriched in intergenic regions (**Supplementary Data 3–6**), of which 29.43–31.09% were located within 2-kb genes flanking regions. Notably, 1,283 (18.76–19.75%) were classified as Gypsy super-family, 1,109 (16.28–16.99%) were classified Copia super-family, and a large number of DNA transposons, especially Helitron elements (3,014; 44.46–45.87%; **Figure 1** and **Supplementary Data 3–6**). Of these DNA transposons, 13.89–15.70% were located within promoters and 11.40–12.61% were in 2-kb-downstream of genes (**Figures 1A,B**). Additionally, we found that Ty1/Copia (26.27–27.78%) and Gypsy (26.03–26.79%) families were frequently located within genes, compared to other TE families (**Supplementary Figures 2**, **3**), indicating that LTR might potentially contribute to the diversification of gene expression in *Populus*. Moreover, an explosive of evolution time indicated by LTR of 18 different species revealed evidence for any species- or section-specific family members (**Figure 1C**), demonstrating that the LTR complement correlated with the divergence of these species. Furthermore, gene ontology enrichment analysis of collinear genes affected by TEs suggested that these genes were involved in stress-responsive (GO:0016301), programmed cell death (GO:0012501), apoptosis (GO:0006915), cell death (GO:0008219), and death (GO:0016265) were particularly affected by these TEs (**Supplementary Figure 4**). These findings suggested that TEs, the lineage-indistinctive components of genomes, might drive the diversification of gene regulation.

Transposable Elements and Small RNAs Contribute to the Divergence of Gene Expression

To explore TEs regulation on nearby genes expression, we used the publicly available RNA-seq data. Therefore, the average gene expression level increased with the distance from the nearest TE, and the gene expression level reached the maximum when the nearest TE distance is 1-kb (**Figure 2A**). Under high temperature treatment, the difference dissipated within a distance of ∼1-kb from the gene across different treatment time points, which confirmed that TEs were more likely to be active under stress treatments. Additionally, the expression at 0 h is significantly lower than 36 h (Student's *t*-test; $P < 0.01$), indicating that temperature could affect transposon activity (**Figure 2**). Remarkably, the expression levels of genes with TEs were uniformity ablated and began to stabilize at 1-kb (**Figure 2**).

One of the many factors affecting TEs accumulation near genes is siRNA guided transcriptional gene silencing. To test this possibility, 7,868,825 24-nt siRNA sequences were clustered to 56,840 siRNA loci, of which 44,735 loci (∼78.70%) perfectly matched the TE sequence. We observed that the expression level was only fluctuated significantly for TEs-24-nt siRNA+ (Student's *t*-test; $P < 0.01$), and not for TEs-24-nt siRNA− (Student's *t*-test; $P = 0.3$). At 0 h, the average expression levels of genes with flanking TEs-24-nt siRNA+ were twofold higher than TEs-24-nt siRNA− ($P < 0.05$). The overall pattern is clear: proximal TEs are related to gene expression level. When TEs are targeted by 24-nt siRNAs, the decline or increasing of expression is statistically supported.

Regulatory Motifs Highly Enriched in Conserved Transposons

To investigate the TE insertions putative regulatory mechanisms, we investigated 25,681 TEs insertions with 246 enriched motifs which were found in promoters of protein encoding genes among the four studied poplar species (**Supplementary Data 8**). The typical CAAT box and TATA box are very important for transcription initiation and occur in all promoter regions. In addition, six motifs were found to be enriched within 2-kb of the transcription start sites (TSS; $P < 0.01$; **Supplementary Figure 5**), revealing that decaying TE sequence might provide *cis* regulatory elements. Top-scoring *cis*-elements in TE insertion regions, such as the WRKYGQK motif (CGTTGACTWWDDYWDWNHH), CACG sequences (AAAGTCAACGN), ethylene-responsive elements (YREGIONNTPRB1B), and W-box (WBOXATNPR1), all having a peak position within 200-bp upstream of TSS, a high *Z-score* (>3) and/or *CF* score, or are highly overrepresented compared with random sequences reaffirming some extent of conservation in plant promoter architecture and TE insertions (**Supplementary Figure 5** and **Supplementary Data 8**). Further analysis identified that the motifs were associated with functions of defense response (GO:0006952), and response to chitin (GO:0010200) (**Supplementary Data 7**).

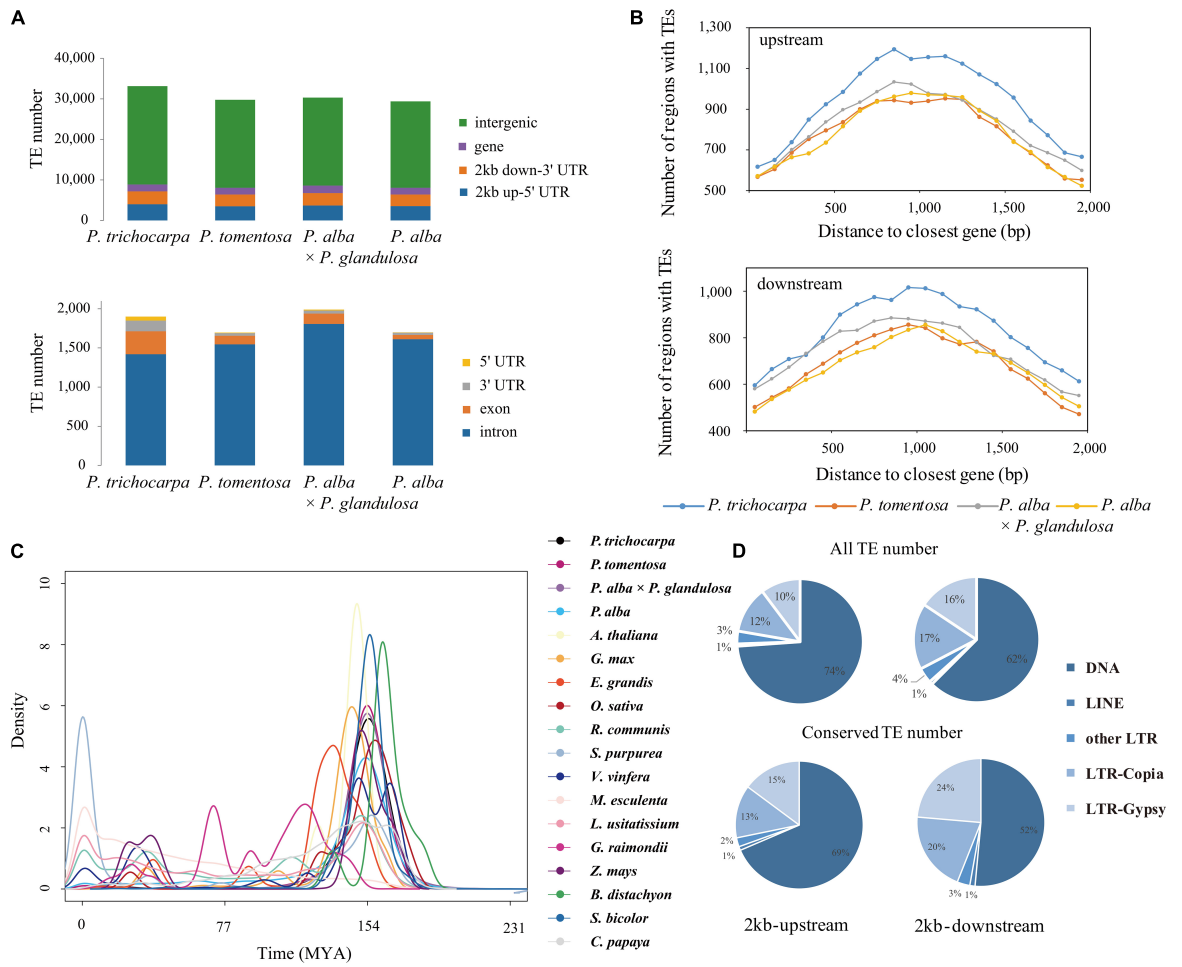


FIGURE 1 | Distribution of transposable elements in *Populus*. **(A)** Number of TEs in different genic regions of *Populus trichocarpa, P. tomentosa, P. alba* × *P. glandulosa* (84K), and *P. alba* genomes. **(B)** TEs enriched in 100-bp bins in the 2-kb upstream and 2-kb downstream region of genes between collinear gene pairs of *P. trichocarpa, P. tomentosa,* 84K, and *P. alba*. The *x*-axis represents the distance from TEs to the start codon (stop codon) of the nearest gene; the *y*-axis represents the number of TEs genes at different distances from the start codon (stop codon). **(C)** Density distributions of the Ks values for LTR transposons of 18 different species. **(D)** Distribution of transposable elements in *P. trichocarpa* of all TEs and conserved TEs (collinear TEs among four poplar genomes) in 2-kb upstream and 2-kb downstream regions.

Similarly, conserved TEs are more abundant in genomic neighborhoods of stress response and immune-related genes (**Supplementary Figure 4**). For instance, we uncovered a intronic Copia-containing the gene Potri.001G152100, which is involved in endoplasmic reticulum membrane organization that was considered as stress responses (Ascencio-Ibáñez et al., 2008). In addition, Helitron DNA transposons were more enriched with TFs than others, such as WRKY, MYB, and AP2/ERF families (**Figure 3B** and **Supplementary Data 9**). Likewise, 29.52% Helitron transposons insertions within LRR-NB-ARC related genes, which confer broad resistance to biotic and abiotic stresses, are associated with transcript truncation and could therefore cause increased susceptibility. Additionally, we measured the trend of different TE families abundance within 24-nt siRNAs from 0∼5-kb distance to nearby genes (**Figure 3A**). Intriguingly, only Helitron DNA transposons had a downward trend with the increasing distance and reached peak at 1-kb, indicating that Helitron transposons made the most significant contribution to donate *cis*-elements and regulated nearby genes with 24-nt siRNAs targeted (**Figures 2**, **3**). Conversely, other TE families with 24-nt siRNAs targeted might affect gene regulation by TEs copy and movement. These findings indicated that the insertion of TEs, especially Helitron, affected the expression of adjacent genes by providing new regulatory signals.

Polymorphic Transposable Elements Insertions Affect Natural Variation in Three Sub-Populations From Different Climatic Zones of *Populus tomentosa*

Based on the inconsistency between mapping distance and read pair insertion size, we used short-read whole-genome resequencing data of *P. tomentosa* accessions representing three sub-populations, resulting in a total of 9,680 polymorphic TE

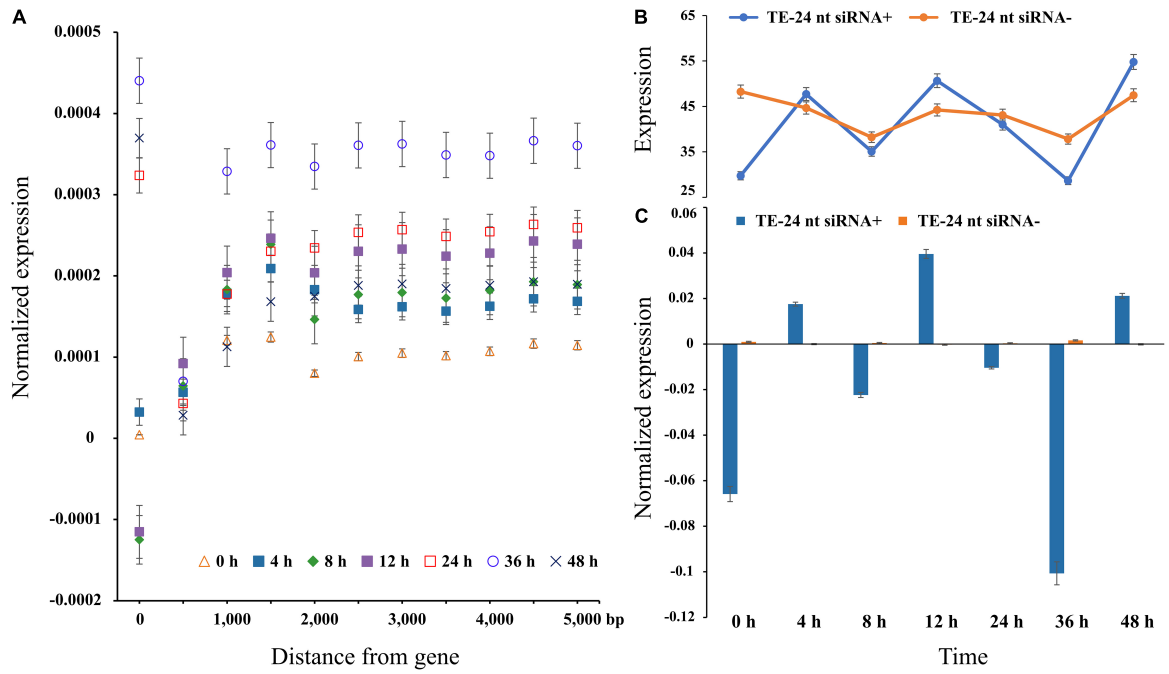


FIGURE 2 | Gene expression level in each bin farther away from the nearest TE in *P. trichocarpa*. **(A)** The average value of gene expression was TE distance, with 500-bp bin and max to 5000-bp. A distance of zero indicates TEs in intron or UTRs. **(B,C)** Gene expression levels of TEs (distances binned in 1000-bp) with and without matching 24-nt siRNAs.

loci (**Supplementary Table 2**, **Supplementary Figures 6A**, **7A**, and **Supplementary Data 10**). Further, we validated 12 identified TE loci by PCR sequencing by using 12 accessions from three subpopulations (**Supplementary Table 6**), leading in a U-shaped of TEs frequency distribution which supported by a previous study (Robert et al., 2015). Significant correlations between genetic distance and polymorphic TE insertions number were detected in our study (Person correlation; $r = -0.25$, $P > 0.01$; **Figure 4A**), among these 5,370 loci were shared. Among all three subpopulations, 3,621, 5,037, and 6,208 polymorphic TE loci in popNE, popNW, and popS, respectively (**Supplementary Data 11–13**). Additionally, we identified 213, 685, and 1,401 population-specific TE loci in popNE, popNW, and popS, respectively (**Supplementary Table 3**), proving that popS had the highest number of polymorphic TEs, population-specific TEs and the highest number of inserted TEs (**Supplementary Table 3** and **Supplementary Data 13**). Furthermore, popNE, popNW, and popS are generally consistent in the composition of polymorphic TE types and the polymorphic TE shared among them. Most of the polymorphic loci are distributed in the intraspecific or specific population level intergenic regions, but nearly 10% of the polymorphic loci still exist in the gene regions of coding sequences or introns (**Figure 4B**). Totally, 245 polymorphic TE insertions distributed in CDS region of 194 genes (**Supplementary Data 14**), which were enriched in defense response, response to stresses, RNA biosynthetic process, and immune response (GO enrichment analysis, FDR < 0.05). Compared with the TE located at intergenic regions, the polymorphic TEs inserted into CDS regions and untranslated regions (UTRs) were significantly biased to low frequency (frequency ≤ 0.1) (**Figure 4C**; Fisher's exact test, multiple testing corrected $P < 0.01$ for TEs in CDS regions and UTR regions), indicating that the diffusion of TE insertions in CDS and UTR regions are limited by purification selection (**Figure 4D** and **Supplementary Table 4**). Of the 9,680 polymorphic TEs, the proportion of DNA transposon-type TEs (46.02%) is substantially higher than that of Gypsy-type TEs (11.39%) and Copia-type TEs (11.85%). Additionally, DNA transposon-type TEs (26.84%) were the most enriched in popNE-specific polymorphic TEs, whereas Helitron-type TEs (20.73%) and other LTR-type (26.55%) were the major components in popNW, and popS-specific polymorphic TEs, respectively (**Figure 4E** and **Supplementary Table 4**). Overall, the expansion and contraction of different TEs types differentiated among the three *P. tomentosa* sub-populations.

Helitron Inserted at 3′UTR Repressed *WRKY18* Expression

Transposable element loci with selective advantage in a specific environment could spread in the population and accelerate its adaptation. Our aim was to identify the adaptive TE in different climatic regions of China (**Supplementary Figure 7B**). Then, we screened for adaptive TEs in sliding windows regions of each population (**Supplementary Figures 6**, **8A–C**, **Supplementary Data 15**, and **Supplementary Method 1**). According to the first 5% of the empirical distribution of the logarithmic ratio ($\pi_{region\_1}/\pi_{region\_2}$) and the population differentiation

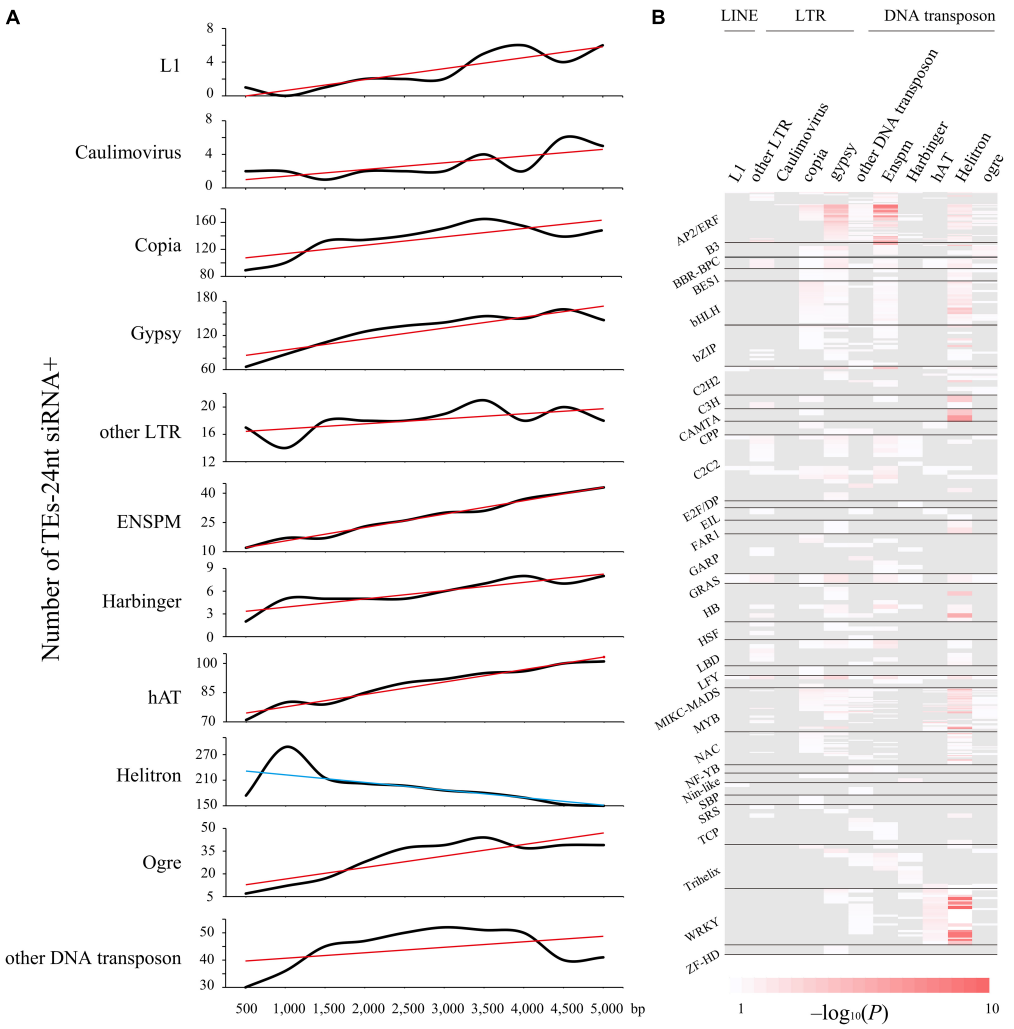


FIGURE 3 | Helitron transposons drive 24-nt siRNAs as a supply of TFBS. **(A)** The trend of quantity variation of TEs with matching 24-nt siRNAs. TEs distance is in 500-bp bin, up to 5,000-bp. Red and blue lines represent the increasing and declining trend, respectively. **(B)** Distribution peaks of *cis*-regulatory elements enriched in the different types of TEs, showing the distribution profile of motifs from 31 transcription factor families significantly ($P < 0.05$) enriched in the Helitron type transposons.

Statistics (F_{st}) value of each paired comparison between climatic regions, 4, 137, and 231 TE sites were obtained, respectively (**Supplementary Data 16**). Among TE sites, 1, 24, and 45 TE sites had significant low π values (2-kb around regions) in popNE, popNW, and popS, respectively (**Supplementary Table 5**), suggesting that the TE insertion alleles might be positive selection targets. In order to further confirm that the identified TEs are positive targets, we screened Tajima's D values of SNP sites in 20-kb regions surrounding each candidate TE (10-kb upstream and 10-kb in downstream of each TE; **Supplementary Method 1**). Finally, two adaptive TE candidates having significantly higher or lower Tajima's D values in their flanking 20-kb regions compared the target population with the reference population (**Figures 5A–C** and **Supplementary Table 5**), indicating that these two adaptive TEs showed higher haplotype homozygosity in TE insertion alleles than those without TEs.

As **Figure 5** shown, an adaptive TE inserted 837-bp in 3′UTR of *PtoWRKY18* in sweep regions of chr6-22–23 Mb. To evaluate whether TE insertion had an effect on the expression level of *PtoWRKY18*, we transiently expressed the 3′UTR of *PtoWRKY18* in protoplasts using a dual luciferase reporter assay (**Figure 5D**). The results showed that the complete 3′UTR (Helitron) and the shortened 3′UTRs containing the Helitron (Del-1 and Del-2) significantly repressed luciferase expression level, but the 3′UTR lacking the Helitron (Del-Helitron) had no repression effect (**Figure 5D**). Constructs with mut-1 (an unrelated sequence) or mut-2 (another retrotransposon) showed an approximately 2-fold increase in the relative LUC/GUS level of Helitron compared to constructs with a complete 3′UTR (**Figure 5D**), confirming that the Helitron transposon repressed gene expression. Another Helitron DNA transposon insertion at the 3′UTR in accession #3601 is uniquely present in popNW, and harbored in the

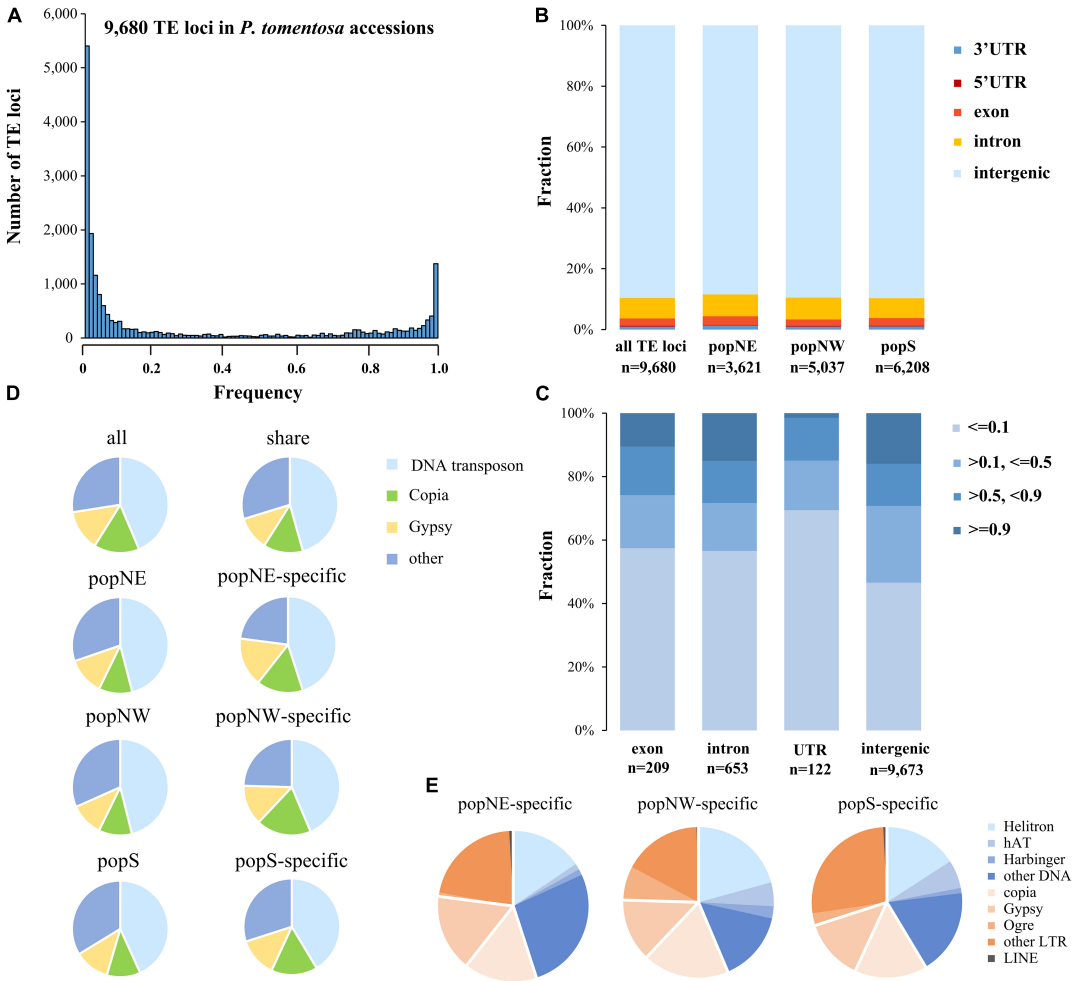


FIGURE 4 | Frequency and distribution of polymorphisms TEs in *P. tomentosa* population. **(A)** Frequency distribution of polymorphic TEs in the genomes of *P. tomentosa* natural population. The *x*-axis represents the number of additions with the same TE insertion. **(B)** Distribution pattern of polymorphic TEs in the genome. **(C)** Frequency of polymorphic TEs located in different genomic regions. **(D)** The composition of polymorphic TEs in different groups (including all groups, shared groups, and specific groups). **(E)** Composition of different types of polymorphic TEs in different specific sub-populations.

upstream of *PtoWRKY28*, which is preferentially expressed in flower bud and mature leaves, encoding a protein involving in defense response (Babitha et al., 2013; Chen et al., 2013; **Supplementary Figures 9A–C**).

DISCUSSION

Transposable Elements as a Contributor of Regulatory Elements

A large number of studies have shown that TEs can directly affect the regulation of nearby gene expression in a variety of ways, including transcriptional level and post transcriptional level (Kobayashi et al., 2004; Studer et al., 2011; Chuong et al., 2017). Similarly with our study, gene expression increased with increasing distance from the nearest TE (**Figure 2A**), and the Helitron inserted into 3′UTR significantly repressed nearby gene expression (**Figure 5D**). However, these mechanisms were found in the laboratory by mutation analysis caused by single TE insertion (Studer et al., 2011; Sun et al., 2014). Actually, it is now clear that in the distant past, the same changes have taken place in gene structure and expression level, which have been preserved by natural selection (Brosius, 2003; Marino-Ramirez et al., 2005; Gonzalez et al., 2009). Jordan et al. (2003) reported that nearly 25% of the experimentally characterized human promoters contain TE derived sequences, including empirically defined *cis* regulatory elements. Further genome-wide analysis showed that there were many promoters in the human genome (Bousios et al., 2016) and mouse genes (van de Lagemaat et al., 2003) are derived from primate specific and rodent specific TE sequences, respectively. Hence, the insertion of these TEs may help to establish lineage specific patterns of gene expression (Marino-Ramirez et al., 2005). Further evidence that TEs generally obtained regulatory function comes from TE fragments that are highly conserved in mammals; however, whether TEs

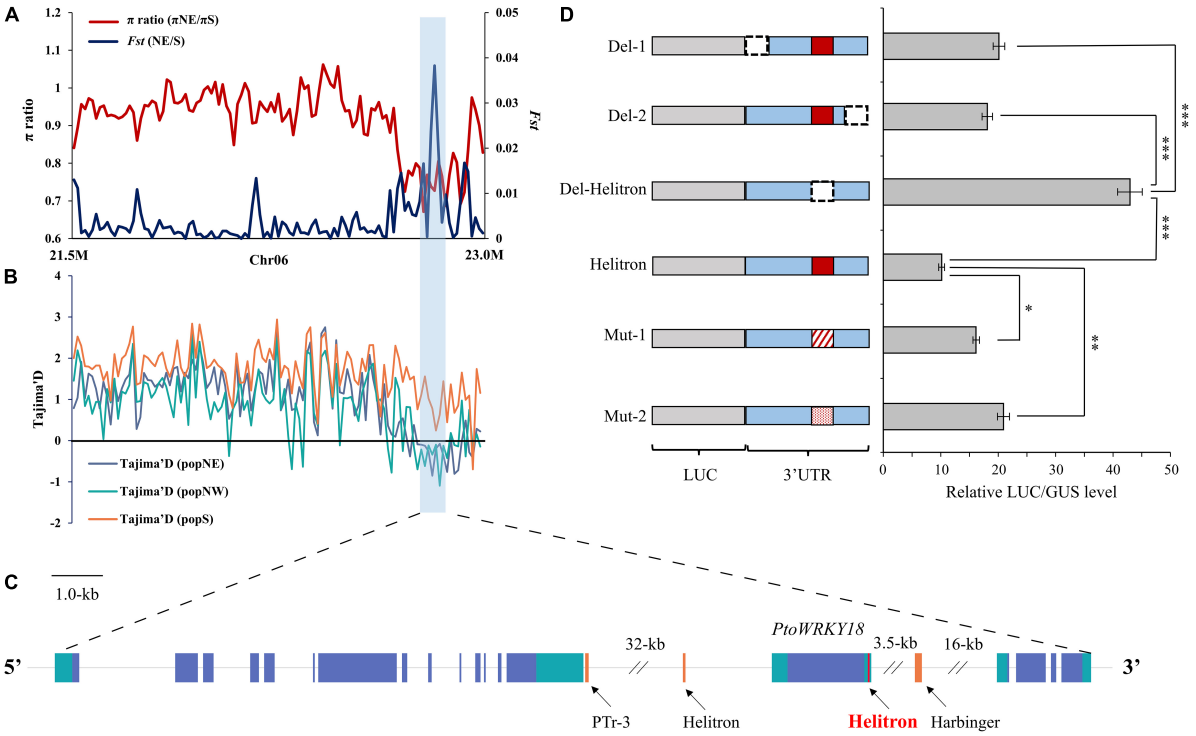


FIGURE 5 | An adaptive TE within selection regions among the three *P. tomentosa* natural climatic regions. **(A–C)** Adaptive TEs with a selection signal between each pairwise comparison. The values of F_{st}, π, and Tajima's D were plotted using a sliding window of 500-bp. Target population (Orange), and reference population (blue). The horizontal dotted line represents the corresponding average value in all selected scanning areas. Genes at the bottom (blue rectangle, coding sequences; black line, introns and intergenic regions; green rectangle, 5′ and 3′ untranslated regions; red rectangle, Helitron insertion TEs; orange rectangle, other TEs). **(D)** Left panel, firefly luciferase reporter constructs fusing the full-length *WRKY18* 3′-UTR (Helitron), deletion variants (Del-1, Del-2, and Del-Helitron), and substitute variants (Mut-1, and Mut-2). Luciferase gene was dark gray with diagonal; 3′UTR is blue, the deleted area is represented by dotted line, and the variant is represented by red line with diagonal and point; The red rectangle in the 3′UTR represents the Helitron transposon. Right panel, relative luminous units and transcription levels quantified by qPCR (Helitron normalized to 1). The data in the right panel represent the mean ± standard deviation, $n = 3$. The *, ** and *** indicate significant difference at $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively (Student's *t*-test). The data show the relative LUC/GUS level calculated by three independent replications.

performed the same function in plants was rarely reported. In our study, The conserved TEs were overexpressed in the predicted *cis* regulatory module, that was, a genomic fragment containing a dense array of TFBSs; such as WBOXATNPR1, which related *WRKY18* was found at a high frequency. Therefore, we hypothesize that these elements tend to cluster around genes that were involved in development and biological pathways of transcriptional regulation. Approximately one-tenth of homologous genes, thousands of which were derived from conserved TEs, overlapped with stress-responsive *cis*-elements that regulated downstream functional genes, which implied that they provided promoter sequences or TFBS for regulatory elements (**Supplementary Data 8**). More and more highly conserved TEs had been recorded as transcriptional enhancers (Bejerano et al., 2006; Santangelo et al., 2007).

Whether regulatory elements produced *de novo* by some mutations or pre-existing in TE sequences, TEs had been a profuse source of new regulatory sequences. The dispersion of the extended TE family throughout the genome may allow the recruitment of the same regulatory motifs at many chromosomal locations and the introduction of multiple genes into the same regulatory network (Slotkin and Martienssen, 2007; Warnefors et al., 2010; Agren et al., 2014). Although our research suggested a correlation with potential adaptive significance, future research should explore whether TEs could drive adaptive evolution in population of woody plants and confer an adaptive advantage.

Expression Levels of Transposable Elements, 24-nt siRNA and Adjacent Genes

Previous studies had showed that TEs in 3′UTRs could alter gene expression level *via* the regulation of methylation in *Arabidopsis* (Le et al., 2015; Li et al., 2018) and could change the translation efficiency in rice (Naito et al., 2009; Shen et al., 2017). Similarly, here we found that a TE insertion in the 3′UTR affected *WRKY18* expression level (**Figure 5**). In addition, the TEs and 24-nt siRNAs distributions were positively correlated with each other in *P. trichocarpa* (**Figures 2**, **3A**), and it is essential to initiate and maintain DNA methylation during TE insertions (Feschotte et al., 2002; Liu et al., 2004; Jia et al., 2009; Casacuberta and Gonzalez, 2013). The gene expression level increased with the distance to the nearest TE, and this relationship was stronger

when the TE closest to a gene was siRNA+ (**Figure 2**). The results of the expression model under heat stress were confirmed by comparing whether there were different genes in the presence of nearby TEs whether to be active under heat stress (**Figure 2**). It has been suggested that decreased gene expression is a direct consequence of TE insertion. Indeed, new insertions of a miniature inverted repeat transposable element (MITE) insertion-*mping* in rice actually enhanced gene expression (Naito et al., 2009), suggesting that the impact may vary by taxon, TE family and individual TE. Nonetheless, for all families shown in **Figure 3A**, the distribution tendency of TEs-24 nt siRNA+ were almost consistent (although not always significantly so), with the exception of Helitron transposons. Therefore, Helitron transposons made the most significant contribution to regulate nearby genes with 24-nt siRNAs targeted. Additionally, the relationship between TE proximity and gene expression level varied under high temperature treatment; in our analysis, gene expression under heat stress seems to be more sensitive to the proximity of TEs (**Figure 2A**). For example, it is not clear whether it reflects greater robustness of gene expression under normal conditions, or perhaps stresses could enhance TE activity. Further study should explore whether high temperature or other stresses affect gene stability by changing nearby TEs activity.

Transposable Elements Promote Adaptive Evolution in Natural Population of *Populus*

Transposable elements are considered to be a rapid adaptation factor because they can produce rich genetic variation in a limited time (Le Rouzic et al., 2007) and can affect phenotypic variation (Martin et al., 2009). In our study, we discussed the interpretation of the evolution of transposable factors and their impact on the host genome from different angles. On the one hand, we investigated the transposon landscape in four poplar genomes and evaluated its dynamic characteristics. This gives us an in-depth understanding of the dynamics of transposon distribution and the selection force that forms the transposon pattern. On the other hand, we deeply studied a special case of poplar, that is, adaptive TEs insertion with selective advantage in a specific environment, which may spread in the population and accelerate the adaptation of organisms. TEs are the main source of genomic mutations and, like any environmental mutagen, occasionally lead to beneficial changes (Chuong et al., 2017). From the perspective of evolution, it is important to study two aspects of TEs, namely, their evolutionary dynamics in the genome and their contribution to adaptation in global climate change.

Because TEs are an important source of genetic variation, they can promote the evolution of organisms in many ways, such as obtaining coding ability and changing coding sequence (Finnegan, 1989; Marino-Ramirez et al., 2005; Agren et al., 2014), and effecting the gene expression level (Wessler et al., 1995; Cui and Fedoroff, 2002; Studer et al., 2011). Specially, TE insertions or mutations may affect adaptation to the environment (Aravin et al., 2007; Joly-Lopez et al., 2012; Casacuberta and Gonzalez, 2013), and TE insertions or mutations that had beneficial effects on the adaptation of natural populations might become fixed. In our study, we used a representative *P. tomentosa* population which distributed a wide geographical area of northern China (30°N–40°N, 105°E–125°E) to explore the relationship between genetic diversity and adaptive evolution. However, the number of natural populations had decreased significantly, mainly due to environmental change, would further threat to the genetic resources. In this study, we found a Helitron TE locus probably become a target of positive choice and therefore contribute to adaptation of *P. tomentosa*. Consistent with that, TEs in 3′UTRs can change gene expression *via* regulating DNA methylation in *Arabidopsis* (Saze and Kakutani, 2014) and can decrease translation efficiency in rice (Shen et al., 2017). In addition, a Helitron-induced RabGDIα variant could cause quantitative recessive resistance to maize rough dwarf disease (Liu et al., 2020). Other studies also found Helitrons, upon heat shock, upregulated of nearby genes in *Caenorhabditis elegans* and promoted adaption on heat environment (Garrigues et al., 2019). Future researches could explore whether this phenotypic diversity will eventually bring adaptation advantages, and whether TEs could promote adaptation to climate change.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: Raw data of RNA-seq and resequencing are available for download at the BIGD Genome Sequence Archive (https://bigd.big.ac.cn) under accession numbers CRA001776 and CRA000903, respectively.

AUTHOR CONTRIBUTIONS

DZ designed the experiments, obtained funding, and was responsible for this article. YZ collected and analyzed the data and wrote the manuscript. YZ, XL, JX, WX, SC, XZ, SL, and JW performed the experiments. YE-K revised the manuscript and provided valuable suggestions to the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by the Major Science and Technology Projects of Inner Mongolia Autonomous Region (2021ZD0008), Project of the National Natural Science Foundation of China (Nos. 31872671 and 32170370), and 111 Project (No. B20050).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.814718/full#supplementary-material

Supplementary Figure 1 | Transposons and genes show negative correlation distribution on *P. trichocarpa* chromosomes.

Supplementary Figure 2 | Transposons distributed on *P. trichocarpa* chromosomes.

Supplementary Figure 3 | Distribution of TEs in four *Populus* genomes.

Supplementary Figure 4 | GO-term analysis of genes with transposons.

Supplementary Figure 5 | Six regulatory motifs significantly enriched in the TEs drive promoter regions.

Supplementary Figure 6 | Flowchart of the polymorphic TE.

Supplementary Figure 7 | Diversity in *P. tomentosa* population.

Supplementary Figure 8 | Adaptive selection signals within polymorphic TE insertions among the three *P. tomentosa* natural climatic regions.

Supplementary Figure 9 | A adaptive TE within selection regions among the three *P. tomentosa* natural climatic regions.

Supplementary Method 1 | Supplementary methods for this study.

REFERENCES

Abrusan, G., Grundmann, N., Demester, L., and Makalowski, W. (2009). TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25, 1329–1330. doi: 10.1093/bioinformatics/btp084

Agren, J. A., Wang, W., Koenig, D., Neuffer, B., Weigel, D., and Wright, S. I. (2014). Mating system shifts and transposable element evolution in the plant genus *Capsella*. *BMC Genomics* 15:602. doi: 10.1186/1471-2164-15-602

Almeida, R., and Allshire, R. C. (2005). RNA silencing and genome regulation. *Trends Cell Biol.* 15, 251–258. doi: 10.1016/j.tcb.2005.03.006

Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318, 761–764. doi: 10.1126/science.1146484

Ascencio-Ibáñez, J. T., Sozzani, R., Lee, T.-J., Chu, T.-M., Wolfinger, R. D., Cella, R., et al. (2008). Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiol.* 148, 436–454. doi: 10.1104/pp.108.121038

Axtell, M. J. (2013). ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 19, 740–751. doi: 10.1261/rna.035279.112

Babitha, K. C., Ramu, S. V., Pruthvi, V., and Mahesh, P. (2013). Co-expression of AtbHLH17 and AtWRKY28 confers resistance to abiotic stress in *Arabidopsis*. *Transgen. Res.* 22, 327–341. doi: 10.1007/s11248-012-9645-8

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49.

Barron, M. G., Fiston-Lavier, A. S., Petrov, D. A., and Gonzalez, J. (2014). Population genomics of transposable elements in *Drosophila*. *Annu. Rev. Genet.* 48, 561–581. doi: 10.1146/annurev-genet-120213-092359

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., et al. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441, 87–90. doi: 10.1038/nature04696

Bousios, A., Diez, C. M., Takuno, S., Bystry, V., Darzentas, N., and Gaut, B. S. (2016). A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. *Genome Res.* 26, 226–237. doi: 10.1101/gr.193763.115

Brant, E. J., and Budak, H. (2018). Plant small non-coding RNAs and their roles in biotic stresses. *Front. Plant Sci.* 9:38. doi: 10.3389/fpls.2018.01038

Brookfield, J. F. Y. (2005). The ecology of the genome - Mobile DNA elements and their hosts. *Nat. Rev. Genet.* 6, 128–136. doi: 10.1038/nrg1524

Brosius, J. (2003). The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* 118, 99–116. doi: 10.1007/978-94-010-0229-5_1

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST plus: architecture and applications. *BMC Bioinform.* 10:421.

Casacuberta, E., and Gonzalez, J. (2013). The impact of transposable elements in environmental adaptation. *Mol. Ecol.* 22, 1503–1517. doi: 10.1111/mec.12170

Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M., and Mittelsten Scheid, O. (2014). How a retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genet.* 10:e1004115. doi: 10.1371/journal.pgen.1004115

Chen, X., Liu, J., Lin, G., Wang, A., Wang, Z., and Lu, G. (2013). Overexpression of *AtWRKY28* and *AtWRKY75* in *Arabidopsis* enhances resistance to oxalic acid and Sclerotinia sclerotiorum. *Plant Cell Rep.* 32, 1589–1599. doi: 10.1007/s00299-013-1469-3

Chow, C. N., Zheng, H. Q., Wu, N. Y., Chien, C. H., Huang, H. D., Lee, T. Y., et al. (2016). PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.* 44, D1154–D1160. doi: 10.1093/nar/gkv1035

Chung, H., Bogwitz, M. R., Mccart, C., Andrianopoulos, A., Ffrench-Constant, R. H., Batterham, P., et al. (2007). Cis-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* 175, 1071–1077. doi: 10.1534/genetics.106.066597

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71–86. doi: 10.1038/nrg.2016.139

Cui, H., and Fedoroff, N. V. (2002). Inducible DNA demethylation mediated by the maize Suppressor-mutator transposon-encoded TnpA protein. *Plant Cell* 14, 2883–2899. doi: 10.1105/tpc.006163

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491. doi: 10.1038/ng.806

Du, Q., Yang, X., Xie, J., Quan, M., Liang, X., Lu, W., et al. (2018). Time-specific and pleiotropic quantitative trait loci coordinately modulate stem growth in *Populus*. *Plant Biotechnol. J.* 17, 608–624. doi: 10.1111/pbi.13002

El Baidouri, M., Do Kim, K., Abernathy, B., Arikit, S., Maumus, F., Panaud, O., et al. (2015). A new approach for annotation of transposable elements using small RNA mapping. *Nucleic Acids Res.* 43:e84. doi: 10.1093/nar/gkv257

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.* 5, 954–965. doi: 10.1093/gbe/evt025

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinform.* 9:18. doi: 10.1186/1471-2105-9-18

Felippes, F. F., Wang, J. W., and Weigel, D. (2012). MIGS: miRNA-induced gene silencing. *Plant J.* 70, 541–547. doi: 10.1111/j.1365-313x.2011.04896.x

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. doi: 10.1038/nrg2337

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* 3, 329–341. doi: 10.1038/nrg793

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5, 103–107. doi: 10.1016/0168-9525(89)90039-5

FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A., and Vinson, C. (2004). Clustering of DNA sequences in human promoters. *Genome Res.* 14, 1562–1574. doi: 10.1101/gr.1953904

Garrigues, J. M., Tsu, B. V., Daugherty, M. D., and Pasquinelli, A. E. (2019). Diversification of the *Caenorhabditis* heat shock response by helitron transposable elements. *Elife* 8:e51139. doi: 10.7554/eLife.51139

Gonzalez, J., Macpherson, J. M., and Petrov, D. A. (2009). A recent adaptive transposable element insertion near highly conserved developmental loci in *Drosophila melanogaster*. *Mol. Biol. Evol.* 26, 1949–1961. doi: 10.1093/molbev/msp107

Hollister, J. D., and Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc. Natl. Acad. Sci. USA* 108:2322. doi: 10.1073/pnas.1018222108

Huang, Z. (1992). The Study on the Climatic Regionalization of the Distributional Region of *Populus tomentosa*. *J. Beijing Fores. Univ.* 14, 26–32.

Jia, Y., Lisch, D. R., Ohtsu, K., Scanlon, M. J., Nettleton, D., and Schnable, P. S. (2009). Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. *PLoS Genet.* 5:e1000737. doi: 10.1371/journal.pgen.1000737

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Joly-Lopez, Z., Forczek, E., Hoen, D. R., Juretic, N., and Bureau, T. E. (2012). A Gene Family Derived from Transposable elements during early *Angiosperm* evolution has reproductive fitness benefits in *Arabidopsis thaliana*. *PloS Genet.* 8:e1002931. doi: 10.1371/journal.pgen.1002931

Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72. doi: 10.1016/S0168-9525(02)00006-9

Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science* 304:982. doi: 10.1126/science.1095011

Kok, S. Y., Ong-Abdullah, M., Cheng-Lian, G. E. E., and Namasivayam, P. (2015). A histological study of oil palm (Elaeis guineensis) endosperm during seed development. *J. Oil Palm Res.* 27, 107–112.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinform.* 11:1117. doi: 10.1002/0471250953.bi1107s32

Le, T. N., Miyazaki, Y., Takuno, S., and Saze, H. (2015). Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana*. *Nucleic Acids Res.* 43, 3911–3921. doi: 10.1093/nar/gkv258

Le Rouzic, A., Boutin, T. S., and Capy, P. (2007). Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. USA* 104, 19375–19380. doi: 10.1073/pnas.0705238104

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Z. W., Hou, X. H., Chen, J. F., Xu, Y. C., Wu, Q., Gonzalez, J., et al. (2018). Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. *Genome Biol. Evol.* 10, 2140–2150. doi: 10.1093/gbe/evy171

Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., Mccombie, W. R., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476. doi: 10.1038/nature02651

Liu, J., He, Y., Amasino, R., and Chen, X. (2004). siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes Dev.* 18, 2873–2878. doi: 10.1101/gad.1217304

Liu, Q., Deng, S., Liu, B., Tao, Y., and Xu, M. (2020). A helitron-induced RabGDIα variant causes quantitative recessive resistance to maize rough dwarf disease. *Nat. Commun.* 11:14372. doi: 10.1038/s41467-020-14372-3

Liu, Q., Ding, C., Chu, Y., Zhang, W., Guo, G., Chen, J., et al. (2017). Pln24NT: a web resource for plant 24-nt siRNA producing loci. *Bioinformatics* 33, 2065–2067. doi: 10.1093/bioinformatics/btx096

Liu, Y. J., Wang, X. R., and Zeng, Q. Y. (2019). *De novo* assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China. *Sci. China Life Sci.* 62, 609–618. doi: 10.1007/s11427-018-9455-2

Ma, S. S., Shah, S., Bohnert, H. J., Snyder, M., and Dinesh-Kumar, S. P. (2013). Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. *PloS Genet.* 9:e1003840. doi: 10.1371/journal.pgen.1003840

Marino-Ramirez, L., Lewis, K. C., Landsman, D., and Jordan, I. K. (2005). Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* 110, 333–341. doi: 10.1159/000084965

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., et al. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461, 1135–1138. doi: 10.1038/nature08498

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801. doi: 10.1126/science.15739260

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., et al. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130–1134. doi: 10.1038/nature08479

Nei, M. (1987). *Molecular evolutionary genetics*. New York, NY: Columbia University Press.

Niu, X. M., Xu, Y. C., Li, Z. W., Bian, Y. T., Hou, X. H., Chen, J. F., et al. (2019). Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc. Natl. Acad. Sci. USA* 116, 6908–6913. doi: 10.1073/pnas.1811498116

Oliver, K. R., Mccomb, J. A., and Greene, W. K. (2013). Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.* 5, 1886–1901. doi: 10.1093/gbe/evt141

Ong-Abdullah, M., Ordway, J. M., Jiang, N., Ooi, S. E., Kok, S. Y., Sarpan, N., et al. (2015). Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 525, 533–537. doi: 10.1038/nature15365

Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., and Scheid, O. M. (2010). Epigenetic regulation of repetitive elements Is attenuated by prolonged heat stress in *Arabidopsis*. *Plant Cell* 22, 3118–3129. doi: 10.1105/tpc.110.078493

Pereira, V., Enard, D., and Eyre-Walker, A. (2009). The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One* 4:e4321. doi: 10.1371/journal.pone.0004321

Pietzenuk, B., Markus, C., Gaubert, H., Bagwan, N., Merotto, A., Bucher, E., et al. (2016). Recurrent evolution of heat-responsiveness in *Brassicaceae* COPIA elements. *Genome Biol.* 17:209. doi: 10.1186/s13059-016-1072-3

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Qiu, D., Bai, S., Ma, J., Zhang, L., Shao, F., Zhang, K., et al. (2019). The genome of *Populus alba* x *Populus tremula* var. *glandulosa* clone 84K. *DNA Res.* 26, 423–431. doi: 10.1093/dnares/dsz020

Robert, K., Viola, N., Christian, S. T., and Petrov, D. A. (2015). Tempo and Mode of Transposable Element Activity in *Drosophila*. *PLoS Genet.* 11:e1005406. doi: 10.1371/journal.pgen.1005406

Santangelo, A. M., De Souza, F. S. J., Franchini, L. F., Bumaschny, V. F., Low, M. J., and Rubinstein, M. (2007). Ancient exaptation of a CORE-SINE Retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PloS Genet.* 3:1813–1826. doi: 10.1371/journal.pgen.0030166

Saze, H., and Kakutani, T. (2014). Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *Embo. J.* 26, 3641–3652. doi: 10.1038/sj.emboj.7601788

Schrader, L., Kim, J. W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., et al. (2014). Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.* 5:5495. doi: 10.1038/ncomms6495

Shen, J., Liu, J., Xie, K., Xing, F., Xiong, F., Xiao, J., et al. (2017). Translational repression by a miniature inverted-repeat transposable element in the 3' untranslated region. *Nat. Commun.* 8:14651. doi: 10.1038/ncomms14651

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285. doi: 10.1038/nrg2072

Stuart, T., Eichten, S. R., Cahn, J., Karpievitch, Y. V., Borevitz, J. O., and Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* 5:e20777. doi: 10.7554/eLife.20777

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* 43, 1160–1163. doi: 10.1038/ng.942

Su, J., Yan, Y., Song, J., Li, J., Mao, J., Wang, N., et al. (2018). Recent fragmentation may not alter genetic patterns in endangered long-lived species: evidence from *Taxus cuspidata*. *Front. Plant Sci.* 9:1571.

Sun, W., Shen, Y. H., Han, M. J., Cao, Y. F., and Zhang, Z. (2014). An adaptive transposable element insertion in the regulatory region of the EO gene in the domesticated silkworm, Bombyx mori. *Mol. Biol. Evol.* 31, 3302–3313. doi: 10.1093/molbev/msu261

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Thiel, T., Graner, A., Waugh, R., Grosse, I., Close, T. J., and Stein, N. (2009). Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol. Biol.* 9:209. doi: 10.1186/1471-2148-9-209

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* 461, 423–426. doi: 10.1038/nature08351

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

van de Lagemaat, L. N., Landry, J. R., Mager, D. L., and Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19, 530–536. doi: 10.1016/j.tig.2003.08.004

Warnefors, M., Pereira, V., and Eyre-Walker, A. (2010). Transposable elements: insertion pattern and impact on gene expression evolution in hominids. *Mol. Biol. Evol.* 27, 1955–1962. doi: 10.1093/molbev/msq084

Wessler, S. R., Bureau, T. E., and White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* 5, 814–821. doi: 10.1016/0959-437x(95)80016-x

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982.

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., et al. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111. doi: 10.1038/nbt.2050

Yoo, S. D., Cho, Y. H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* 2, 1565–1572. doi: 10.1038/nprot.2007.199

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

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

frontiers in Plant Science

REVIEW
published: 24 February 2022
doi: 10.3389/fpls.2022.822217

Landscape Genomics in Tree Conservation Under a Changing Environment

Li Feng[1†] *and Fang K. Du*[2*†]

[1] School of Pharmacy, Xi'an Jiaotong University, Xi'an, China, [2] School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China

Understanding the genetic basis of how species respond to changing environments is essential to the conservation of species. However, the molecular mechanisms of adaptation remain largely unknown for long-lived tree species which always have large population sizes, long generation time, and extensive gene flow. Recent advances in landscape genomics can reveal the signals of adaptive selection linking genetic variations and landscape characteristics and therefore have created novel insights into tree conservation strategies. In this review article, we first summarized the methods of landscape genomics used in tree conservation and elucidated the advantages and disadvantages of these methods. We then highlighted the newly developed method "Risk of Non-adaptedness," which can predict the genetic offset or genomic vulnerability of species *via* allele frequency change under multiple scenarios of climate change. Finally, we provided prospects concerning how our introduced approaches of landscape genomics can assist policymaking and improve the existing conservation strategies for tree species under the ongoing global changes.

Keywords: changing environment, genotype-environment associations (GEAs), landscape genomics, local adaptation, tree conservation

OPEN ACCESS

Edited by:
Jue Ruan,
Chinese Academy of Agricultural Sciences (CAAS), China

Reviewed by:
Jun Chen,
Zhejiang University, China
Chengjun Zhang,
Kunming Institute of Botany (CAS), China
Suhua Shi,
Sun Yat-sen University, China

***Correspondence:**
Fang K. Du
dufang325@bjfu.edu.cn

[†]***ORCID:***
Li Feng
orcid.org/0000-0002-8252-9463
Fang K. Du
orcid.org/0000-0002-7377-5259

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *25 November 2021*
Accepted: *10 January 2022*
Published: *24 February 2022*

Citation:
Feng L and Du FK (2022) Landscape Genomics in Tree Conservation Under a Changing Environment. Front. Plant Sci. 13:822217. doi: 10.3389/fpls.2022.822217

INTRODUCTION

Forest trees cover *ca.* 30% of the terrestrial surface of the earth from boreal to tropical latitudes and contain approximately three-quarters of the terrestrial biomass of the earth, which tightly links them with the global carbon cycle (Holliday et al., 2017; Isabel et al., 2020). They generally have higher levels of genetic diversity and experience rapid microevolution, which often show distinguishable adaptation to local environments (Hamrick et al., 1992; Petit and Hampe, 2006; Neale and Kremer, 2011). In addition, quantitative traits with high heritability make trees exhibit stronger signals of local adaptation (clinal variation); however, a large genome, long generation time makes it not suitable for quantitative trait loci (QTL) and related analysis even though great progress have been achieved on quantitative genetics study on trees (Savolainen et al., 2013; Milesi et al., 2019 and references therein). Therefore, understanding the genetic basis of adaptation to the environment *via* landscape genomics studies is essential for management interventions of tree species related to conservation and reforestation under climate change (Allendorf et al., 2010; Savolainen et al., 2013; Anderson and Song, 2020).

Empirical studies had already suggested that adaptation in tree species primarily arises from standing genetic variations, facilitating more rapid adaptation to climate change than that *via*

new mutations (Barrett and Schluter, 2008; Alberto et al., 2013; Savolainen et al., 2013). However, rapid climate change can break this association and create a mismatch between population climatic optima and current climate (Jump and Peñuelas, 2005; Aitken et al., 2008). Additional challenges such as gene flow, eco-evolutionary dynamics on the species range margins, and variation in climate changes across the landscape may also impact the adaptation of species (Savolainen et al., 2007; Alberto et al., 2013; Aitken and Bemmels, 2016). Still, in practice, very few conservation strategies consider genetic resources, especially for forest species (Lefèvre et al., 2013), with some applications in forest restoration program, e.g., it is recommended that *Quercus mongolica* seeds should not be transferred from their provenances because the genetic cline was determined between the northeastern and southwestern Japan by neutral genetic markers (Nagamitsu and Shuri, 2021).

Great progress had been achieved on the topic of the adaptive potential of natural populations (Hoffmann et al., 2017; Gaitán-Espitia and Hobday, 2021). In recent years, integrated interdisciplinary methods such as landscape genetics or genomics which are used to disentangle the impacts of environmental conditions on forest trees might provide guides for forest conservations (Rellstab et al., 2015; Isabel et al., 2020). The classical method for detecting the genetic basis of adaptation relies on population genetics (Wright and Gaut, 2004). This method attempts to find out outlier single-nucleotide polymorphisms (SNPs) by the comparisons of the genetic differentiation (F_{ST}) between populations and to hypothesize that these outliers are most likely to be affected by natural selection (Excoffier et al., 2009; Hohenlohe et al., 2010). However, this method suffers a high ratio of false-positive results due to the ignorance of environmental heterogeneity (Eveno et al., 2008; Hoban et al., 2016). Another approach is landscape genomics, which uncovers the molecular mechanism of adaptation on the basis of the genotype-environment associations (GEAs) by integrating genetic variation and spatial models (Holderegger et al., 2006; Sork et al., 2013; Sork, 2017; De Lafontaine et al., 2018). Recently, the evaluation of the genomic vulnerability (Bay et al., 2018), genetic offset (Fitzpatrick and Keller, 2015), or risk of non-adaptedness (RONA) (Rellstab et al., 2016) was used to predict the climate-driven population shifts. At present, a growing number of studies focus on tree species already utilized the population-level genomic data to evaluate the genomic vulnerability of the species in a changing climate.

The next-generation sequencing makes landscape genomic studies currently possible for detecting adaptive signals and uncovering the genomic basis of adaptation in many organisms. Although landscape genomics has been pursued for a decade and the advances of theoretical frameworks and applications are promising (Schoville et al., 2012; Joost et al., 2013; Bragg et al., 2015; Ćalić et al., 2015; Rellstab et al., 2015; Capblancq et al., 2020), molecular ecologists and evolutionists are currently awash with data, and the analytical methods in landscape genomics have lagged behind.

As an emerging approach used for the conservation genetics of trees, it is essential to understand its advanced trend. However, the existing approaches belong to landscape genomics for detecting adaptive signatures and predicting genetic offset of adaptive allelic frequencies under multiple climates have different assumptions, advantages, and limitations. An effective integrative framework shortage and how to utilize the results from these variable methods to improve management interventions of forest trees are big challenges for landscape genomics studies in the genomic era. Therefore, we first surveyed recent literature on the landscape genomics approach used for tree conservation study. We checked the Original Journal articles in the *Molecular Ecology*, *Evolutionary Applications*, *Global Change Biology*, *New Phytologist*, *Ecology Letters*, and *Nature Climate Change* from 2015 to 2021 (**Table 1**). Publications were selected based on four criteria: (i) the research was performed on forest tree species; (ii) an SNP dataset was used; (iii) adaptive SNPs were detected, and (iv) articles must predict the optimal composition to the future climate to evaluate genetic offset. Second, we summarized and depicted the advantages and disadvantages of utilizing related methods and genomic tools involved in detecting GEAs (**Table 2**) to quantify and/or map the disruption in local adaptation of forest trees under climate change. Then, we established a general framework (**Figure 1**) integrated methods of landscape genomics and population genomics for local adaptation analysis in forest trees. Finally, we provided suggestions on how these approaches can be used in making conservation strategies for tree species under climate change.

EXISTING APPROACHES OF LANDSCAPE GENOMICS

Mixed-Effects Models

Mixed-effects models provide a unified analytical framework to indicate robust and powerful evidence for adaptation (Rellstab et al., 2015). The advantage of the mixed-effects model is that it can reduce false-positive results by considering the influence of pairwise genetic distances and population structure. In mixed models, the genetic structure is incorporated as a random factor, allele frequencies are defined as response variables, and environmental factors are used as fixed factors. In this "Mixed-effects models" section, we illuminated the principles and methodologies using mixed models to detect signals of local adaptation based on BAYENV (Coop et al., 2010), Bayesian population association analysis (BayPass) (Gautier, 2015; Olazcuaga et al., 2020), latent factor mixed models (LFMMs) (Frichot et al., 2013), and spatial analysis method (SAM) (Joost et al., 2007, 2008).

The BAYENV is a method under the Bayesian framework employed to evaluate correlations between loci and environmental variables, and it can incorporate the uncertainty of allele frequencies from uneven sample sizes (Coop et al., 2010). The advantage of this program is that it applies a covariance matrix to take account for population structure, which is similar to an F_{ST} or kinship matrix. BAYENV requires a null model based on neutral loci and then determines the covariance matrix of estimated allele frequencies across populations. The significance test of each locus-variable combination utilizes Bayes factors calculated automatically by the program. However, cautions

TABLE 1 | Short overview of recent studies of landscape genomics for forest trees.

Species	Spatial scale	Data	Adaptive signature identification	Predictive model	References	Journal
Populus balsamifera	North America	Targeted genotyping	F_{ST} outlier tests, Bayenv, GPA	GF, GDM	Fitzpatrick and Keller, 2015	Ecology letters
Populus balsamifera	North America	Targeted genotyping	LFMM, Bayenv	GDM	Gougherty et al., 2021	Nature Climate change
Quercus lobate	United States	GBS	F_{ST} outlier test, LFMM	GF	Gugger et al., 2021	Molecular Ecology
Quercus rugose	Mexico	GBS	F_{ST} outlier test, LFMM	GF, GDM	Martins et al., 2018	Molecular Ecology
Quercus spp.	Switzerland	Poolseq	LFMM	RONA	Rellstab et al., 2016	Molecular Ecology
Quercus suber	Western Mediterranean	GBS	F_{ST} outlier test, SelEstim	RONA	Pina-Martins et al., 2019	Global Change Biology
Betula nana	United Kingdom	RADseq	F_{ST} outlier test, RDA, Bayenv2	RONA	Borrell et al., 2020	Evolutionary Applications
Euptelea polyandra and *Euptelea pleiosperma*	Japan and China	RAD	F_{ST} outlier test	GF	Cao et al., 2020	Evolutionary Applications
Platycladus orientalis	China	GBS	F_{ST} outlier test, Bayenv2	GF	Jia et al., 2020	Evolutionary Applications
Quercus aquifolioides	Western China	Poolseq	F_{ST} outlier test, Bayenv, LFMM	RONA	Du et al., 2020	Evolutionary Applications
Pinus densata	Western China	Exome capture sequencing	Bayenv, Pcadapt, RDA	GF	Zhao et al., 2020	New Phytologist
Eucalyptus microcarpa	Australia	DArTseq	F_{ST} outlier tests	RONA	Jordan et al., 2017	Molecular Ecology
Corymbia calophylla	Western Australia	DArTseq	Bayenv2, LFMM	GDM	Ahrens et al., 2019	Molecular Ecology
Melaleuca rhaphiophylla and *Nuytsia floribunda*	Southwestern Australia	DArTseq	F_{ST} outlier test, LFMM	GDM	Walters et al., 2020	Molecular Ecology

DArTseq, diversity arrays technology sequencing; GBS, genotype-by-sequencing; GDM, generalized dissimilarity modeling; GF, gradient forest; GPA, genotype-phenotype association; LFMM, latent factor mixed model; Poolseq, whole-genome sequencing of pools of individuals; RADseq, restriction-site associated DNA sequencing; RDA, redundancy analysis; RONA, risk of non-adaptiveness.

TABLE 2 | Overview of methods and software available for environmental associations and genomic offset analyses in landscape genomics.

Software	Method	Purpose	Data type	Specifics and limitations	References
BAYENV, BAYPASS	Bayes	detecting GEAs	Allele frequencies and environmental variable	Less sensitive to population demography; but calibration with neutral SNPs is needed and significance thresholds need to be determined from simulated datasets.	Günther and Coop, 2013; Gautier, 2015
LFMM, R (LEA)	Bayes	detecting GEAs	Allele frequencies and environmental variable	Corrects for population structure using latent factors; but only performs association with environment.	Frichot et al., 2013; Frichot and Francois, 2015
SAMβADA, R (R.SamBada)	Spatial analysis	detecting GEAs	Allele frequencies and environmental variable	Underlying models are simple, allows correction for population structure; but possibly has high false-positive rates.	Stucki et al., 2017; Duruz et al., 2019
R (vegan)	Ordination	detecting GEAs	SNPs, environmental and geographic datasets	Finds the linear combinations of genetic and environmental datasets *via* RDA or CCA; but exists strong multicollinearity and doesn't allow missing data.	XLSTAT, 2012; Oksanen et al., 2013
R (gdm)	GDM	projecting GF	Allele frequencies, environmental and geographic datasets	Provides genomic offset based on numbers of adaptive loci simultaneously *via* distance-based method; but result should be validated by additional datasets.	Manion et al., 2014; Fitzpatrick and Keller, 2015
R (gradientForest)	RF	projecting GF	Allele frequencies and environmental variables	Provides genomic offset based on numbers of adaptive loci simultaneously *via* machine-learning algorithm; but result should be validated by additional datasets.	Ellis et al., 2012; Fitzpatrick and Keller, 2015
pyRona	SLR	projecting GF	Allele frequency and environmental variable	Provides genomic offset based on average change in allele frequency at multiple adaptive loci; but result should be validated by additional datasets.	Rellstab et al., 2016; Pina-Martins et al., 2019

CCA, canonical correlation analysis; GDM, generalized dissimilarity modeling; GEAs, genotype-environment associations; GF, genomic offset; RDA, redundancy analysis; RF, random forest; SLR, simple linear regression.

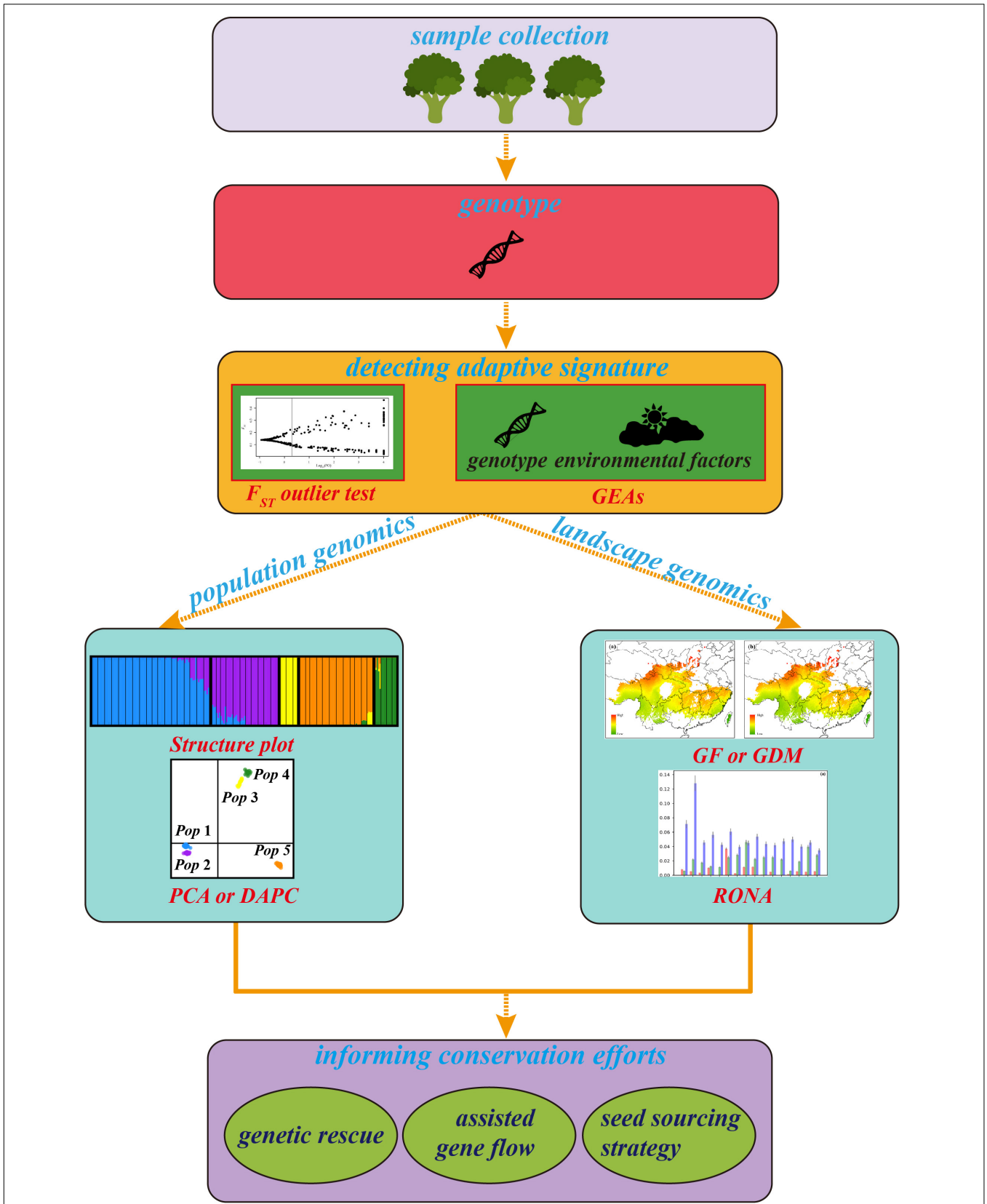


FIGURE 1 | The general framework of landscape genomics for tree conservation. The plots of cluster, F_{ST} outlier test and RONA are modified from Du et al. (2020) and Feng et al. (2020), respectively.

should be noticed here that these factors may not be directly compared across environmental variables due to variable-specific value ranges. In 2013, Günther and Coop developed an updated program called BAYENV2, which added non-parametric tests in the options and could be robustly applied for the Pool-Seq data (Günther and Coop, 2013).

BayPass (Gautier, 2015) is an extension of the Bayesian outlier detection model implemented in BAYENV to execute GEAs or environmental association analysis (EAA). It takes demographic effects into account by the estimation of the covariance matrix of allele frequency between populations. The core model (i.e., multivariate generalization model) in the BayPass reports locus XtX that is analogous to F_{ST} but explicitly corrected for this covariance matrix, accounting for the neutral correlations of allelic frequencies. Simulation studies suggested that the BayPass provided a robust framework to detect adaptive SNP signals (Gautier, 2015). However, a recent study revealed that the assumed linear relationships between allele frequencies utilized in EAA in line with the algorithm proposed by Gautier (2015) are unsatisfactory and even problematic when dealing with small datasets (Olazcuaga et al., 2020). Hence, they proposed a new approach that does not necessitate considering the uncertainty of the allele frequency estimation but assumes the exchangeability of SNPs both across the populations and along the genome. It is effective for gaining well-behaved *P*-values, avoiding intensively computational calibration, and providing reasonable numbers of SNPs analyzed (Olazcuaga et al., 2020). The advantage of BayPass is that it improves test performances by the estimations of the covariance matrix **Ω** (Olazcuaga et al., 2020).

The LFMMs rely on the Markov Chain Monte Carlo algorithms and integrate fixed effects to model environmental variables (Frichot et al., 2013). This algorithm is an extension of principal component analysis (PCA). LFMMs incorporate fixed effects to model environmental variables, and natural genetic structure is introduced as a random factor (i.e., latent factor). The computational speed is fast, and in addition, this approach does not need any *a priori* knowledge, making it attractive for determining adaptive signals with genomic data (Frichot et al., 2013; Rellstab et al., 2015). LFMMs can be implemented by the software LFMM (Frichot et al., 2013; Caye et al., 2019) or the R package *LEA* (Frichot and Francois, 2015).

The SAM is developed to assess putative associations between molecular markers and environmental variables using multiple univariate logistic regressions (Joost et al., 2007). It detects signatures of selection based on an integrative application of geographical information systems (GIS), environmental variables, and molecular data (Joost et al., 2007, 2008) implemented in MATLAB. The significance is determined by the likelihood ratio and Wald tests. Simulation studies implied that SAM might provide false-positive results if tested species endure complicated demography (De Mita et al., 2013; Frichot et al., 2013). Recently, an improved version of SAM called SAMβADA was developed (Stucki et al., 2017). This new approach allows for rapidly analyzing large genomics datasets by parallel processing. Compared with the early analysis method (i.e., SAM), the advantages of this new algorithm include that it (i) incorporates multivariate analyses to assess the impacts of many environmental predictor variables, (ii) allows to split the datasets and merges the results *via* parallel processing of SAMβADA, and (iii) enables the inclusion of explanatory variables representing population structure into the models to decrease false-positive results. However, pre- and post-processing of data will be labor-intensive when using the SAMβADA. In view of these facts, Duruz et al. (2019) published the R.SamBada landscape genomics pipeline to ease the identification and interpretation of candidate genes underlying local adaptation.

Multivariate Statistical Analysis

The multivariate statistical analysis usually integrates environmental variables and spatial genetic structure into the analytical framework to detect the adaptive variation. Traditionally, isolation by environment (IBE) is commonly used to detect selection signatures (Wright and Gaut, 2004; Wang and Bradburd, 2014; Manthey and Moyle, 2015). However, this Mantel-based method had poor performance in detecting true-positive results (Harmon and Glor, 2010; Hardy and Pavoine, 2012; Legendre et al., 2015), and its estimation bias might be amplified in the genomic era. Instead, the multivariate statistical analysis such as canonical correlation analysis (CCA) (Ter Braak, 1986) and redundancy analysis (RDA) (Van Den Wollenberg, 1977; Legendre and Legendre, 2012) may be more realistic for detecting selection signatures than univariate methods (Forester et al., 2016, 2018), because the selection is always a polygenic process driven by multiple environmental factors.

The CCA aimed to find a linear relationship between multiple loci and environmental factors. The loadings consist of loci and environmental variables indicate which loci respond to which environmental factors. However, we need caution to infer the outcomes if strong patterns of multicollinearity exist within datasets (Rellstab et al., 2015; Fenderson et al., 2020). RDA is another ordination approach that is effective to detect adaptive variation based on allele frequency data (Legendre and Legendre, 2012; Capblancq et al., 2018). First, RDA produces a matrix of fitted values based on the multivariate linear regression between genetic and environmental data, and then, the PCA of the fitted values produces canonical axes that are linear combinations of the original explanatory variables (Forester et al., 2016). In addition, partial RDA (pRDA) that stems from RDA also allows for constructing and testing complicated models to avoid the impacts of neutral genetic structure or spatial effects on detecting loci underlying adaptive variation (Legendre and Legendre, 2012). Another analogous approach, called the distance-based redundancy analysis (dbRDA), which differs in associations between genetic data and principal coordinate analysis and the procedure of emerging response variables compared with the RDA, can also enable to detect the adaptive evolution (Legendre and Anderson, 1999). However, when using the abovementioned methods, an important caveat exists is that the explanatory variables within these methods are uncorrelated and the number of loci examined might be at least three times as large as the number of putative explanatory variables (Jombart et al., 2009).

Simulation and empirical studies suggested that the RDA-based method could detect lower false-positive and higher true-positive rates when compared with generalized linear models

(GLM) or LFMM (Forester et al., 2018). Even the powers to identify adaptive loci associated with environmental variables are similar *via* RDA and LFMM, the former has the advantage to identify the main selective gradients as a combination of environmental variables (Capblancq et al., 2018). Additionally, the constrained ordination methods have robust performance and enable to avoid spurious GEAs when the tested species has isolation-by-distance (IBD) pattern or low dispersal capability (Forester et al., 2016, 2018; Capblancq et al., 2018).

PREDICTING GENOMIC VULNERABILITY UNDER ALTERNATIVE CLIMATE SCENARIOS

Traditionally, the vulnerability of species mainly relied on the prediction of species distribution models (SDMs) (Elith and Leathwick, 2009) or its extensions, such as climate-niche factor analysis (CNFA) (Rinnan and Lawler, 2019). However, the abovementioned methods are unable to account for the continuous, multidimensional nature of genomic variation (Fitzpatrick and Keller, 2015). The expanding omics and statistical tools enable us to generate robust predictions of plant adaptive potential under climate change. In this study, we introduced three new methods used for predicting genomic vulnerability under alternative climate scenarios based on the linear or non-linear functions: generalized dissimilarity modeling (GDM) (Ferrier et al., 2007), gradient forests (GFs) (Ellis et al., 2012), and RONA (Rellstab et al., 2016).

Predicting Genomic Vulnerability Using Non-linear Regressions

The GDM is used for estimating and predicting the spatial pattern of turnover in community composition (Ferrier et al., 2002, 2007). Fitzpatrick and Keller (2015) extended the application of GDM to forecast the genetic offset of *Populus balsamifera*, and they concluded that the changes of genetic composition are required if it tries to mitigate maladaptation and maintain genetic diversity in the future. GDM accounts for spatial patterns in genetic data caused by demographic processes, accommodates varied factors (e.g., geographic or ecological separation, barriers to dispersal) as predictors, and also enables to deal with numerous SNP loci (Fitzpatrick and Keller, 2015). The functions can be implemented *via* the R package *gdm* (Manion et al., 2014). A recent study evaluated the local, forward, and reverse genetic offsets (**Figure 2**) of balsam poplar using the GDM and incorporated migration and dispersal into predictive genomic models to show the adaptive potential of balsam poplar in future climates (Gougherty et al., 2021). This study provides a new way to assess population-level risk at alternative climate scenarios that accounts for local adaptation and breaks through the prediction limitations at the species level.

The GF is an extension of random forests based on the non-parametric, machine-learning regression tree approach (Ellis et al., 2012). This method enables to estimate and map the frequency changes of SNPs associated with environmental tolerance at different spatial-temporal scales (Fitzpatrick and Keller, 2015). It can be executed *via* the R package *gradientForest* (Ellis et al., 2012). Unlike GDM, GF can handle complicated

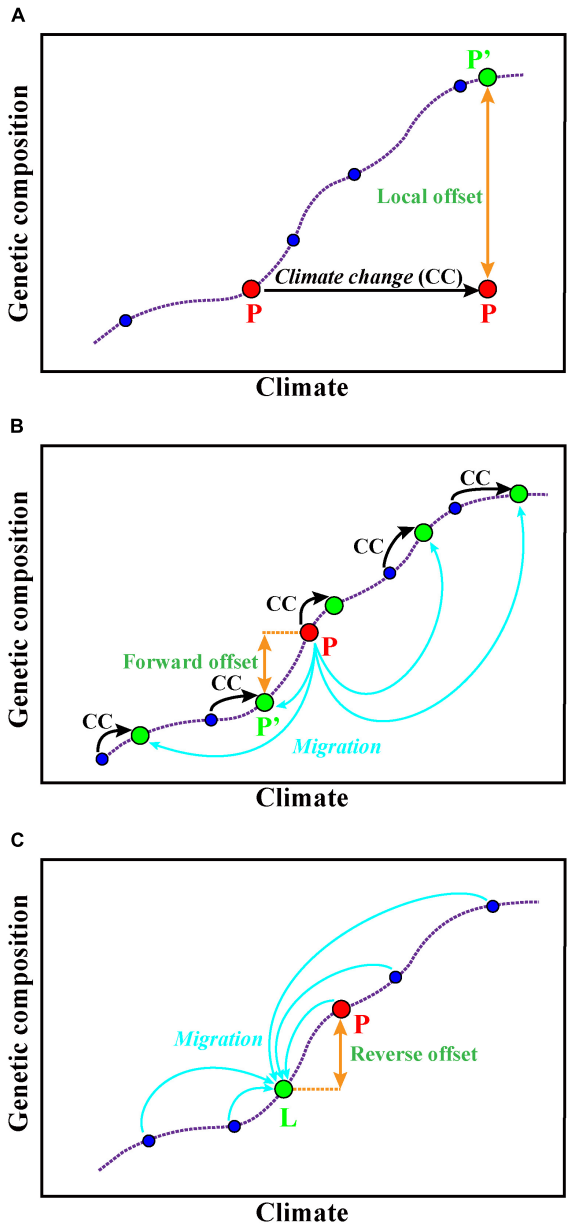


FIGURE 2 | Predictions of potential adaptation to alternative climate scenarios. **(A)** Local offset means the specific population P (red color) to the theoretically required changes of allele frequency under a future climate *in situ* [P′ (green)]. **(B)** Forward genetic offset means that a specific contemporary population P (red color) can migrate (blue arrows) to the habitat whose future climate best matches its genetic composition [P′ (green)]. **(C)** Reverse genetic offset means for a specific location L (green color) and its future climate, the minimum genetic distance of a contemporary population P (red color) to the theoretically required population for location L. The purple dotted line represents the association between the genetic composition of several populations (blue circles) and their local, contemporary climate. These figures are modified from Rellstab (2021).

associations between predictors and accommodate these correlated predictors, providing a means to determine the response of individual SNPs to environmental gradients.

Both GDM and GF can handle large genomic datasets that include numerous rare alleles and accommodate pronounced non-linearities in the exploration of GEAs, providing unprecedented insights into genome regions under local selection and predicting the changes of adaptive genomic diversity across landscape. For instance, Martins et al. (2018) revealed a strong association between the genetic variation of *Quercus rugosa* and the precipitation seasonality in Mexico *via* the GDM and GF, and they predicted that future populations of *Q. rugosa* might be at risk due to the high rate of climate change. However, considering that the actual evolutionary responses of populations to climate change will be more complex than the simplified projections based on the two abovementioned approaches, we must consider caveats when explaining the result of genetic offset arising from the GDM and GF.

Predicting Genomic Vulnerability Using Linear Regression

Rellstab et al. (2016) developed a method called RONA to evaluate genomic vulnerabilities of populations under alternative climate scenarios based on linear regressions inspired by the study of assessing the relative risk of maladaptation in Douglas fir (Bradley St Clair and Howe, 2007). RONA represents the average change in allele frequency at adaptive SNPs required to keep pace with the change of a given environmental factor in future (**Figure 3**). The average absolute difference of the changes in allele frequencies of these loci between the current and future climate conditions represents RONA under a given environmental variable (Rellstab et al., 2016; Pina-Martins et al., 2019). Recently, Borrell et al. (2020) utilized the current RONA (c-RONA, **Figure 3**) to define the average change in allele frequency at climate-associated loci required to match the estimation of the optimum for a given environmental factor, using the future RONA (f-RONA; **Figure 3**) to define the original concept of RONA proposed by Rellstab et al. (2016).

Theoretically, if the difference between the current and the prediction values is high, then more conservation efforts are needed for persisting of focal species. Empirical studies in trees and invertebrates show that if the expected allele frequency changes are less than 0.1 per decade, it might keep pace with climate change, while if the changes are greater than 0.1–0.2 per decade, it may cause a lag between allele frequency and climate adaptation (Jump, 2006; Egan et al., 2015; Jump et al., 2017). However, this simplified approach does not take gene flow and migration into account and assumes that the best model profiling the GEAs is resulted from local adaptation. Furthermore, the predictions of RONA for adaptive loci based on this method have multiple values, and each RONA arises from a given environmental variable, which will not account for the interactive effects of loci contributing to climate adaptation (Rellstab et al., 2016; Capblancq et al., 2020). Therefore, we must keep in mind that the genomic vulnerability approaches are still in their infancy and face numerous challenges and uncertainties, and they have yet to be tested and validated in real conservation applications (Rellstab, 2021).

APPLICATIONS OF LANDSCAPE GENOMICS IN TREE CONSERVATION

Landscape genomics significantly improves our understanding of ecological and evolutionary processes in tree species and offers guidelines for conservation efforts and management applications. The potential of landscape genomics for forest management is discussed in the following sections.

Using Landscape Genomics to Inform Genetic Rescue

Genetic rescue aims to increase population fitness and avoid population declines by introducing immigration of new alleles (Tallmon et al., 2004; Whiteley et al., 2015; Bell et al., 2019; Fitzpatrick and Funk, 2021). Landscape genomics studies will increase the effectiveness of genetic rescue by identifying which populations are most likely to increase fitness and population growth rate (Whiteley et al., 2015). The population with the lowest level of adaptive differentiation would be chosen in order to minimize outbreeding impression. A recent study on dwarf birch suggested that the genetic rescue should be applied for the populations with small population sizes that occurred in the margins of the distribution of species (Borrell et al., 2020). However, previous studies had indicated that genetic rescue can only improve fitness and increase population sizes in the short term rather than save imperiled populations over the long term (Whiteley et al., 2015). Additionally, if the inbreeding depression in small populations resulted from the recent effect of human-caused fragmentation, assisted migration is more appropriate than genetic rescue (Hohenlohe et al., 2021).

Using Landscape Genomics to Inform Assisted Gene Flow

Assisted gene flow (AGF) means managed translocation of individuals within the current species range to mitigate local maladaptation (Aitken and Whitlock, 2013; Aitken and Bemmels, 2016). AGF is equivalent to the genetic rescue when target populations are small and maladapted, with genetic diversity therein dominantly threatened by drift (Aitken and Whitlock, 2013). Compared with genetic rescue, AGF emphasizes the introduced alleles that are preadapted to new local environments and thus increase the frequency of these adaptive loci in existing populations. AGF has already been applied for some forest trees. For example, Browne et al. (2019) suggested that AGF might be applied to mitigate adaptation lag of temperature for California oak according to a landscape genomic survey of the species. Another fascinating case is dissecting the associations of GEAs in balsam poplar (Gougherty et al., 2021). Gougherty et al. found that the eastern populations of the balsam poplar might face the greatest vulnerability and risk of future extirpation to climate change, and the conservation efforts *via* AGF are needed for

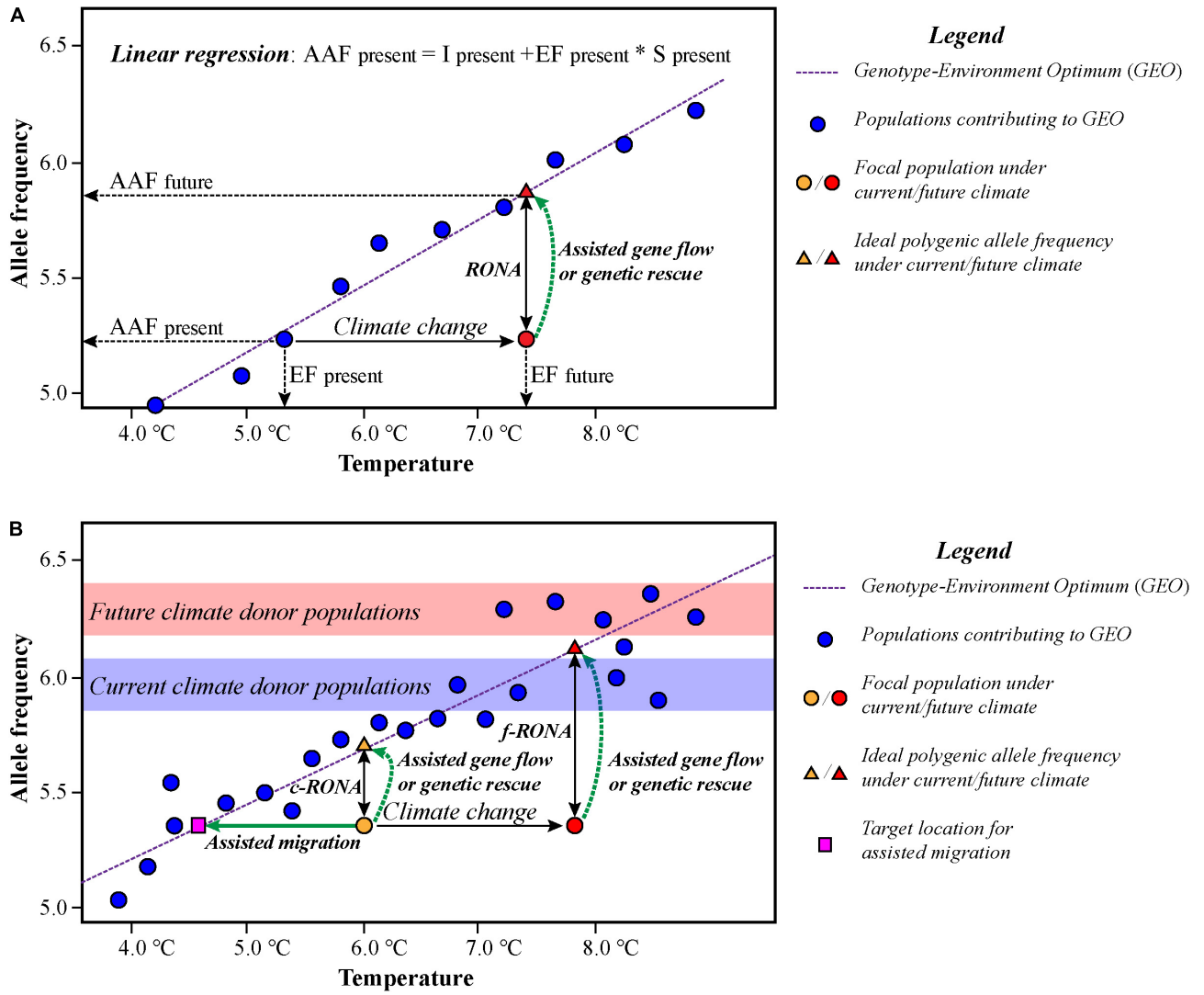


FIGURE 3 | Schematic illustration of the risk of non-adaptedness (RONA) to alternative climate change. **(A)** RONA is the average change of allele frequency in a set of adaptive loci that are required under future climate scenario according to a simple linear regression of the relationship of allele frequency and environments. AAF, alternative allele frequency; EF, environmental factor; I, intercept of the regression; S, slope of the regression. **(B)** The current and future RONA (c-RONA and f-RONA); c-RONA/f-RONA is the average change in allele frequency required under current environmental conditions. Blue and red bands indicate suitable candidate donor populations for assisted gene flow under current and future scenarios, respectively. The figures **(A,B)** are modified from Rellstab et al. (2016) and Borrell et al. (2020), respectively.

those populations by estimating the local, forward, and reverse genetic offsets of the species. However, outbreeding depression might occur if source and recipient populations are isolated for a long time (Aitken and Whitlock, 2013). Additionally, high levels of gene flow introduced by AGF might result in biotic homogenization between source and target populations and consequently prevent them from adapting to novel climate conditions (Gaitán-Espitia and Hobday, 2021).

Using Landscape Genomics to Inform Seed Sourcing Strategy

Seed sourcing strategy aims to capture the adaptive diversity and improve the adaptive potential of species under climate change and has been proposed for ecological restoration during past decades (Broadhurst et al., 2008; Breed et al., 2013, 2019). Landscape genomics is an ideal approach to inform seed sourcing strategies for species persisting. Jordan et al. (2017) detected 81 putatively adaptive SNPs in *Eucalyptus microcarpa*, and 62 of which are associated with mean annual temperature by a combination of four F_{ST} outlier tests and one EAA (i.e., BAYENV2) as the general framework of landscape genomics (**Figure 1**). They found that the expected allelic frequency changes of these adaptive SNPs in the New South Wales (NSW) populations were greater than that of other sites, suggesting that the warmer, northern end of the range (i.e., NSW) of *E. microcarpa* might not suitable for seed source. Recently, a

provenance decision-making framework proposed by Carvalho et al. (2021) offers a comprehensive perspective for seed source guidelines based on the information that arises from neutral and adaptive variation *via* integrative analyses of population genomics and landscape genomics, which can also be applied for informing seed sourcing strategy of forest species.

CHALLENGES AND FUTURE DIRECTIONS

Landscape genomics provides unprecedented insights into understanding the mechanism of adaptive variation of tree species by dissecting the impacts of environmental variables and landscape characteristics on their adaptive evolution (Browne et al., 2019; Pina-Martins et al., 2019; Borrell et al., 2020; Du et al., 2020; Zhao et al., 2020). Common challenges, such as the sampling strategies and using only a single analysis for detecting adaptive signatures, still exist in landscape genomic studies although many reviews discuss the abovementioned topics (Balkenhol et al., 2019). Instead of discussing the abovementioned common challenges in this study, we focused on the challenge of landscape genomics studies on tree species in identifying adaptive variation and their spatial patterns facing the changing climate.

First, the levels of commonality in genes or SNPs associated with climates that arose in landscape genomics studies are quite low. Although the large majority of landscape genomics studies utilize integrative methods for detecting putatively adaptive loci to illuminate the GEAs, few loci are shared between these approaches. These inconsistent patterns by different methods were detected in many studies, for example, in Mexican oak *Q. rugose* (Martins et al., 2018: only one SNP associated with temperature seasonality was identical between LFMM and BAYESCAN test) or Norway spruce in three independent landscape genomic studies across the Italian Alps sharing similar sampling areas and climates (Scalfi et al., 2014; Ćalić, 2015; Di Pierro et al., 2016: no identical adaptive genes were detected in more than two studies).

The low commonality in adaptive signals might be the evidence for lacking parallel evolution of adaptive traits in forest trees (Geraldes et al., 2014) or just because of false-positive results (Ćalić et al., 2015). Additionally, we believed that the varied methods applied for detecting adaptive SNPs have different assumptions, advantages, and limitations, which responded to the low commonality. However, the commonality levels for projecting genomic vulnerability under alternative climate scenarios using GF or GDM are relatively high. Fitzpatrick and Keller (2015) found that using GDM and GF for the projections of genomic vulnerability, the genetic offset of the circadian clock gene GIGANTEA-5 (*GI5*) associated with plant circadian clock and light perception pathways in balsam poplar is similar, although slight differences existed in its marginal area, both methods predict the range core of balsam poplar likely suffered minimal disruption of the existing GEAs. In the future, efforts by a combination of simulations, genomic data, and common garden experiments might be applied to demonstrate the high effectivity and accuracy of genomic offset under alternative climates (Fitzpatrick et al., 2021).

Second, the current studies of landscape genomics for evaluating and uncovering the adaptive variation in tree species focus only on a single species rather than at the community level (**Table 1**). Analyzing multiple species within the same landscapes makes it possible to assess the commonality of their eco-evolutionary dynamics across species and landscapes and thereby depict a thorough picture of how local adaptation is originated in nature (Bragg et al., 2015; Hand et al., 2015; Balkenhol et al., 2019). However, eco-evolutionary models require new data and methods for assessing the adaptive potential of species, which have only been possible for a few model species so far (Waldvogel et al., 2020). In addition, the present challenge of illuminating ecological adaptation at the community level is how to simulate the patterns of local adaptation of species or populations and their adaptive potential under future climate changes, while a possible way to overcome these inconveniences is integrating the prediction methods including GDM, GF, or RONA into the analytical framework of landscape community genomics.

Finally, the investigators of landscape genomics must consider the genomic sequencing strategy employed and the genomic resources available for their focal species. Prevalent sequencing methods in the landscape genomics studies of non-model species currently take advantage of reduced-representation methods [e.g., genotype-by-sequencing (GBS) and restriction-site associated DNA sequencing (RADseq)] and RNA sequencing. However, the number of SNPs obtained and the ability to detect genes underlying local adaptation from the abovementioned methods may be influenced due to the differences in library preparation, SNP densities, and the bioinformatics parameters applied to SNP filtering (Hoban et al., 2016; Lowry et al., 2017; McKinney et al., 2017). As more and more forest tree genomes have been published (e.g., Table 1 in Ingvarsson et al., 2016) and sequencing costs fall, whole-genome resequencing is thriving and becoming an option for landscape genomics studies (Lin et al., 2018; Zhu et al., 2020), which can provide unprecedented marker density and determine other genetic variation such as structural variants and mutations in regulatory elements, increasing power for the detection of local adaptation and providing novel insights into the role of selection, recombination, and gene flow in promoting or impairing local adaptation to new habitats compared with reduced-representation methods (Fuentes-Pardo and Ruzzante, 2017; Bourgeois and Warren, 2021). In addition, the degrees of linkage disequilibrium (LD) in the studied species will also influence the power of detecting adaptive SNPs. Considering the low LD rates in tree species, using these methods such as reduced-representation methods will fail to detect loci that underlie most local adaptation and adaptive phenotypic variation (Bragg et al., 2015; Hoban et al., 2016). We advocated obtaining detailed LD information of focal species using whole-genome sequencing before the studies of landscape genomics in future because the resources of reference genome are critical to fully address the issues of local adaptation (Manel et al., 2016). Moreover, prior knowledge about LD decay from the reference genome of focal species can inform sampling strategies

and sequencing selections to maximize opportunities to identify adaptive SNPs (Bragg et al., 2015).

CONCLUSION

Understanding the genetic mechanism of adaptation is the key issue for molecular ecology and evolutionary biology. We reviewed the existing theories and methods that belong to landscape genomics for detecting adaptive evolution in species and advocated utilizing an integrated analytical framework to illuminate the GEAs between genetic and environmental data. We particularly emphasized the effectivity and necessity of multiple methods for detecting signatures of local adaptation combined with models for predicting adaptation potential in tree conservation. With the low sequencing cost, ease availability of high-solution environmental data, and newly developed genomic tools in the near future, we believe that the conservation efforts and management interventions for forest trees will benefit from advancing studies of landscape genomics.

AUTHOR CONTRIBUTIONS

LF and FD conceived the study and wrote the manuscript. Both authors designed the focus, structure and content of the review.

FUNDING

This study was financially supported by the 111 Project (No. B20050) of BJFU, the National Natural Science Foundation of China (42071060 and 31901075), and the China Postdoctoral Science Foundation (2018M633490).

ACKNOWLEDGMENTS

The authors thank three reviewers and Rong Wang in the East China Normal University for useful comments on the early draft, and Keke Liu and Min Qi for their contributions on **Table 1**.

REFERENCES

Ahrens, C. W., Byrne, M., and Rymer, P. D. (2019). Standing genomic variation within coding and regulatory regions contributes to the adaptive capacity to climate in a foundation tree species. *Mol. Ecol.* 28, 2502–2516. doi: 10.1111/mec.15092

Aitken, S. N., and Bemmels, J. B. (2016). Time to get moving: assisted gene flow of forest trees. *Evol. Appl.* 9, 271–290. doi: 10.1111/eva.12293

Aitken, S. N., and Whitlock, M. C. (2013). Assisted gene flow to facilitate local adaptation to climate change. *Annu. Rev. Ecol. Evol. Syst.* 44, 367–388. doi: 10.1146/annurev-ecolsys-110512-135747

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., and Curtis-Mclane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol. Appl.* 1, 95–111. doi: 10.1111/j.1752-4571.2007.00013.x

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., et al. (2013). Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biol.* 19, 1645–1661. doi: 10.1111/gcb.12181

Allendorf, F. W., Hohenlohe, P. A., and Luikart, G. (2010). Genomics and the future of conservation genetics. *Nat. Rev. Genet.* 11, 697–709. doi: 10.1038/nrg2844

Anderson, J. T., and Song, B.-H. (2020). Plant adaptation to climate change-Where are we? *J. Syst. Evol.* 58, 533–545. doi: 10.1111/jse.12649

Balkenhol, N., Dudaniec, R. Y., Krutovsky, K. V., Johnson, J. S., Cairns, D. M., Segelbacher, G., et al. (2019). "Landscape genomics: Understanding relationships between environmental heterogeneity and genomic characteristics of populations," in *Population Genomics: Concepts, Approaches and Applications*, ed. O. P. Rajora (Cham: Springer International Publishing), 261–322. doi: 10.1186/s12868-016-0283-6

Barrett, R. D. H., and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends Ecol. Evol.* 23, 38–44. doi: 10.1016/j.tree.2007.09.008

Bay, R. A., Harrigan, R. J., Underwood, V. L., Gibbs, H. L., Smith, T. B., and Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science* 359, 83–85. doi: 10.1126/science.aan4380

Bell, D. A., Robinson, Z. L., Funk, W. C., Fitzpatrick, S. W., Allendorf, F. W., Tallmon, D. A., et al. (2019). The exciting potential and remaining uncertainties of genetic rescue. *Trends Ecol. Evol.* 34, 1070–1079. doi: 10.1016/j.tree.2019.06.006

Borrell, J. S., Zohren, J., Nichols, R. A., and Buggs, R. J. A. (2020). Genomic assessment of local adaptation in dwarf birch to inform assisted gene flow. *Evol. Appl.* 13, 161–175. doi: 10.1111/eva.12883

Bourgeois, Y. X. C., and Warren, B. H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Mol. Ecol.* 30, 6036–6071. doi: 10.1111/mec.15989

Bradley St Clair, J., and Howe, G. T. (2007). Genetic maladaptation of coastal Douglas-fir seedlings to future climates. *Global Change Biol.* 13, 1441–1454. doi: 10.1111/j.1365-2486.2007.01385.x

Bragg, J. G., Supple, M. A., Andrew, R. L., and Borevitz, J. O. (2015). Genomic variation across landscapes: insights and applications. *New Phytol.* 207, 953–967. doi: 10.1111/nph.13410

Breed, M. F., Harrison, P. A., Blyth, C., Byrne, M., Gaget, V., Gellie, N. J. C., et al. (2019). The potential of genomics for restoring ecosystems and biodiversity. *Nat. Rev. Genet.* 20, 615–628. doi: 10.1038/s41576-019-0152-0

Breed, M. F., Stead, M. G., Ottewell, K. M., Gardner, M. G., and Lowe, A. J. (2013). Which provenance and where? Seed sourcing strategies for revegetation in a changing environment. *Conserv. Genet.* 14, 1–10. doi: 10.1007/s10592-012-0425-z

Broadhurst, L. M., Lowe, A., Coates, D. J., Cunningham, S. A., Mcdonald, M., Vesk, P. A., et al. (2008). Seed supply for broadscale restoration: maximizing evolutionary potential. *Evol. Appl.* 1, 587–597. doi: 10.1111/j.1752-4571.2008.00045.x

Browne, L., Wright, J. W., Fitz-Gibbon, S., Gugger, P. F., and Sork, V. L. (2019). Adaptational lag to temperature in valley oak (Quercus lobata) can be mitigated by genome-informed assisted gene flow. *Proc. Natl. Acad. Sci. U.S.A.* 116, 25179–25185. doi: 10.1073/pnas.1908771116

ĆaliĆ, I. (2015). *Estimation of Adaptive Genetic Variation in Norway spruce (Picea abies (L.) Karst) to Climate Change*. PhD Doctoral dissertation. Florence: University of Florence-Italy.

ĆaliĆ, I., Bussotti, F., Martínez-García, P., and Neale, D. (2015). Recent landscape genomics studies in forest trees-what can we believe? *Tree Genet. Genomes* 12, 1–7. doi: 10.1007/s11295-015-0960-0

Cao, Y. N., Zhu, S. S., Chen, J., Comes, H. P., Wang, I. J., Chen, L. Y., 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. doi: 10.1111/eva.12960

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., and Keller, S. R. (2020). Genomic prediction of (mal)adaptation across current and future climatic landscapes. *Annu. Rev. Ecol. Evol. Syst.* 51, 245–269. doi: 10.1146/annurev-ecolsys-020720-042553

Capblancq, T., Luu, K., Blum, M. G. B., and Bazin, E. (2018). Evaluation of redundancy analysis to identify signatures of local adaptation. *Mol. Ecol. Resour.* 18, 1223–1233. doi: 10.1111/1755-0998.12906

Carvalho, C. S., Forester, B. R., Mitre, S. K., Alves, R., Imperatriz-Fonseca, V. L., Ramos, S. J., et al. (2021). Combining genotype, phenotype, and environmental data to delineate site-adjusted provenance strategies for ecological restoration. *Mol. Ecol. Resour.* 21, 44–58. doi: 10.1111/1755-0998.13191

Caye, K., Jumentier, B., Lepeule, J., and François, O. (2019). LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Mol. Biol. Evol.* 36, 852–860. doi: 10.1093/molbev/msz008

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–1423. doi: 10.1534/genetics.110.114819

De Lafontaine, G., Napier, J. D., Petit, R. J., and Hu, F. S. (2018). Invoking adaptation to decipher the genetic legacy of past climate change. *Ecology* 99, 1530–1546. doi: 10.1002/ecy.2382

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

Di Pierro, E. A., Mosca, E., Rocchini, D., Binelli, G., Neale, D. B., and La Porta, N. (2016). Climate-related adaptive genetic variation and population structure in natural stands of Norway spruce in the South-Eastern Alps. *Tree Genet. Genomes* 12, 1–15. doi: 10.1007/s11295-016-0972-4

Du, F. K., Wang, T., Wang, Y., Ueno, S., and De Lafontaine, G. (2020). Contrasted patterns of local adaptation to climate change across the range of an evergreen oak, *Quercus aquifolioides*. *Evol. Appl.* 13, 2377–2391. doi: 10.1111/eva.13030

Duruz, S., Sevane, N., Selmoni, O., Vajana, E., Leempoel, K., Stucki, S., et al. (2019). Rapid identification and interpretation of gene-environment associations using the new R.SamBada landscape genomics pipeline. *Mol. Ecol. Resour.* 19, 1355–1365. doi: 10.1111/1755-0998.13044

Egan, S. P., Ragland, G. J., Assour, L., Powell, T. H., Hood, G. R., Emrich, S., et al. (2015). Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecol. Lett.* 18, 817–825. doi: 10.1111/ele.12460

Elith, J., and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697. doi: 10.1146/annurev.ecolsys.110308.120159

Ellis, N., Smith, S. J., and Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. *Ecology* 93, 156–168. doi: 10.1890/11-0252.1

Eveno, E., Collada, C., Guevara, M. A., Léger, V., Soto, A., Díaz, L., et al. (2008). Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Mol. Biol. Evol.* 25, 417–437. doi: 10.1093/molbev/msm272

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298. doi: 10.1038/hdy.2009.74

Fenderson, L. E., Kovach, A. I., and Llamas, B. (2020). Spatiotemporal landscape genetics: investigating ecology and evolution through space and time. *Mol. Ecol.* 29, 218–246. doi: 10.1111/mec.15315

Feng, L., Ruhsam, M., Wang, Y. H., Li, Z. H., and Wang, X. M. (2020). Using demographic model selection to untangle allopatric divergence and diversification mechanisms in the *Rheum palmatum* complex in the Eastern Asiatic Region. *Mol. Ecol.* 29, 1791–1805. doi: 10.1111/mec.15448

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. community-level modelling. *Biodivers. Conserv.* 11, 2309–2338. doi: 10.1023/A:1021302930424

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* 13, 252–264. doi: 10.1111/j.1472-4642.2007.00341.x

Fitzpatrick, M. C., and Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecol. Lett.* 18, 1–16. doi: 10.1111/ele.12376

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., and Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Mol. Ecol. Resour.* 21, 2749–2765. doi: 10.1111/1755-0998.13374

Fitzpatrick, S. W., and Funk, W. C. (2021). "Genomics for genetic rescue," in *Population Genomics: Wildlife*, eds P. A. Hohenlohe and O. P. Rajora (Cham: Springer International Publishing), 437–471. doi: 10.1007/13836_2019_64

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120. doi: 10.1111/mec.13476

Forester, B. R., Lasky, J. R., Wagner, H. H., and Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Mol. Ecol.* 27, 2215–2233. doi: 10.1111/mec.14584

Frichot, E., and Francois, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi: 10.1111/2041-210x.12382

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi: 10.1093/molbev/mst063

Fuentes-Pardo, A. P., and Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. *Mol. Ecol.* 26, 5369–5406. doi: 10.1111/mec.14264

Gaitán-Espitia, J. D., and Hobday, A. J. (2021). Evolutionary principles and genetic considerations for guiding conservation interventions under climate change. *Global Change Biol.* 27, 475–488. doi: 10.1111/gcb.15359

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201, 1555–1579. doi: 10.1534/genetics.115.181453

Geraldes, A., Farzaneh, N., Grassa, C. J., McKnown, A. D., Guy, R. D., Mansfield, S. D., et al. (2014). Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution* 68, 3260–3280. doi: 10.1111/evo.12497

Gougherty, A. V., Keller, S. R., and Fitzpatrick, M. C. (2021). Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. *Nat. Clim. Change* 11, 166–171. doi: 10.1038/s41558-020-00968-6

Gugger, P. F., Fitz-Gibbon, S. T., Albarrán-Lara, A., Wright, J. W., and Sork, V. L. (2021). Landscape genomics of *Quercus lobata* reveals genes involved in local climate adaptation at multiple spatial scales. *Mol. Ecol.* 30, 406–423. doi: 10.1111/mec.15731

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi: 10.1534/genetics.113.152462

Hamrick, J. L., Godt, M. J. W., and Sherman-Broyles, S. L. (1992). "Factors influencing levels of genetic diversity in woody plant species," in *Population Genetics of Forest Trees*, eds W. T. Adams, S. H. Strauss, D. L. Copes, and A. R. Griffin (Dordrecht: Springer), 95–124.

Hand, B. K., Lowe, W. H., Kovach, R. P., Muhlfeld, C. C., and Luikart, G. (2015). Landscape community genomics: understanding eco-evolutionary processes in complex environments. *Trends Ecol. Evol.* 30, 161–168. doi: 10.1016/j.tree.2015.01.005

Hardy, O. J., and Pavoine, S. (2012). Assessing phylogenetic signal with measurement error: a comparison of mantel tests, blomberg et al.'s K, and phylogenetic distograms. *Evolution* 66, 2614–2621. doi: 10.1111/j.1558-5646.2012.01623.x

Harmon, L. J., and Glor, R. E. (2010). Poor statistical performance of the mantel test in phylogenetic comparative analyses. *Evolution* 64, 2173–2178. doi: 10.1111/j.1558-5646.2010.00973.x

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., et al. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397. doi: 10.1086/688018

Hoffmann, A. A., Sgrò, C. M., and Kristensen, T. N. (2017). Revisiting adaptive potential, population size, and conservation. *Trends Ecol. Evol.* 32, 506–517. doi: 10.1016/j.tree.2017.03.012

Hohenlohe, P. A., Funk, W. C., and Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Mol. Ecol.* 30, 62–82. doi: 10.1111/mec.15720

Hohenlohe, P. A., Phillips, P. C., and Cresko, W. A. (2010). Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *Int. J. Plant Sci.* 171, 1059–1071. doi: 10.1086/656306

Holderegger, R., Kamm, U., and Gugerli, F. (2006). Adaptive vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecol.* 21, 797–807. doi: 10.1007/s10980-005-5245-9

Holliday, J. A., Aitken, S. N., Cooke, J. E., Fady, B., Gonzalez-Martinez, S. C., Heuertz, M., et al. (2017). Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. *Mol. Ecol.* 26, 706–717. doi: 10.1111/mec.13963

Ingvarsson, P. K., Hvidsten, T. R., and Street, N. R. (2016). Towards integration of population and comparative genomics in forest trees. *New Phytol.* 212, 338–344. doi: 10.1111/nph.14153

Isabel, N., Holliday, J. A., and Aitken, S. N. (2020). Forest genomics: advancing climate adaptation, forest health, productivity, and conservation. *Evol. Appl.* 13, 3–10. doi: 10.1111/eva.12902

Jia, K. H., Zhao, W., Maier, P. A., Hu, X. G., Jin, Y., Zhou, S. S., et al. (2020). Landscape genomics predicts climate change-related genetic offset for the widespread *Platycladus orientalis* (Cupressaceae). *Evol. Appl.* 13, 665Ű676. doi: 10.1111/eva.12891

Jombart, T., Pontier, D., and Dufour, A. B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity* 102, 330–341. doi: 10.1038/hdy.2008.130

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., et al. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–3969. doi: 10.1111/j.1365-294X.2007.03442.x

Joost, S., Kalbermatten, M., and Bonin, A. (2008). Spatial analysis method (sam): a software tool combining molecular and environmental data to identify candidate loci for selection. *Mol. Ecol. Resour.* 8, 957–960. doi: 10.1111/j.1755-0998.2008.02162.x

Joost, S., Vuilleumier, S., Jensen, J. D., Schoville, S., Leempoel, K., Stucki, S., et al. (2013). Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Mol. Ecol.* 22, 3659–3665. doi: 10.1111/mec.12352

Jordan, R., Hoffmann, A. A., Dillon, S. K., and Prober, S. M. (2017). Evidence of genomic adaptation to climate in *Eucalyptus microcarpa*: implications for adaptive potential to projected climate change. *Mol. Ecol.* 26, 6002–6020. doi: 10.1111/mec.14341

Jump, A. S. (2006). Genetic effects of chronic habitat fragmentation in a wind-pollinated tree. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8096–8100. doi: 10.1073/pnas.0510127103

Jump, A. S., and Peñuelas, J. (2005). Running to stand still: adaptation and the response of plants to rapid climate change. *Ecol. Lett.* 8, 1010–1020. doi: 10.1111/j.1461-0248.2005.00796.x

Jump, A. S., Ruiz-Benito, P., Greenwood, S., Allen, C. D., Kitzberger, T., Fensham, F., et al. (2017). Structural overshoot of tree growth with climate variability and the global spectrum of drought-induced forest dieback. *Global Change Biol.* 23, 3742–3757. doi: 10.1111/gcb.13636

Lefèvre, F., Koskela, J., Hubert, J., Kraigher, H., Longauer, R., Olrik, D. C., et al. (2013). Dynamic conservation of forest genetic resources in 33 European countries. *Conserv. Biol.* 27, 373–384. doi: 10.1111/j.1523-1739.2012.01961.x

Legendre, P., and Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1–24. doi: 10.2307/2657192

Legendre, P., Fortin, M.-J., Borcard, D., and Peres-Neto, P. (2015). Should the Mantel test be used in spatial analysis? *Methods Ecol. Evol.* 6, 1239–1247. doi: 10.1111/2041-210x.12425

Legendre, P., and Legendre, L. (2012). *Numerical Ecology*. Amsterdam: Elsevier.

Lin, Y. C., Wang, J., Delhomme, N., Schiffthaler, B., Sundström, G., Zuccolo, A., et al. (2018). Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10970–E10978. doi: 10.1073/pnas.1801437115

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., et al. (2017). Breaking RAD: an evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17, 142–152. doi: 10.1111/1755-0998.12635

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., et al. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Mol. Ecol.* 25, 170–184. doi: 10.1111/mec.13468

Manion, G., Lisk, M., Ferrier, S., Nieto-Lugilde, D., and Fitzpatrick, M. (2014). *GDM: Functions for Generalized Dissimilarity Modeling. R Package Version 1.2.3*. Available at online: http://CRAN.R-project.org/package=gdm (accessed 18 January 2017).

Manthey, J. D., and Moyle, R. G. (2015). Isolation by environment in white-breasted nuthatches (*Sitta carolinensis*) of the Madrean Archipelago sky islands: a landscape genomics approach. *Mol. Ecol.* 24, 3628–3638. doi: 10.1111/mec.13258

Martins, K., Gugger, P. F., Llanderal-Mendoza, J., González-Rodríguez, A., Fitz-Gibbon, S. T., Zhao, J.-L., et al. (2018). Landscape genomics provides evidence of climate-associated genetic variation in Mexican populations of *Quercus rugosa*. *Evol. Appl.* 11, 1842–1858. doi: 10.1111/eva.12684

McKinney, G. J., Larson, W. A., Seeb, L. W., and Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on breaking RAD by Lowry et al. (2016). *Mol. Ecol. Resour.* 17, 356–361. doi: 10.1111/1755-0998.12649

Milesi, P., Berlin, M., Chen, J., Orsucci, M., Li, L. L., Jansson, G., et al. (2019). Assessing the potential for assisted gene flow using past introduction of Norway spruce in southern Sweden: local adaptation and genetic basis of quantitative traits in trees. *Evol. Appl.* 12, 1946–1959. doi: 10.1111/eva.12855

Nagamitsu, T., and Shuri, K. (2021). Seed transfer across geographic regions in different climates leads to reduced tree growth and genetic admixture in *Quercus mongolica* var. *crispula*. *Forest Ecol. Manag.* 482:118787. doi: 10.1016/j.foreco.2020.118787

Neale, D. B., and Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* 12, 111–122. doi: 10.1038/nrg2931

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., et al. (2013). *Package 'Vegan'. Community Ecology Package, Version 2*. Available online at: http://CRAN.R-project.org/package=vegan (accessed September 23, 2021).

Olazcuaga, L., Loiseau, A., Parrinello, H., Paris, M., Fraimout, A., Guedot, C., et al. (2020). A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure. *Mol. Biol. Evol.* 37, 2369–2385. doi: 10.1093/molbev/msaa098

Petit, R. J., and Hampe, A. (2006). Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Evol. Syst.* 37, 187–214. doi: 10.1146/annurev.ecolsys.37.091305.110215

Pina-Martins, F., Baptista, J., Pappas, G., and Paulo, O. S. (2019). New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Global Change Biol.* 25, 337–350. doi: 10.1111/gcb.14497

Rellstab, C. (2021). Genomics helps to predict maladaptation to climate change. *Nat. Clim. Change* 11, 85–86. doi: 10.1038/s41558-020-00964-w

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., et al. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present and future climatic conditions. *Mol. Ecol.* 25, 5907–5924. doi: 10.1111/mec.13889

Rinnan, D. S., and Lawler, J. (2019). Climate-niche factor analysis: a spatial approach to quantifying species vulnerability to climate change. *Ecography* 42, 1494–1503. doi: 10.1111/ecog.03937

Savolainen, O., Lascoux, M., and Merila, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.* 14, 807–820. doi: 10.1038/nrg3522

Savolainen, O., Pyhäjärvi, T., and Knürr, T. (2007). Gene flow and local adaptation in trees. *Annu. Rev. Ecol. Evol. Syst.* 38, 595–619. doi: 10.2307/30033873

Scalfi, M., Mosca, E., Di Pierro, E. A., Troggio, M., Vendramin, G. G., Sperisen, C., et al. (2014). Micro and macro-geographic scale effect on the molecular imprint of selection and adaptation in Norway spruce. *PLoS One* 9:e115499. doi: 10.1371/journal.pone.0115499

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Annu. Rev. Ecol. Evol. Syst.* 43, 23–43. doi: 10.1146/annurev-ecolsys-110411-160248

Sork, V. L. (2017). Genomic studies of local adaptation in natural plant populations. *J. Hered.* 109, 3–15. doi: 10.1093/jhered/esx091

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., and Neale, D. B. (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. doi: 10.1007/s11295-013-0596-x

Stucki, S., Orozco-Terwengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., et al. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Mol. Ecol. Resour.* 17, 1072–1089. doi: 10.1111/1755-0998.12629

Tallmon, D. A., Luikart, G., and Waples, R. S. (2004). The alluring simplicity and complex reality of genetic rescue. *Trends Ecol. Evol.* 19, 489–496. doi: 10.1016/j.tree.2004.07.003

Ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167–1179. doi: 10.2307/1938672

Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42, 207–219. doi: 10.1007/BF02294050

Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., et al. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evol. Lett.* 4, 4–18. doi: 10.1002/evl3.154

Walters, S. J., Robinson, T. P., Byrne, M., Wardell-Johnson, G. W., and Nevill, P. (2020). Contrasting patterns of local adaptation along climatic gradients between a sympatric parasitic and autotrophic tree species. *Mol. Ecol.* 29, 3022–3037. doi: 10.1111/mec.15537

Wang, I. J., and Bradburd, G. S. (2014). Isolation by environment. *Mol. Ecol.* 23, 5649–5662. doi: 10.1111/mec.12938

Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C., and Tallmon, D. A. (2015). Genetic rescue to the rescue. *Trends Ecol. Evol.* 30, 42–49. doi: 10.1016/j.tree.2014.10.009

Wright, S. I., and Gaut, B. S. (2004). Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* 22, 506–519. doi: 10.1093/molbev/msi035

XLSTAT (2012). *Leading Data Analysis and Statistical Solution for Microsoft Excel*. New York, NY: Addinsoft SRL.

Zhao, W., Sun, Y.-Q., Pan, J., Sullivan, A. R., Arnold, M. L., Mao, J.-F., et al. (2020). Effects of landscapes and range expansion on population structure and local adaptation. *New Phytol.* 228, 330–343. doi: 10.1111/nph.16619

Zhu, S. S., Chen, J., Zhao, J., Comes, H. P., Li, P., Fu, C. X., 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. doi: 10.1111/nph.16798

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

frontiers | Frontiers in Plant Science

CORRECTION
published: 28 April 2022
doi: 10.3389/fpls.2022.911163

Corrigendum: Landscape Genomics in Tree Conservation Under a Changing Environment

Li Feng[1†] *and Fang K. Du*[2*†]

[1] School of Pharmacy, Xi'an Jiaotong University, Xi'an, China, [2] School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China

Keywords: changing environment, genotype-environment associations (GEAs), landscape genomics, local adaptation, tree conservation

A Corrigendum on

Landscape Genomics in Tree Conservation Under a Changing Environment
by Feng, L., and Du, F. K. (2022). Front. Plant Sci. 13:822217. doi: 10.3389/fpls.2022.822217

In the original article, there was a mistake in the legend for **Figure 1** titled as **The general framework of landscape genomics for tree conservation** as published. We missed the proper citations about plots of F_{ST} outlier test, cluster and RONA which were depicted in **Figure 1**. The correct legend appears below.

In the original article, there was a mistake in **Figure 1** as published. "In the early **Figure 1** of this article, it contained some elements that debate might arise." The corrected **Figure 1** appears below.

In the original article Feng, L., Ruhsam, M., Wang, Y. H., Li, Z. H., and Wang, X. M. (2020). Using demographic model selection to untangle allopatric divergence and diversification mechanisms in the *Rheum palmatum* complex in the Eastern Asiatic Region. *Mol. Ecol.* 29, 1791-1805. doi: 10.1111/mec.15448 was not cited in the article. The citation has now been inserted in the legend of **Figure 1** and should read:

The general framework of landscape genomics for tree conservation. The plots of cluster, F_{ST} outlier test and RONA are modified from Du et al. (2020) and Feng et al. (2020), respectively.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

REFERENCES

Du, F. K., Wang, T., Wang, Y., Ueno, S., and De Lafontaine, G. (2020). Contrasted patterns of local adaptation to climate change across the range of an evergreen oak, *Quercus aquifolioides*. *Evol. Appl.* 13, 2377–2391. doi: 10.1111/eva.13030

Feng, L., Ruhsam, M., Wang, Y. H., Li, Z. H., and Wang, X. M. (2020). Using demographic model selection to untangle allopatric divergence and diversification mechanisms in the *Rheum palmatum* complex in the Eastern Asiatic Region. *Mol. Ecol.* 29, 1791–1805. doi: 10.1111/mec.15448

OPEN ACCESS

Edited and reviewed by:
Jue Ruan,
Chinese Academy of Agricultural Sciences (CAAS), China

***Correspondence:**
Fang K. Du
dufang325@bjfu.edu.cn

†ORCID:
Li Feng
orcid.org/0000-0002-8252-9463
Fang K. Du
orcid.org/0000-0002-7377-5259

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *02 April 2022*
Accepted: *12 April 2022*
Published: *28 April 2022*

Citation:
Feng L and Du FK (2022) Corrigendum: Landscape Genomics in Tree Conservation Under a Changing Environment. Front. Plant Sci. 13:911163. doi: 10.3389/fpls.2022.911163

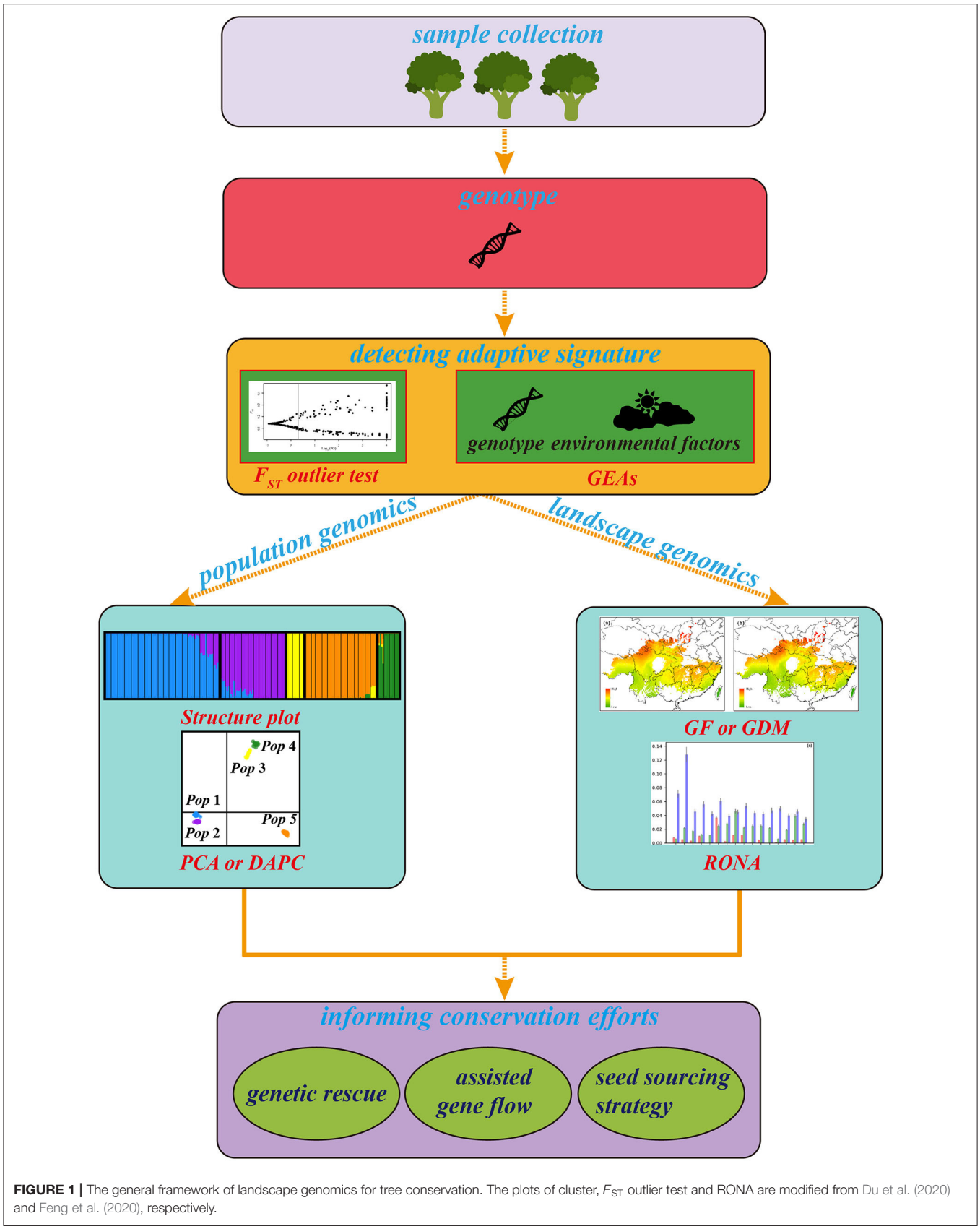


FIGURE 1 | The general framework of landscape genomics for tree conservation. The plots of cluster, F_{ST} outlier test and RONA are modified from Du et al. (2020) and Feng et al. (2020), respectively.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 March 2022
doi: 10.3389/fpls.2022.822942

A High-Quality Reference Genome Sequence and Genetic Transformation System of *Aralia elata*

Wenxuan Liu[1†], Wenhua Guo[2†], Song Chen[1], Honghao Xu[2], Yue Zhao[2], Su Chen[1] and Xiangling You[2]*

[1]State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, China, [2]Key Laboratory of Saline-Alkali Vegetation Ecology Restoration, Ministry of Education, Northeast Forestry University, Harbin, China

OPEN ACCESS

Edited by:
Fang Du,
Beijing Forestry University, China

Reviewed by:
Liangsheng Zhang,
Zhejiang University, China
Sunil Kumar Sahu,
Beijing Genomics Institute (BGI), China
Guangpeng Ren,
Lanzhou University, China

***Correspondence:**
Xiangling You
youxiangling@nefu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *26 November 2021*
Accepted: *14 February 2022*
Published: *01 March 2022*

Citation:
Liu W, Guo W, Chen S, Xu H, Zhao Y, Chen S and You X (2022) A High-Quality Reference Genome Sequence and Genetic Transformation System of Aralia elata. Front. Plant Sci. 13:822942. doi: 10.3389/fpls.2022.822942

Aralia elata is a perennial woody plant of the genus *Aralia* in the family Araliaceae. It is rich in saponins and therefore has a wide range of pharmacological effects. Here, we report a high-quality reference genome of *A. elata*, with a genome size of 1.21 Gb and a contig N50 of 51.34 Mb, produced by PacBio HiFi sequencing technology. This is the first genome assembly for the genus *Aralia*. Through genome evolutionary analysis, we explored the phylogeny and whole genome duplication (WGD) events in the *A. elata* genome. The results indicated that a recent WGD event occurred in the *A. elata* genome. Estimation of the divergence times indicated that the WGD may be shared by Araliaceae. By analyzing the genome sequence of *A. elata* and combining the transcriptome data from three tissues, we discovered important genes related to triterpene saponins biosynthesis. Furthermore, based on the embryonic callus induction system of *A. elata* established in our laboratory, we set up the genetic transformation system of this plant. The genomic resources and genetic transformation system obtained in this study provide insights into *A. elata* and lays the foundation for further exploration of the *A. elata* regulatory mechanism.

Keywords: *Aralia elata*, genome assembly, evolutionary analysis, terpenoid biosynthesis pathway, transgenic system

INTRODUCTION

Aralia elata (Miq.) Seem. (Araliaceae), also known as Chinese angelica-tree, is widely distributed in Northeast China (mainly Heilongjiang and Jilin province), Korea, Japan, Russia, the south of Far East, the south of Sakhalin, and Kuril Islands (Ahn, 1998; Reunov et al., 2007). It is one of the most desirable mountain wild vegetables in Asia benefitting from the rich nutrients in its young shoots (Cheng et al., 2021). Furthermore, as a Chinese traditional medicinal plant, *A. elata* plays roles on rheumatism, diabetes, hepatitis, neurasthenia, and stomach spasms (Zhang et al., 2018), especially the anti-tumor role (Duan et al., 2019). Those medicinal potentials were depended on the bioactive components of saponins. More than 100 kinds of saponins belonging to triterpene saponins have been reported in *A. elata* and they are mainly oleanane-type saponins (Cheng et al., 2021).

The triterpenoid biosynthesis is initiated from isopentenyl diphosphate (IPP) that is derived from the metabolism of cytosolic mevalonic acid (MVA) or the plastid methylerythritol phosphate (MEP; Sawai and Saito, 2011). This biosynthesis process is catalyzed by a series of key enzymes. The enzymes before triterpenoid structural skeleton formation include farnesyl diphosphate (FPP) synthase (FPS), squalene synthase (SS), and squalene epoxidase (SE). Oxidosqualene cyclases (OSCs) catalyze oxidosqualene to different triterpenoid backbones. *β-AeAS* has been identified to encode the OSC in *A. elata* (Wu, 2011). Subsequently, the key enzymes are cytochrome P450 monooxygenases (P450), which mediate oxidations. Uridine diphosphate-dependent glycosyl transferases (UDT) finally catalyze glycosylations to generate different triterpenoid saponins (Sawai and Saito, 2011). Recent studies revealed that subfamilies of CYP450, such as CYP71, CYP72, CYP88, CYP93, CYP716, and CYP749 are extensively involved in the oxidative stress response (Heitz et al., 2012) and the biosynthesis of triterpenes (Carelli et al., 2011; Fukushima et al., 2011; Han et al., 2011), sterols, indole alkaloids (Irmler et al., 2000; Collu et al., 2001; Nafisi et al., 2007), geraniol iridoid (Höfer et al., 2013), etc.

In *A. elata*, genes that potentially encode these key enzymes, including *AeFPS* (Wu, 2011), *AeSS* (Cheng, 2011), *AeSE* (Zhao, 2012), and *β-AeAS* (Wu, 2011) have been cloned and investigated by real-time qRT-PCR. In addition, 254 members of P450 and 122 UGT families were identified by the RNA-sequencing analysis (Cheng et al., 2020). But for the complex pathway of triterpenoid synthesis, the information of these key enzyme encoding genes is still limited due to the lack of a genome reference of this species.

With the rapid development of sequencing technology and reduction of sequencing cost, more and more plant genomes have been sequenced and published. The third-generation sequencing, especially the High Fidelity (HiFi) technology, has greatly reduced the cost and shortened the circle of genome sequencing. In this study, we sequenced, assembled, and annotated a high-quality genome of *A. elata* using HiFi data. This is the first genome of the genus *Aralia*. Using comparative genomics, we explored the evolutionary trajectory and whole genome duplication (WGD) events of *A. elata*. We also identified the key enzyme encoding genes involved in the triterpenoid biosynthesis pathway in the genome. The expressional patterns of these genes were preliminarily investigated. Genetic transformation is the most efficient way to further explore the functions of the annotated genes in *A. elata*. However, no transformation system has been established for this non-model plant species. We therefore established an *Agrobacterium tumefaciens* mediated genetic transformation system for *A. elata*, which laid a solid foundation for plant genetic engineering. The genomic resources of *A. elata* provided here will be valuable for biological and breeding research on *Aralia* species and will provide new tools for Araliaceae geneticists and breeders.

MATERIALS AND METHODS

Plant Materials, DNA Extraction, and Library Construction

Fresh and healthy leaves of 2-month-old tissue culture plantlets of *A. elata* were harvested and immediately frozen in liquid nitrogen and preserved at −80°C. The samples were then sent to the company (Annoroad Gene Technology, China) for DNA extraction. The quality and quantity of the isolated DNA were assessed using a NanoDrop 2000&8000 spectrophotometer and a Qubit 2.0 Fluorometer, respectively. Illumina and PacBio libraries were constructed using the eligible DNA following the instruction for each technology, respectively.

Genome Sequencing, Assembly, and Quality Assessment

We integrated Illumina HiSeq and PacBio HiFi sequencing data to achieve the complete genome sequence of *A. elata*. The Illumina library was sequenced on the Illumina HiSeq X Ten platform following standard Illumina protocols. After filtering out adapter sequences, low-quality reads, and duplicated reads, the clean reads were used to investigate the genomic features including genome size and heterozygosity by *k*-mer distribution analysis using GenomeScope (Vurture et al., 2017). Two libraries were constructed for PacBio HiFi sequencing. The subreads generated from the PacBio libraries were assembled into contigs using hifiasm with the default parameters (Cheng et al., 2008). The Illumina sequencing reads were aligned to the genome assembly using BWA (Li, 2013) to assess its completeness. Benchmarking Universal Single-Copy Orthologs (BUSCO) was also used to assess the quality of the final genome assembly (Simão et al., 2015).

Genome Annotation

The *A. elata* genome was annotated by the integration of multiple strategies including *de novo*, homology-based, and transcriptome-based predictions. Repeat Masker and Repeat Modeler were used to identify the repetitive sequences in the genome based on repeat sequence database. Augustus was used for *de novo* prediction of protein coding genes based on the repeat masked genome. For similarity-based gene prediction, eight species including *Arabidopsis thaliana*, *Oryza sativa*, *Daucus carota*, *Populus trichocarpa*, *Apium graveolens*, *Vitis vinifera*, *Panax notoginseng*, and *Coriandrum sativum* were selected, and the protein sequences of these species were downloaded from Phytozome.[1] Annotation of coding genes in the genome was subsequently performed using these homologous proteins. BLAST with identity ≥ 0.95 and coverage ≥ 0.90 as thresholds was used to identify genes with significant similarity in the *A. elata* genome. To carry out the RNA-Seq aided gene prediction, we downloaded the transcriptome data of *A. elata* from NCBI SRA database (BioProject: PRJNA555256). The clean reads were assembled into transcripts using Trinity (Haas et al., 2013), which were aligned against the genome assembly for gene

[1]https://phytozome-next.jgi.doe.gov/

structure prediction using Program to Assemble Spliced Alignments (PASA; Haas et al., 2008). The gene sets predicted by the various strategies were integrated into a non-redundant and more complete gene set by Evidence Modeler (EVM; Haas et al., 2008). BUSCO was used to evaluate the integrity and completeness of the predicted gene set.

Analysis of Genomic Evolution and WGD Events

We used OrthoFinder to identify the orthologous groups in 12 species: five species from Apiales including *A. elata*, *Panax notoginseng*, *Daucus carota*, *Apium graveolens*, and *Coriandrum sativum*, two species from Asterales including *Lactuca sativa* and *Taraxacum mongolicum*, one species from Tubiflorae (*Capsicum annuum*), three other dicot species including *Arabidopsis thaliana*, *Carica papaya* and *Populus trichocarpa*, and one monocot *Oryza sativa*, which was used as the outgroup (Guo et al., 2021). MUSCLE was used for multiple sequence alignment for each single-copy orthologous group identified by OrthoFinder (Emms and Kelly, 2019). All the alignment blocks were then manually concatenated, and substitution model for each alignment block was estimated using ModelTest-NG (Darriba et al., 2020) program. The results were subsequently used to construct a phylogenetic tree using maximum-likelihood algorithm. Divergence times of these species in the phylogenetic tree were estimated with MCMCtree (v4.0) using the Bayesian Relaxed Molecular Clock (BRMC) approach (Yang, 2007). The parameters of MCMCtree were set as follows: burn-in = 2,000; sample-frequency = 10; and sample-number = 20,000. *Oryza sativa* was designated as an outgroup of the phylogenetic tree. The calibration times of each divergent nodes were obtained from the TimeTree website (Kumar et al., 2017). Gene family amplification and contraction was analyzed by CAFÉ using the phylogenetic tree and gene numbers in each orthogroup (De Bie et al., 2006).

Identification and Tissue Specific Expression of Genes Involved in Triterpene Saponins Biosynthesis

BLASTP, with E-value of 1e−5 as a threshold, was used to identify candidate enzymes that catalyze triterpene saponins biosynthesis. The NCBI Conserved Domain Database (Marchler-Bauer et al., 2010) was used to scan conserved domains in these candidates. Only the protein sequences containing canonical domains were identified as authentic enzymes. IQ-TREE (Nguyen et al., 2015) was used to construct phylogenetic trees for these protein sequences. The expressional profiles of genes encoding these enzymes were investigated using RNA-seq data retrieved from public databases (PRJNA555256). The gene expression analysis was performed using the nf-core/rnaseq v3.2 (Ewels et al., 2020) pipeline in nextflow v21.04.1 (Di Tommaso et al., 2017). The sequencing reads were mapped to the reference genome using Spliced Transcripts Alignments to a Reference V2.7.6a (STAR) as an aligner. Gene expression levels were then determined by using RNA-Seq by Expectation-Maximization v1.3.1 (RSEM). Trimmed mean of M value (TMM; Robinson and Oshlack, 2010) method was used to normalize and measure the expression levels of these samples.

Establishment of *Agrobacterium* Mediated Genetic Transformation for *Aralia elata*

The system of vegetable propagation of *A. elata* was built in our lab (Dai et al., 2011). The somatic embryogenic callus was induced from the roots of the somatic embryo plants in the induction medium (1/2 SH medium with 3.0 mg/L of IBA and 0.2 mg/L of KT) for 3 weeks. When the callus was transformed to the re-differentiation medium: 1/2 SH with 1.0 mg/L IBA and 0.2 mg/L KT, after 6 weeks, lots of somatic embryo or plants emerged.

The roots of the above somatic embryo plants were used for *Agrobacterium tumefaciens* mediated genetic transformation. After 3 days of pre-culture, the roots were infected by *A. tumefaciens* for 5, 10, and 15 min, respectively and then co-cultured in medium for 3 days. Next, they were transformed to the selection medium with 50 mg/L kanamycin and 200 mg/L timentin. Then after 8 weeks, the calli were checked by PCR using the primers 5′-CGC ACA ATC CCA CTA TCC TT-3′, and 5′-AAG ACC GGC AAC AGG ATT C-3′ to choose the callus line of gene transformation. The positive callus lines were transformed into above plant-medium.

RESULTS

Genome Sequencing, Assembly, and Annotation

To investigate the genomic features of *A. elata*, 17, 21, 25, and 27 *K*-mer distribution analysis was performed (**Figure 1A**; **Supplementary Figure 1**), respectively, using 56.89 Gb of the Illumina reads. The Illumina reads representing 50.79× coverage based on the estimated genome size of 1.12 Gb (*K*-mer analysis; **Supplementary Table 1**). The *K*-mer distributions followed a Poisson distribution, with two peaks corresponding to homozygous and heterozygous sequences, respectively (**Supplementary Figure 1**). According to the *K*-mer distribution analysis, the genome size of *A. elata* was estimated as 1.08–1.14 Gb and the heterozygosity ratio of the genome was estimated as 1.60–1.69% (**Supplementary Table 2**). The results indicated that the genome of *A. elata* is highly heterozygous and repetitive. We then used HiFi technologies to sequence the *A. elata* genome. A total of 51.14 Gb of HiFi reads from two libraries were obtained for the genome assembly. A total of 25.75 and 25.39 Gb data were generated from the two libraries, respectively (**Supplementary Table 3**). The HiFi reads were assembled into contigs using hifiasm. The final assembled genome was 1.21 Gb in size with a contig N50 length of 51.34 Mb. The genome assembly contained 1,350 contigs, the longest contig was 100.88 Mb, and the average contig length was 0.89 Mb. The GC content of the *A. elata* genome is 36.13% (**Table 1**).

To evaluate the completeness of the genome assembly, short reads generated for the genomic survey were mapped to the genome. In total, 99.95% of the short reads were mapped to

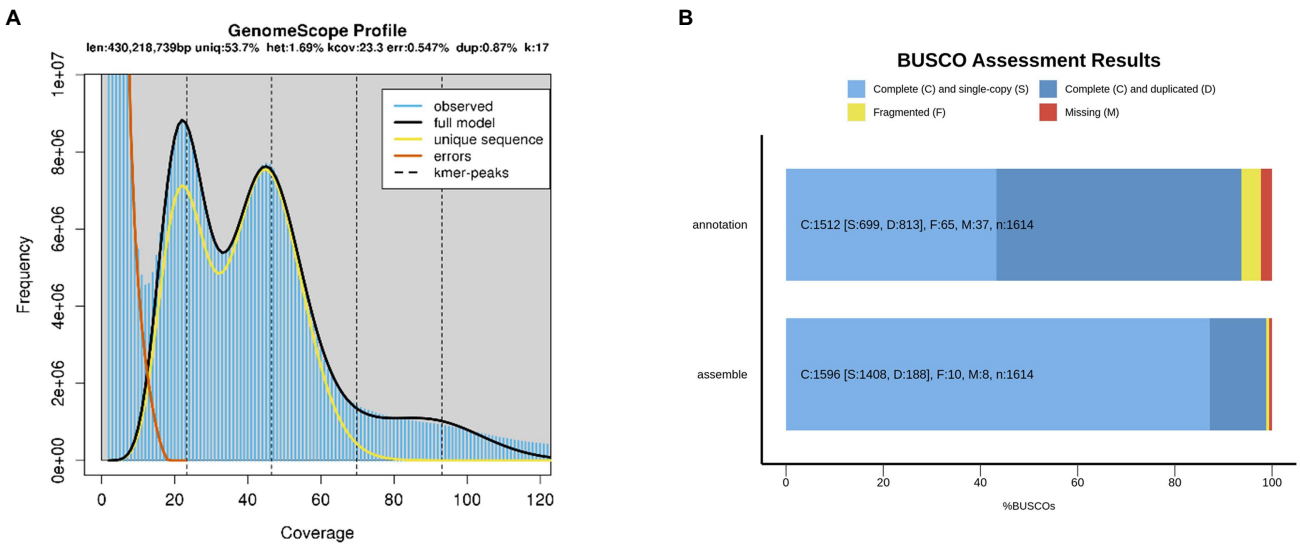


FIGURE 1 | (A) 17-mer analysis for estimating the genome size of *Aralia elata*. **(B)** Assessment the gene coverage rate of genome assembly and annotation using Benchmarking Universal Single-Copy Orthologs (BUSCO).

TABLE 1 | Assembly statistics of the *Aralia elata* genome.

N50 contig size (bp)	51,342,398
L50 contig number	8
N75 contig size (bp)	38,678,071
L75 contig number	15
N90 contig size (bp)	17,066,756
L90 contig number	22
Longest contig (bp)	100,882,612
Shortest contig (bp)	10,694
Average contig (bp)	893,717
Total length (bp)	1,206,518,707
Total N length (bp)	0
Number of contigs	1,350
GC content (%)	36.13

the contigs, 99.48% of which were properly pair-end mapped (**Supplementary Table 4**). The completeness of the genome assembly was also evaluated by BUSCO. The result revealed that the genome covered at least 98.8% of the BUSCO genes, 87.2% of which were classified as "complete and single-copy," 11.6% as "complete and duplicated," 0.6% as "fragmented," and 0.6% as "missing" (**Figure 1B**; **Supplementary Table 5**). All the results suggested a high quality of the *A. elata* genome assembly.

Repetitive sequences, including tandem repeat and interspersed repeats, are important parts of genomes. In this study, two strategies, *de novo* prediction and homology-based identification, were used to annotate the repetitive sequences in the *A. elata* genome. According to the integrated results obtained above, the proportion of repetitive sequences in the genome was 71.69%, which was higher than carrot (45.95%; Iorizzo et al., 2016). The most abundant type of repetitive elements was long terminal repeat (LTR), which accounted for 49.15% of the genome, while DNA transposon repetitive sequences accounted for only 3.86% of the genome (**Supplementary Table 6**).

To annotate the protein coding genes in the *A. elata* genome, we used a combination of *ab initio* prediction, homology-based search, and transcript evidence from RNA-seq data. Finally, a total of 37,016 genes were annotated in the genome. We evaluated the completeness and quality of the annotated proteome through BUSCO using Embryophyta_odb10 as database. The results indicated that 97.7% of the conserved genes were annotated in the genome, which included 93.7 and 4.0% complete and fragmented BUSCO genes, respectively. The BUSCO assessment indicated that the annotation of genome was of high accuracy (**Figure 1B**; **Supplementary Table 7**).

Genome Evolution of *Aralia elata*

In order to reveal the evolutionary position of *A. elata*, we compared the genome assembly with genomes from 11 other plants. A total of 250 single-copy gene families were identified among these species by OrthoFinder. These single-copy genes were used to construct a phylogenetic tree using a maximum likelihood method. Consistent with Angiosperm Phylogeny Group, *A. elata* was closed to *P. notoginseng*, another Araliaceae species and these two species were classed into a clade. This clade was most closely related to the species from Apiales family (**Figure 2A**). The divergent times of these species were then estimated based on the phylogenetic tree. We estimated that *A. elata* and *P. notoginseng* diverged from Apiaceae at approximately 80.1 million years ago (mya). *Aralia elata* and *P. notoginseng* subsequently diverged into two species at around 24.2 mya. These results showed that the relationship between *A. elata* and *P. notoginseng* is very close. In addition, we performed a comparative analysis of gene family evolution in the phylogenetic tree. A total of 1,925 gene families were expanded in the *A. elata* lineage, whereas 1,832 gene families had undergone contraction (**Figure 2A**).

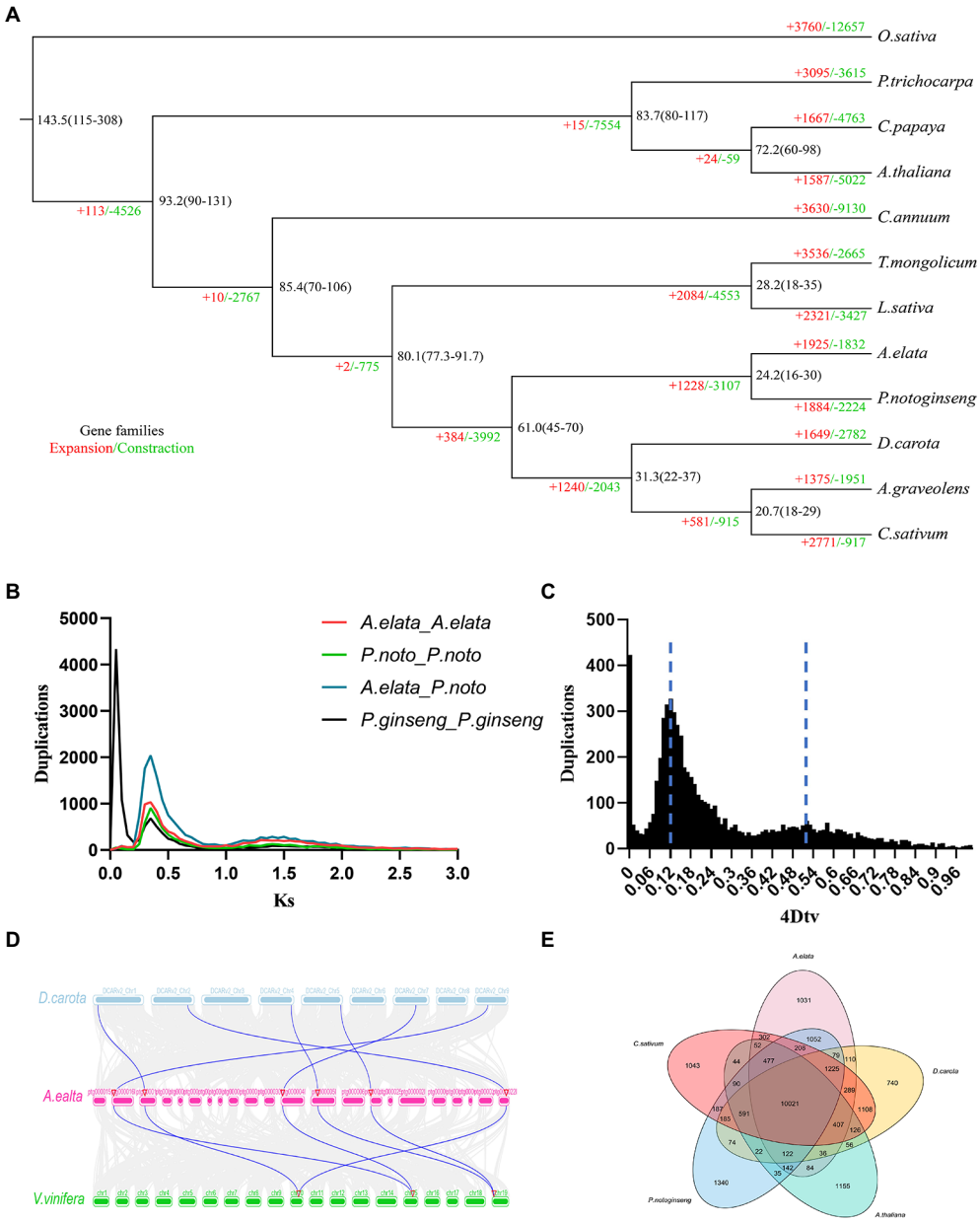


FIGURE 2 | (A) Inferred phylogenetic tree of *Aralia elata* and 11 plant species based on protein sequences of single-copy orthologous genes. Numbers at each node represent the estimated divergence time of each node in million years ago (mya). Gene family expansions are indicated in red, and gene family contractions are indicated in green. **(B)** Ks distribution of paralogous gene pairs in the *A. elata*, *Panax notoginseng*, and *Panax ginseng* genome. The probability density of Ks was estimated using the "density" function in the R language. **(C)** Distribution diagram of 4DTv values. The dark black-filled part indicates the 4DTv analysis inside *A. elata*, and the peaks marked by the dotted line indicate where the two whole genome duplication (WGD) events of *A. elata* occurred. **(D)** Collinear analysis among *Daucus carota*, *A. elata*, and *Vitis vinifera* genome. The blue lines in the genomes of *A. elata* and *V. vinifera* indicate that the 2:1 correspondence between the two collinear regions. **(E)** Venn diagram showing the cluster distribution of shared gene families among *A. elata*, *C. sativum*, *P. notoginseng*, *D. carota*, and *A. thaliana*.

The gene family analysis among these species revealed that the 33,499 genes in the *A. elata* genome were clustered into 15,637 gene families with an average size of 2.14. The members in the gene families varied greatly, and the largest family contained 277 genes. We then investigated the specific and shared gene families among the species of *A. elata*, *C. sativum*, *P. notoginseng*, *A. thaliana*, and *D. carota*. The results indicated that 10,021 gene families were observed in all the investigated species, and 1,031 gene families appeared to be lineage specific to *A. elata* (**Figure 2E**).

Whole genome duplication occurs widely in flowering plants and plays important roles in genome evolution, the formation

of new species, and gene neofunctionalization (Piegu et al., 2006; Van de Peer et al., 2009). The previous results indicated that two species in Araliaceae, *P. notoginseng* and *Panax ginseng*, have experienced one and two recent WGD events, respectively (Kim et al., 2018; Jiang et al., 2021). To further explore the evolutional trajectory of *A. elata*, we investigated the WGD events in its genome. The protein sequences from the *A. elata* genome were searched against themselves using BLASTP (E < 1e−5) to identify homologous gene pairs (Camacho et al., 2009). We calculated the 4DTv (4-fold degenerate synonymous sites of the third codons) for the optimal gene pairs and plotted the distribution of the 4DTv values (**Figure 2C**). Two peaks were observed at approximately 0.12 and 0.50, respectively. The right peak at approximately 0.50 revealed the core eudicot gamma triplication event. The left peak at approximately 0.12 indicated that *A. elata* underwent a recent WGD event. We then investigated the syntenic blocks between *V. vinifera* and *A. elata* using McscanX to further confirm the WGD event in *A. elata*, because *V. vinifera* does not undergo any recent WGDs. A 2:1 syntenic relationship between *A. elata* and *V. vinifera* (**Figure 2D**) was observed, which confirmed the recent WGD event occurred in the *A. elata* genome.

Ks (synonymous substitution rate) values can be used to estimate the timing of large-scale duplications (Blanc and Wolfe, 2004). We calculated the *Ks* values of the gene pairs and plotted the distributions to estimate the occurrence time of the WGD events of *A. elata*, *P. notoginseng*, and *P.ginseng*, respectively (**Figure 2B**; Chen et al., 2020). Two peaks were observed in the *Ks* distributions of the *P. notoginseng* and *A. elata* genomes, whereas the *P.ginseng* genome contained three *Ks* peaks. The *Ks* distribution result of *A. elata* was consistent with the 4DTv values. The main peak at approximately 0.38 indicated that a recent WGD event occurred in the *A. elata* genome. Similar *Ks* peaks around 0.38 were also found in the *P. notoginseng* and *P.ginseng* genomes, which indicated that the recent WGD event may be shared by *A. elata*, *P. notoginseng*, and *P.ginseng*. Then we calculated the occurring time of the WGD event of *A. elata* according to the method reported (Qin et al., 2014). The WGD event was estimated to occur at approximately 29.6 mya in the *A. elata* genome. Because the divergence time of *A. elata* and *P. notoginseng* was estimated to be 24.2 mya, this WGD event may occur before the differentiation of the two species. This is consistent with the published result (Jiang et al., 2021). All the results above indicated that unlike *P. ginseng*, who experienced an extra genus specific WGD, *A. elata* and *P. notoginseng* genome experienced only one recent WGD (Kim et al., 2018). In addition, this WGD may be shared by species in Araliaceae.

Analysis of Key Enzyme Encoding Genes Involved in Triterpene Saponins Biosynthesis

The biosynthesis pathways of terpenoids in plants have been comprehensively explained. The research on *Aralia* Linn. plants have attracted extensive interest from researchers. By integrating sequence similarity, conserved domain, and phylogenetic relationship results (**Figure 3**; **Supplementary Figure 2**; **Supplementary Table 8**), we identified 22 candidate genes encoding the enzymes that may catalyze the biosynthesis processes of terpenoids in the *A. elata* genome (**Figure 4**). We used transcriptome sequencing data of *A. elata* downloaded from public database to investigate the expressional profiles of these genes. The RNA-seq reads were aligned to the genome assembly and obtained their expression levels in roots, stems, and leaves. **Figure 4** illustrates the normalized expressional levels of these enzyme-coding genes in each tissue. The results indicated that many genes appeared to be expressed in tissue-specific manners. For example, genes encoding CYP450s are abundant in stems and roots.

Previous studies have shown that CYP72A and CYP716A subfamily members are the main CYP450s involved in the biosynthesis of pentacyclic triterpene saponins. Therefore, we pay special attention to the four CYP716a and three CYP72a coding genes identified in the *A. elata* genome. At the same time, 12 genes related to the terpene skeleton and triterpene biosynthesis pathway were identified, including MVA pathway and 2,3-oxysqualene biosynthesis pathway. Among them, six enzymes (AACT, HMGS, HMGR, MK, PMK, and MVD) were associated with the MVA pathway.

In the MVA pathway, Acetyl-CoA is synthesized into Acetoacetyl-CoA, which is catalyzed by the AACT enzyme (encoded by *Arel.002085*). The expression level of *Arel.002085* in leaves and roots are slightly higher than that in stems. Acetoacetyl-CoA is synthesized into 3-Hydroxy-3-methylglutaryl-CoA, which is catalyzed by HMGS enzyme (encoded by *Arel.020097*). HMGR enzyme (encoded by *Arel.004178*) subsequently catalyzes the synthesis of Mevalonic acid, and then MK enzyme (encoded by *Arel.022041*) is used to catalyze the synthesis of Mevalonic acid-5P. The expressional levels of *Arel.020097*, *Arel.004178*, and *Arel.022041* in roots were higher than those in stems and leaves, indicating that the biosynthetic reaction mainly occurred in the roots of *A. elata*. Then, under the catalysis of PMK enzyme (encoded by *Arel.027136*), Mevalonic acid-5-pyrophosphate was generated, and then Isopentenyl pyrophosphate is catalyzed by MVD enzyme (encoded by *Arel.003662*), and then catalyzed by IPPI enzyme (encoded by *Arel.030181*) to form Dimethylallyl pyrophosphate, further condensation of Isopentenyl pyrophosphate and Dimethylallyl pyrophosphate to form various terpenoids. Except for *Arel.003662*, whose expression levels in leaves and roots are slightly higher than those in stems, the expression levels of *Arel.027136* and *Arel.030181* in stems are slightly higher than those in leaves and roots. The results indicated that these biosynthetic reactions in this part may occur in the stems.

After that, Isopentenyl pyrophosphate and Dimethylallyl pyrophosphate is catalyzed by GPS enzyme (encoded by *Arel.001014*) to produce Geranyl diphosphate. FPS enzyme (encoded by *Arel.030122*) then catalyzes Geranyl diphosphate into Farnesyl pyrophosphate. SS enzyme (encoded by *Arel.022914*) catalyzes Farnesyl pyrophosphate into Squalene. Finally, 2,3-oxidosqualene is synthesized by the catalysis of SE enzyme (encoded by *Arel.013186*). In the synthetic pathway, *Arel.001014*

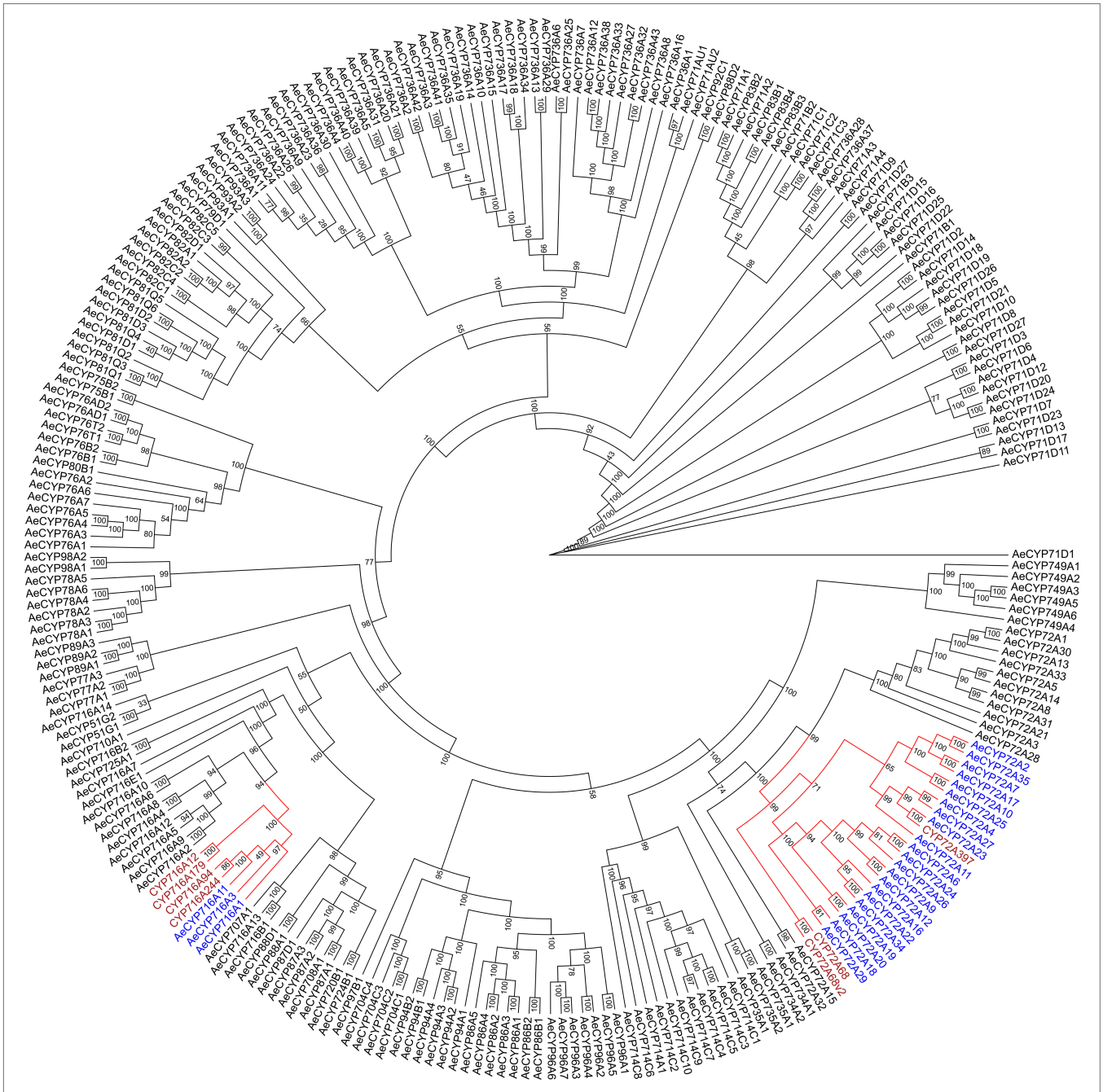


FIGURE 3 | Phylogenetic analysis of P450 genes. Genes that putatively encode Cytochromes P450 were identified by the integration of sequence similarities, conserved domains, and the phylogeny. The colored clades in the phylogenetic tree include genes encoding the subfamilies of CYP72A and CYP716A, which are the key enzymes of P450 participated in the biosynthetic pathway of triterpene saponins. Genes marked in purple are reported to catalyze the synthesis of triterpene saponins in other species.

and *Arel.030122* genes are expressed at higher levels in leaves and stems, while *Arel.002914* and *Arel.013186* genes are expressed at higher levels in stems and roots. Based on these results, it is speculated that the key triterpene skeleton biosynthesis reaction mainly occurs in the stem.

Finally, 2,3-oxidosqualene is catalyzed by β-AS (encoded by *Arel.032339*, *Arel.032335*, and *Arel.020446*) to form β-amyrin, and then oleanolic acid is formed under the catalysis of CYP716A subfamily members (including CYP716A244, CYP716A12, CYP716A179, and CYP716A94), and then Hederagenin is formed under the catalysis of CYP72A subfamily members (including CYP72A68, CYP72A68v2, and CYP72A397). Among them, genes encoding CYP72A and CYP716A subfamily members have higher expression levels in roots and stems than in leaves.

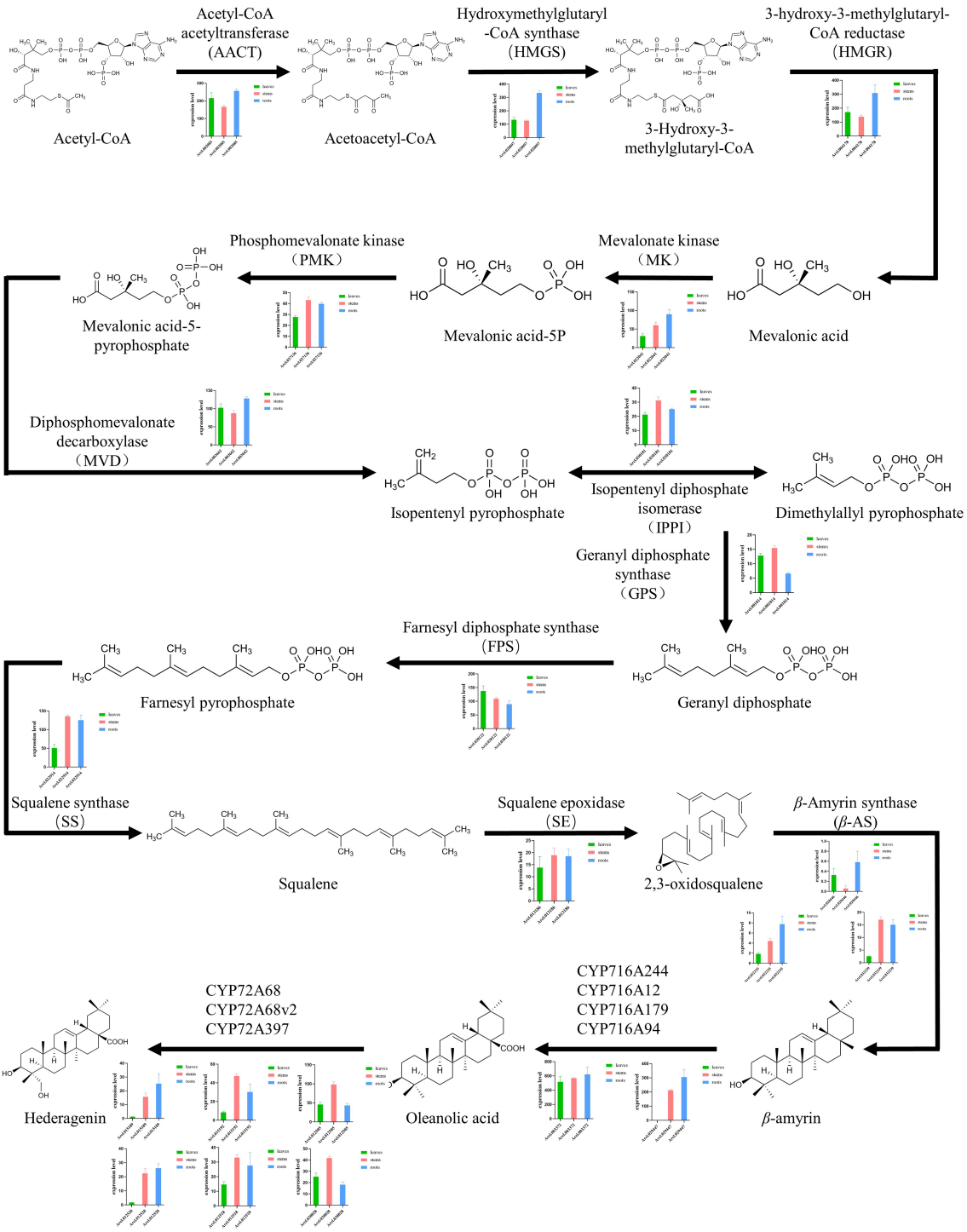


FIGURE 4 | Overview of the saponin biosynthetic pathway in *Aralia elata* and expression profiles of key enzyme encoding genes. The histogram shows the expression levels of different genes in different tissues. Green indicates the amount of expression in the leaves, red indicates the amount of expression in the stem, and blue indicates the amount of expression in the root. The most genes of enzymes involved in the saponin biosynthesis pathway were highly expressed in roots higher than in leaves and stems.

Based on the expression levels of these genes, we explored the secondary metabolism in *A. elata* plants at the spatial levels. We compared the expressional patterns of these genes in different tissues. We found that most of the genes involved in the saponin biosynthesis were specifically expressed in roots, and a few were highly expressed in leaves and stems (**Figure 5**).

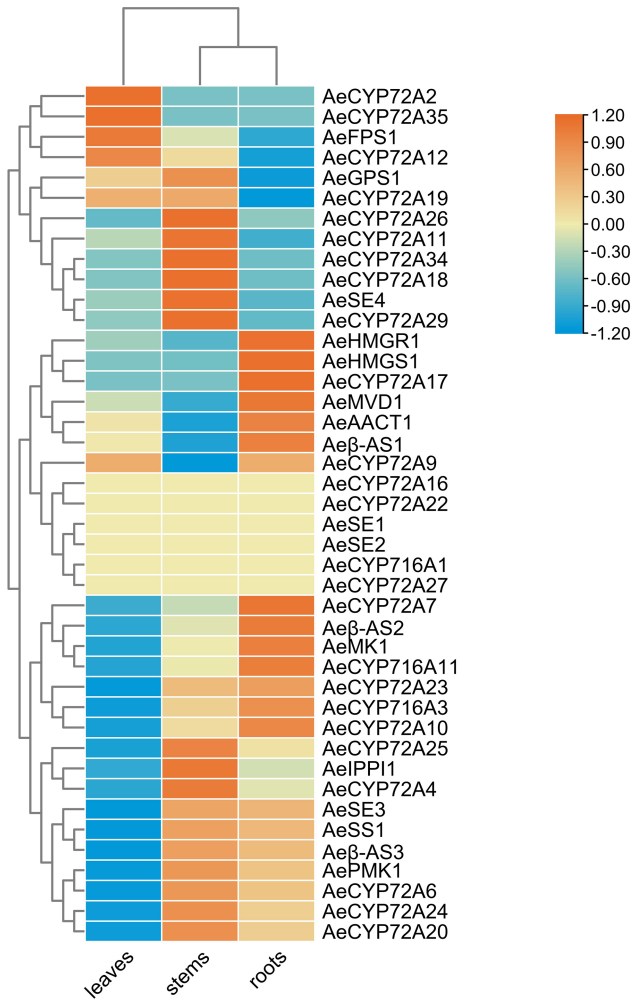


FIGURE 5 | Heatmap of gene transcript abundance in the saponin biosynthetic pathway of key enzyme genes at three different parts in *Aralia elata*. RPKM values are log2-based. Yellow and blue indicate high and low expression levels, respectively.

Establishment of the *Agrobacterium* Mediated Transformation System for *Aralia elata*

Biotechnology is an efficient way to increase the contents of secondary metabolites in plants. The annotation of *A. elata* genome will provide many candidate genes for the generation of genetically modified *A. elata* plants. However, it is still difficult for genetic transformation in *A. elata*. Based on the embryonic callus induction system of *A. elata* established in our laboratory, we set up the genetic transformation system of this plant. Roots of well-grown tissue culture seedlings were used as explants for *A. tumefacien* infestation, and the roots were precultured, co-cultured, and selection cultured (kanamycin resistance) to obtain resistant callus. DNA extracted from the resistant callus was examined by PCR. As shown in **Figure 6**, the target fragment was successfully detected in the positive transgenic plants and found to be better transformed at 10 min

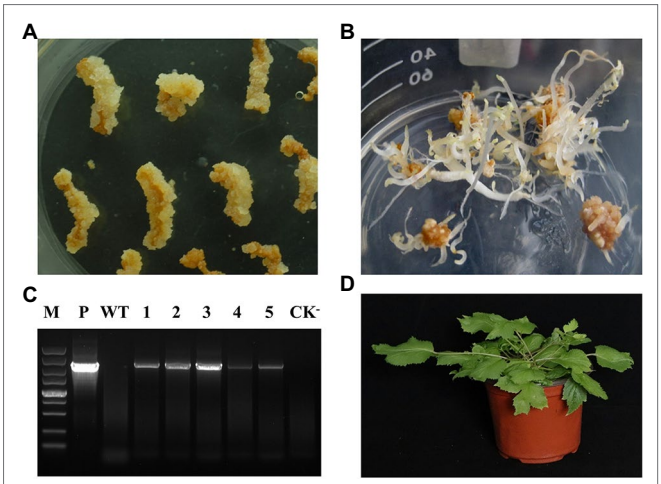


FIGURE 6 | Establishment systems of the propagation and gene information of *Aralia elata*. **(A)** The induction of the somatic embryogenic callus from the root in the 1/2SH medium with 3.0 mg/L of IBA and 0.2 mg/L of KT for 3-week culture. **(B)** The somatic embryogenesis from that callus in the 1/2SH medium with 1.0 mg/L of IBA and 0.2 mg/L of KT for 3-week culture. **(C)** Introduction of the exogenous genes in the plant. P, positive control; The WT was not transformed with *Agrobacterium*; 1–5, the transgenic plant with target gene; CK^-, negative control. **(D)** The somatic embryo plant of gene transformation.

of infection time. We transferred the transgenic callus to differentiation medium to obtain somatic embryonic seedlings. Next, the somatic embryo seedlings were transferred to WPM medium containing 20 g/L sucrose and cultured under 16 h light and 8 h dark conditions for 4 weeks, and then the plants were moved into soil and cultured in a greenhouse for 2 months, as shown in **Figure 6**, the transgenic plants grew well, and finally we obtained transgenic plants.

DISCUSSION

Aralia elata is one of the most widely used Chinese medicinal plants from the family Araliaceae and is well known in China and worldwide for its good efficacy. Triterpenoid saponins are widely existed in Araliaceae and are the most studied active ingredients of *A. elata*. Most of the aglycones are oleanolic acid, ivy, and their derivatives. Kochetkoy (1963) reported its chemical composition for the first time and obtained three saponins. The research on saponins of *A. elata* has become a hot topic, and many studies have reported its chemical components. Up to now, more than 100 saponins have been isolated and identified from *A. elata*. However, the complete biosynthetic pathway of saponins of *A. elata* has not been determined and further research is needed. Here, we briefly analyzed the terpenoid biosynthesis pathway of *A. elata*, to provide a reference for follow-up research. The contents of triterpenoid saponins in *A. elata* could be increased by genetically modification of the candidate genes involved in this pathway. The annotated genome and the genetic transformation system established in this study would be used for the further functional genome analysis in this species.

In the family Aralialeae, the genomes of some species including *Eleuthorcoccus senticusus* (Yang et al., 2021), *P. ginseng* (Kim et al., 2018), and *P. notoginseng* (Jiang et al., 2021) have been reported. The high-quality genomic analysis of *A. elata* will provide a valuable extensive information for studying the evolutionary landscape of other species in Araliaceae. Gene mining of high-quality genomic and transcriptomic data can provide resources for further exploration of plant growth and secondary metabolism mechanisms (Tu et al., 2020). So, we produced the first high-quality genome reference for *A. elata* with the latest sequencing technologies and bioinformatics methods. The size of the assembled genome is very close to the predicted result of *K*-mer, reflecting no obvious expansion or collapse occurred during the assembly process. Benefit from the long lengths and high accuracy of HiFi reads, the continuity and completeness of the *A. elata* genome obtained in this study are at a high-quality level. The evolutional process of the genome was studied based on the genome. Our results combined with the published genomes revealed the WGD trajectory in Araliaceae. A recent WGD event occurred before the divergence of species in Araliaceae.

In conclusion, the high-quality *A. elata* genome sequence described in this article, combined with comparative genome analysis, identification and tissue species expression analysis of putative genes involved in saponins biosynthesis, and the establishment of an efficient genetic transformation system of *A. elata* will contribute to *A. elata* breeding and cultivation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

XY conceived the project. XY and SuC designed the experiments. WL, SoC, and WG performed most of the experiments and analyzed the data. The other authors assisted in the experiments and discussed the results. XY, SuC, WL, and WG wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the National Natural Science Foundation of China (No. 30972390) and the Fundamental Research Funds for the Central Universities (2572018CL02).

ACKNOWLEDGMENTS

The authors extremely appreciate the suggestions and comments from the editors and reviewers for improving the quality of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.822942/full#supplementary-material

REFERENCES

Ahn, D. (1998). *Illustrated Book of Korean Medicinal Herbs*. Seoul (Korea): Kyo-hak Publishing Co., 107.

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi: 10.1105/tpc.021345

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Carelli, M., Biazzi, E., Panara, F., Tava, A., Scaramelli, L., Porceddu, A., et al. (2011). *Medicago truncatula* CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell* 23, 3070–3081. doi: 10.1105/tpc.111.087312

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Cheng, H. (2011). *Study on Genes Related to Triterpenoid Saponin Biosynthesis Pathway in Aralia elata*. ChangChun, China: Jilin University.

Cheng, H., Concepcion, G., Feng, X., Zhang, H., and Li, H. (2008). Haplotype-resolved de novo assembly with phased assembly graphs. arXiv. [Preprint].

Cheng, Y., Liu, H., Tong, X., Liu, Z., Zhang, X., Chen, Y., et al. (2021). Effects of shading on triterpene saponin accumulation and related gene expression of *Aralia elata* (Miq.) seem. *Plant Physiol. Biochem.* 160, 166–174. doi: 10.1016/j.plaphy.2021.01.009

Cheng, Y., Liu, H., Tong, X., Liu, Z., Zhang, X., Li, D., et al. (2020). Identification and analysis of CYP450 and UGT supergene family members from the transcriptome of *Aralia elata* (Miq.) seem reveal candidate genes for triterpenoid saponin biosynthesis. *BMC Plant Biol.* 20:214. doi: 10.1186/s12870-020-02411-6

Collu, G., Unver, N., Peltenburg-Looman, A. M., Van Der Heijden, R., Verpoorte, R., and Memelink, J. (2001). Geraniol 10-hydroxylase1, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. *FEBS Lett.* 508, 215–220. doi: 10.1016/S0014-5793(01)03045-9

Dai, J.-L., Tan, X., Zhan, Y.-G., Zhang, Y.-Q., Xiao, S., Gao, Y., et al. (2011). Rapid and repetitive plant regeneration of *Aralia elata* seem. via somatic embryogenesis. *Plant Cell Tissue Organ Cult.* 104, 125–130. doi: 10.1007/s11240-010-9801-x

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294. doi: 10.1093/molbev/msz189

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Duan, H., Wang, D., Wang, S., and Wu, S. (2019). Research progress of *Aralia continentalis* kitag of bioactive components. *Farm Product. Process.*

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/s13059-019-1832-y

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi: 10.1038/s41587-020-0439-x

Fukushima, E. O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., et al. (2011). CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant Cell Physiol.* 52, 2050–2061. doi: 10.1093/pcp/pcr146

Guo, X., Fang, D., Sahu, S. K., Yang, S., Guang, X., Folk, R., et al. (2021). Chloranthus genome provides insights into the early diversification of angiosperms. *Nat. Commun.* 12, 1–14. doi: 10.1038/s41467-021-26922-4

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, 1–22. doi: 10.1186/gb-2008-9-1-r7

Han, J.-Y., Kim, H.-J., Kwon, Y.-S., and Choi, Y.-E. (2011). The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in Panax ginseng. *Plant Cell Physiol.* 52, 2062–2073. doi: 10.1093/pcp/pcr150

Heitz, T., Widemann, E., Lugan, R., Miesch, L., Ullmann, P., Désaubry, L., et al. (2012). Cytochromes P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone jasmonoyl-isoleucine for catabolic turnover. *J. Biol. Chem.* 287, 6296–6306. doi: 10.1074/jbc.M111.316364

Höfer, R., Dong, L., André, F., Ginglinger, J.-F., Lugan, R., Gavira, C., et al. (2013). Geraniol hydroxylase and hydroxygeraniol oxidase activities of the CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the (seco) iridoid pathway. *Metab. Eng.* 20, 221–232. doi: 10.1016/j.ymben.2013.08.001

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* 48, 657–666. doi: 10.1038/ng.3565

Irmler, S., Schröder, G., St-Pierre, B., Crouch, N. P., Hotze, M., Schmidt, J., et al. (2000). Indole alkaloid biosynthesis in *Catharanthus roseus*: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.* 24, 797–804. doi: 10.1046/j.1365-313x.2000.00922.x

Jiang, Z., Tu, L., Yang, W., Zhang, Y., Hu, T., Ma, B., et al. (2021). The chromosome-level reference genome assembly for *Panax notoginseng* and insights into ginsenoside biosynthesis. *Plant Commun.* 2:100113. doi: 10.1016/j.xplc.2020.100113

Kim, N. H., Jayakodi, M., Lee, S. C., Choi, B. S., Jang, W., Lee, J., et al. (2018). Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Plant Biotechnol. J.* 16, 1904–1917. doi: 10.1111/pbi.12926

Kochetkoy, H. K. (1963). Chemical constituent of *Aralia elata*. *DoklAkadNauk* 50:1289.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [Preprint].

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., Deweese-Scott, C., et al. (2010). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Nafisi, M., Goregaoker, S., Botanga, C. J., Glawischnig, E., Olsen, C. E., Halkier, B. A., et al. (2007). *Arabidopsis* cytochrome P450 monooxygenase 71A13 catalyzes the conversion of indole-3-acetaldoxime in camalexin synthesis. *Plant Cell* 19, 2039–2052. doi: 10.1105/tpc.107.051383

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into capsicum domestication and specialization. *Proc. Natl. Acad. Sci. U. S. A.* 111, 5135–5140. doi: 10.1073/pnas.1400975111

Reunov, A., Reunova, G., and Zhuravlev, Y. N. (2007). Morphological study of pollen grains in mature anthers of *Aralia elata*, *A. continentalis*, and *A. cordata* (Araliaceae). *Dokl. Biol. Sci.* 417, 465–468. doi: 10.1134/s0012496607060166

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Sawai, S., and Saito, K. (2011). Triterpenoid biosynthesis and engineering in plants. *Front. Plant Sci.* 2:25. doi: 10.3389/fpls.2011.00025

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T., et al. (2020). Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* 11:971. doi: 10.1038/s41467-020-14776-1

Van De Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wu, Y. (2011). *Clong and Fundational Analysis of Triterpenoid Saponin Related Genes in Aralia elata*. ChuangChun,China: Jilin University.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., Chen, S., Wang, S., Hu, Y., Zhang, G., Dong, Y., et al. (2021). Chromosomal-scale genome assembly of *Eleutherococcus senticosus* provides insights into chromosome evolution in Araliaceae. *Mol. Ecol. Resour.* 21, 2204–2220. doi: 10.1111/1755-0998.13403

Zhang, Y., Wang, W., He, H., Song, X.-Y., Yao, G.-D., and Song, S.-J. (2018). Triterpene saponins with neuroprotective effects from a wild vegetable *Aralia elata*. *J. Funct. Foods* 45, 313–320. doi: 10.1016/j.jff.2018.04.026

Zhao, C. Y. (2012). *Clong and Genetic Transfermaiton of Key Genes Related to Triterpenoid Saponin Biosynthetic Pathway in Aralia elata*. ChuangChun, China: Jilin University.

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpls.2022.830413

Integrated Metabolome and Transcriptome Analyses Reveal Dissimilarities in the Anthocyanin Synthesis Pathway Between Different Developmental Leaf Color Transitions in *Hopea hainanensis* (Dipterocarpaceae)

Guihua Huang[1†], Xuezhu Liao[2†], Qiang Han[1], Zaizhi Zhou[1], Kunnan Liang[1], Guangyou Li[1], Guang Yang[3], Luke R. Tembrock[4], Xianbang Wang[1] and Zhiqiang Wu[2]**

[1] State Key Laboratory of Tree Genetics and Breeding, Research Institute of Tropical Forestry, Chinese Academy of Forestry, Guangzhou, China, [2] Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, [3] Guangdong Eco-Engineering Polytechnic, Guangzhou, China, [4] Department of Agricultural Biology, Colorado State University, Fort Collins, CO, United States

Changes in plant leaf color during development are directly related to the accumulation or degradation of certain phytochemicals such as anthocyanins. Since some anthocyanins can be beneficial to human health and provide insights into the biology of leaves, the underlying processes and timing by which plants produce these molecules has been the focus of numerous studies. The tree species *Hopea hainanensis* generally produces green leaves at all growth stages; however, a few explored individuals have been identified possessing red leaves on the top of the seedlings at a young stage. While the phenomenon of leaf color varying with age has been studied in several species, the underlying mechanisms are largely unknown in *H. hainanensis*. Using a metabolomics approach, the young red leaves in *H. hainanensis* were found to contain higher levels of anthocyanins and flavonoids than the young green-leaved individuals. Among anthocyanins, pelargonidin and cyanidin were the most likely candidates contributing to the red color of the young leaves. Transcriptome results indicated the genes related to the production of these anthocyanins were significantly upregulated, leading to greater accumulation of red pigments. Specifically, the expression of several *MYB* and *bHLH* genes in young red leaf lines was significantly higher than that in the young green leaf lines, especially *HhMYB66*, *HhMYB91*, *HhMYB6*, and *HhbHLH70*. As such these four transcription factors are probably the main regulatory genes resulting in young red leaves in *H. hainanensis*. From these results, comparative analyses with other species can be made to better understand the evolution of pigment biosynthesis and how anthocyanins function in plant metabolism and evolution/adaptation.

Keywords: plant pigments, leaf maturation, trees, anthocyanins, transcription factors

OPEN ACCESS

Edited by:
Saneyoshi Ueno,
Forestry and Forest Products Research Institute, Japan

Reviewed by:
Xiaopeng Fu,
Huazhong Agricultural University, China
Jinhuan Chen,
Beijing Forestry University, China

****Correspondence:***
Xianbang Wang
wangxb@caf.ac.cn
Zhiqiang Wu
wuzhiqiang@caas.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *07 December 2021*
Accepted: *07 February 2022*
Published: *03 March 2022*

Citation:
Huang G, Liao X, Han Q, Zhou Z, Liang K, Li G, Yang G, Tembrock LR, Wang X and Wu Z (2022) Integrated Metabolome and Transcriptome Analyses Reveal Dissimilarities in the Anthocyanin Synthesis Pathway Between Different Developmental Leaf Color Transitions in Hopea hainanensis (Dipterocarpaceae). Front. Plant Sci. 13:830413. doi: 10.3389/fpls.2022.830413

INTRODUCTION

Differences in leaf color have been observed by humans for millennia, especially in temperate climates where many trees reveal colorful pigments in autumn as chlorophyll is degraded and secondary pigments are made evident. Beyond the aesthetics of leaf coloration, pigment molecules play an essential role in leaf physiology and metabolism through harvesting light and providing protection from the damaging effects of UV radiation (Demmig-Adams and Iii, 1996; Chen, 2014; Rahnasto-Rilla et al., 2018). The protective effects of some plant pigment molecules, such as anthocyanins, are known to provide similar functions in human biochemistry, and thus a great deal of work has been conducted to better understand these molecules and how they can be used medicinally (Santos-Buelga et al., 2014; Demmig-Adams et al., 2020; Jideani et al., 2021). Therefore, improving our understanding of plant pigment biosynthetic pathways could have profound influences on both plant and human health. For example, the development of cultivars with specific leaf or fruit pigmentation could be improved by molecular breeding and/or genetic engineering of specific steps in pigment biosynthesis gene networks only after the pathways are properly mapped.

Hopea hainanensis Merr. et Chun (坡垒 po lei) is a tropical evergreen tree species in the Dipterocarpaceae family. The current range of this species is restricted to scattered occurrences in dense tropical forests on the island of Hainan and a small number of climactically similar locations in Northern Vietnam. This species is well-known due to the fine-grained durable hardwood it possesses, which has historically been used for a wide variety of applications such as in the production of railroad sleepers, mechanical appliances, fishing vessels, docks, bridges, and in constructing buildings. Since this species is now uncommon in the wild due to intensive logging and habitat loss (Ly et al., 2018), the uses of the wood are now mainly restricted to making decorative furniture. In addition to timber production, white aromatic resin from the trunk of *H. hainanensis* has been used in making medicine, perfume, and paint. The characterization of *H. hainanensis* for medicinal compounds has resulted in the discovery of several novel and bioactive compounds, including the acetylcholinesterase inhibitor hopeahainol (Ge et al., 2008, 2009). Moreover, *H. hainanensis* is a deep-rooted tree with tolerance to rocky and shallow soils making it highly valuable in ecosystem services related to soil and water preservation and as such it has been often used in landscaping and ecological conservation projects.

Most recent research on *H. hainanensis* has focused on conservation biology and ecology, seed germination, breeding, reintroduction, and silvicultural attributes (Xiaoying et al., 2017; Lu et al., 2020). Although some chemistry work has been conducted regarding the medical chemistry of *H. hainanensis* (Ge et al., 2008, 2009), relatively little work has been completed regarding the developmental biology of leaf pigment phytochemistry in this species.

In the course of previous studies on *H. hainanensis*, we found that some trees possessed green leaves through the developmental process, whereas others produced leaves that were initially red when first developing and eventually became green as the leaves matured. Breeding experiments between the two leaf forms indicated that the trait was heritable and as such likely not the results of a plastic gene by environment response. Given the desirable nature of this trait in the development of ornamental cultivars and the potential underlying adaptive characteristics associated with pigmentation (such as adaption to different light levels), it was determined that further work should be conducted to map the genetic pathways of these two color forms. Young red leaf forms have been studied in several plant species (Han et al., 2019; Waterland et al., 2019; El-Nakhel et al., 2020); however, it has not been characterized in *H. hainanensis*. Given this gap in knowledge, we carried out metabolome and transcriptome analyses of different colored leaves at different stages of development for *H. hainanensis* to answer the following questions regarding red leaf pigmentation: (1) what phytochemicals are responsible for the red coloration in young red leaves, (2) which genes are upregulated in the biosynthesis of red pigments, and (3) how do these attributes differ between the two young leaf forms and to other known pigment pathways? The results from these investigations will elucidate the mechanisms responsible for young red leaves in *H. hainanensis* and provide a thoroughly described trait for testing hypotheses regarding the ecology and evolution of the two leaf forms and improve our understanding of how increased red pigmentation in young leaves could influence other biosynthetic pathways especially as pertains to medicinal compounds, environmental adaptation, and photosynthetic efficiency. Young red leaf lines are considered desirable for horticultural uses, and thus knowledge about the formation of this trait can improve cultivar selection and could lead to the development of new more desirable and resilient cultivars for use in landscaping and ecological restoration. Except for cultivar characterization, the inclusion of different phytochemical variants into the germplasm repositories (especially among rare trees like *H. hainanensis*) is essential to preserve the diversity of traits for breeding and introduction.

RESULTS

Metabolites in *Hopea hainanensis*

Hopea hainanensis is an economically and ecologically important species in Dipterocarpaceae, which was found to possess young red-leaved individuals in about 10% of the individuals sampled (**Figure 1**).

To investigate what compounds contribute to the red-colored young leaves in *H. hainanensis,* the total metabolites of RU (red young leaves of red lines), RL (green mature leaves of red lines), GU (green young leaves green lines), and GL (green mature leaves of green lines) were detected by UPLC-MS/MS followed by hierarchical cluster analysis (HCA) to analyze the differences in accumulated metabolites of the different samples. The metabolites possessing similar accumulation patterns were clustered together and indicated that there are more differences in young leaves than mature leaves between the comparison of RU vs. GU and RL vs. GL. However, even in the mature leaves of the two lines, several metabolites were differentially accumulated. The HCA results also showed that some of the metabolites of RU

FIGURE 1 | Different leaf color phenotypes in *Hopea hainanensis*. Young red-leaved example on the left and standard green-leaved example on the right.

were significantly different from those of the other three sample types. Nearly half of all detected metabolites were significantly accumulated in RU, some of which are likely contribute to red young leaves (**Figure 2A**).

In addition, the results of principal component analysis (PCA) with all samples showed that principal component 1 (PC1) and principal component 2 (PC2) accounted for 59.77 and 21.57% of the total variation, respectively. In the PCA, the biological replicates were tightly clustered while the four sample types were broadly separated on the graph. The RU samples were clearly differentiated from the other three samples, yet the mature leaves from the different lines were also plotted separately (and non-overlapping) in the PCA plot. These results not only indicate the repeatability and reliability of the data, but also indicate metabolic differences by stage and type across all samples (**Figure 2B**). To resolve the key metabolites responsible for the young red leaves in *H. hainanensis*, 719 metabolites were detected when all RU, RL, GU, and GL were considered. The most abundant compounds were phenolic acids (114), flavonoids (109), lipids (105), amino acids and derivatives (77), and organic acids (74), among which the flavonoids are considered the most important in this study because of their crucial role in plant pigmentation (**Figure 2C**). Of the 719 metabolites, 636 metabolites were shared across all four sample types, with only four metabolites (including flavonoids related to red color) unique to RU (**Figure 2D**), and no specific metabolites were uniquely accumulated in the other three samples. Given that flavonoids as well as several unique flavonoids are found in higher abundance in RU, it is reasonable to suggest that compounds of this type are involved with red coloration among these samples (**Figures 2A,C,D**).

Differential Accumulation of Flavonoid Metabolites Between RU and GU

To further confirm that flavonoid metabolites were responsible for the formation of red young leaves in *H. hainanensis*, we compared the metabolites between groups (GL vs. GU, RU vs. GU, RL vs. GL, and RL vs. RU), selected different metabolites in each group based on the screening condition that the variable importance in the projection (VIP) $\geq$ 1 with $|$ Log2FC $|$ $\geq$ 1, and successfully obtained 424 differentially accumulated metabolites (DAMs) in RL vs. RU and 236 DAMs in RU vs. GU, among which 189 metabolites were shared between these two comparisons (**Figure 3A**). We found that RU had 119 metabolites with decreased accumulation and 117 with increased accumulation compared with GU. The RL samples had 224 metabolites with decreased accumulation and 200 metabolites with increased accumulation compared with RU. The RL samples had 49 metabolites with decreased accumulation and 46 metabolites with increased accumulation compared with GL. The GL samples had 183 metabolites with decreased accumulation and 151 metabolites with increased accumulation compared with GU (**Figure 3A**). We further performed k-means clustering on the DAMs in the comparison group of all the samples and divided these differentially accumulated metabolites into 9 subclasses. Among the 9 subclasses 3, 8, and 9 had unique patterns of metabolites in the RU samples. The DAMs in subclasses 3 and 9 were mainly phenolic and organic acids, while subclass 8 was mainly made up of flavonoids, indicating a high accumulation of these compounds in RU relative to GU, GL, and RL (**Figures 3B,C**). This strongly indicates that flavonoid metabolites are the primary compounds responsible for young leaves in *H. hainanensis*.

Transcriptome Analyses

To further confirm metabolite differences between samples RU, RL, GU, and GL, RNA-seq was used to assess the differences in transcript abundance and type. A total of 86,454 transcripts were obtained by Trinity. The longest cluster sequence obtained by Corset was defined as a unigene for subsequent analysis with 83,078 such unigenes generated. Comparisons of GL vs. GU, RL vs. GL, RL vs. RU, and RU vs. GU were used to select differentially expressed genes (DEGs) with the threshold "$|$ log2Fold Change$|$ $\geq$ 1 and FDR $<$ 0.05." From this 14,195 DEGs (8,203 upregulated and 5,992 downregulated) from GL vs. GU; 7,718 (4,150 upregulated and 3,568 downregulated) from RL vs. GL; 19,488 (10,839 up-regulated and 8,649 downregulated) from RL vs. RU; and 11,584 (6,289 upregulated and 5,295 downregulated) from GL vs. GU were identified (**Figure 4A**). The RL vs. RU comparison contained the greatest number of DEGs and was further scrutinized for pigment-related pathways. The KEGG enrichment results of RU vs. GU showed that DEGs

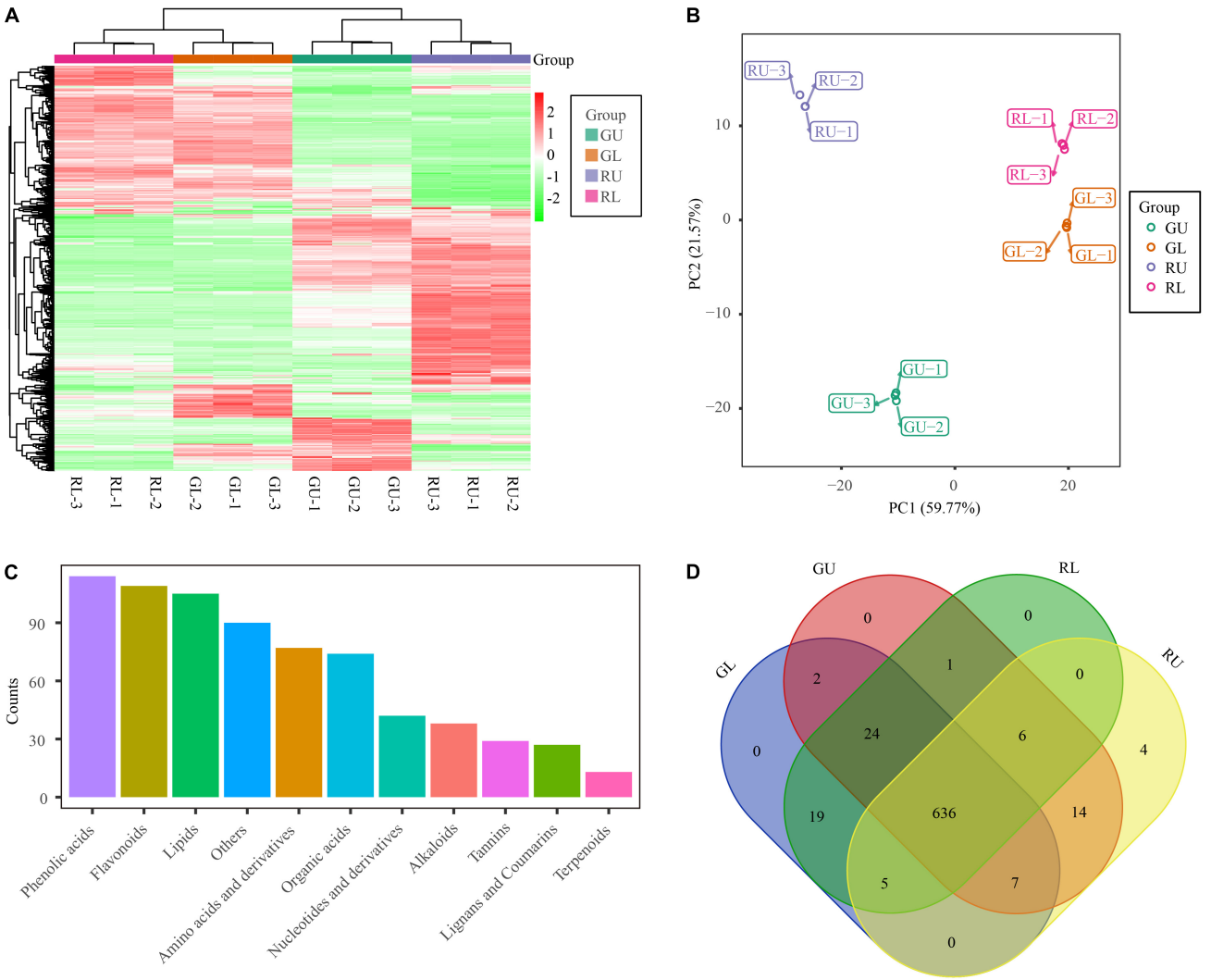


FIGURE 2 | Metabolites of *H. hainanensis*. **(A)** Heatmap of metabolites distributed in all samples and replicates. **(B)** PCA analysis of metabolites. **(C)** Abundance of different metabolite types from all samples. **(D)** Venn diagram of metabolites from all samples.

were significantly enriched in pathways related to flavonoid, stilbenoid, diarylheptanoid, and gingerol biosynthesis as well as photosynthesis (**Figure 4B**). Among the DEGs, the genes related to the flavonoid pathway were the most highly enriched in RU, providing additional evidence that these compounds were responsible for red leaf coloration in *H. hainanensis*.

Gene expression in some instances can be controlled by transcription factors (TFs). The TFs known as trans-acting factors, often regulate plant growth and development as well as the biosynthesis of secondary metabolites by activating or inhibiting gene expression (Latchman, 1993; Schwechheimer and Bevan, 1998). Among them, MYB and bHLH play an important role in the regulation of plant pigment (flavonoid-anthocyanin) biosynthesis (Zhuang et al., 2019). In our study, a total of 1,151 TFs were identified, with the most abundant including AP2/ERF-ERF (88), MYB-related (80), C2H2 (76), WRKY (72), bHLH (55), MYB (52), GRAS (52), NAC (51), and bZIP (44). Within the top 20 TF types, the transcription factors related to flower color in plants such as MYB accounted for 6.11% and bHLH 6.46% of all TFs (**Figure 4C**). Given that MYB and bHLH are known to be involved with anthocyanin accumulation, it was not surprising to find that some of these TFs were found in significantly higher abundance in RU. Specifically, three MYB TFs *HhMYB66* (Cluster-15967.19622), *HhMYB91* (Cluster-15967.12496), and *HhMYB6* (Cluster-15967.32152) and one bHLH TF *HhbHLH70* (Cluster-15967.18674) were found to be significantly upregulated in RU. The other MYB and bHLH TFs were also upregulated in RU (**Figure 4D**). Additionally, two bHLH TFs, Cluster-15967.47825 and Cluster-27423.0, were highly expressed in RL and GL, respectively, suggesting a putative role in leaf development (**Figure 4D**). Thus, increases in red leaf pigments among RU samples appears to be driven in part by the increased abundance of anthocyanin-related TFs.

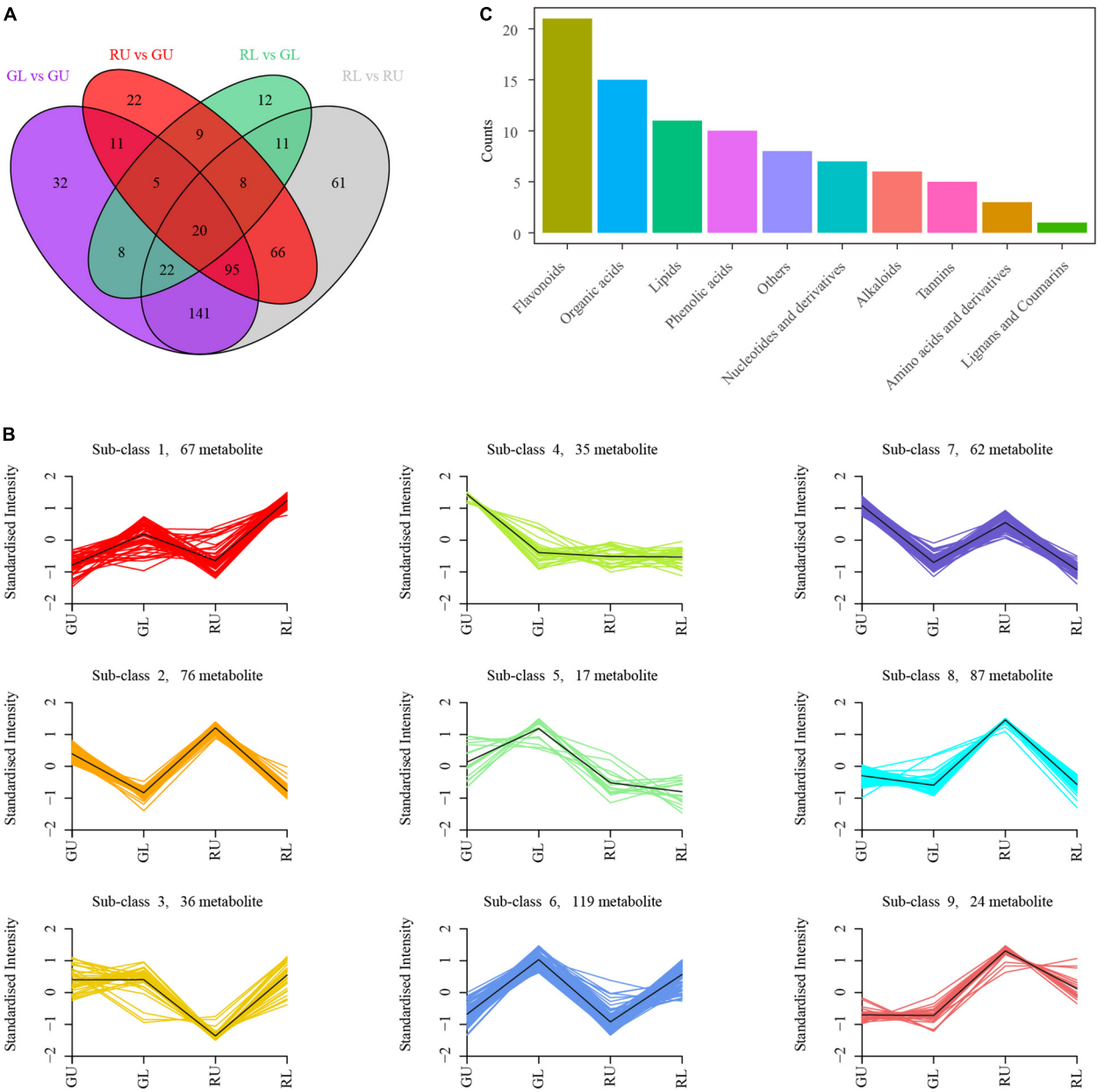


FIGURE 3 | DAMs in *H. hainanensis*. **(A)** Venn diagram of DAMs from all comparisons. **(B)** k-means clustering analysis of total differential metabolites. **(C)** DAMs abundance by type in subclass 8.

Association Analysis of Genes and Metabolites Related to Red Young Leaf Formation in *Hopea hainanensis*

As flavonoids are the main DAMs in the red leaves of *H. hainanensis*, we selectively compared DAMs of phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), and anthocyanin biosynthesis (ko00942) to assess RU vs. GU and RL vs. RU. From these comparisons, 10 metabolites were found to be significantly accumulated in RU, including the anthocyanin pelargonidin-3-O-glucoside, the dihydroflavonol naringenin-7-O-glucoside, and the flavonoid myricetin as well as others in each of these three classes (**Figure 5**). Metabolite pme3392 was annotated as anthocyanin, pelargonidin-3-O-glucoside, which is a pigment known to produce red coloration and is found in the highest relative abundance in RUs (**Figure 5B**). Therefore, we suggest that this may be the most important metabolite contributing to the coloration of red young leaves.

Based on the results of metabolomic analysis, anthocyanins were the main metabolites found to differ in the red young leaves

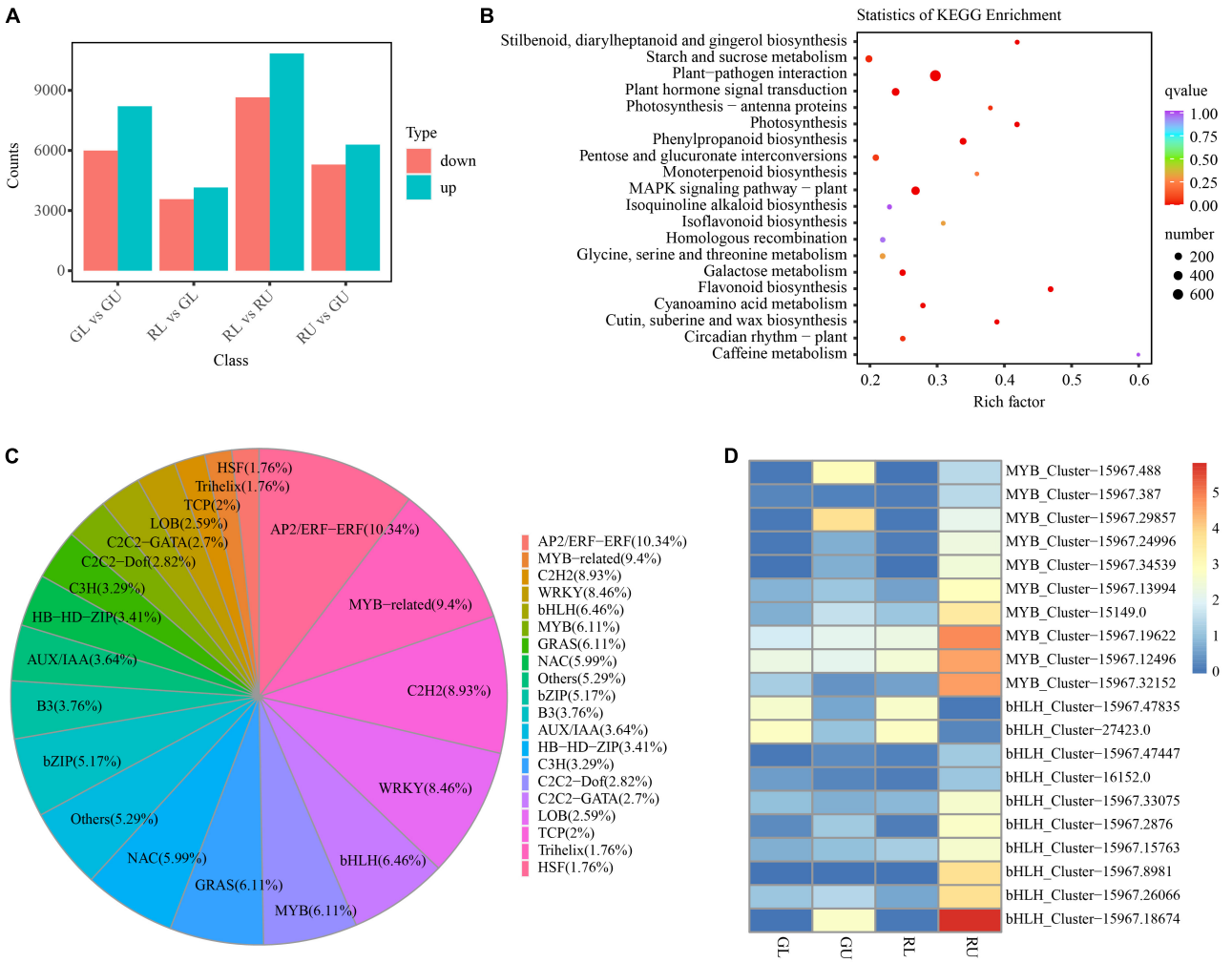


FIGURE 4 | DEGs in *H. hainanensis*. **(A)** DEGs from different comparisons. **(B)** The KEGG enrichment of DEGs in RU vs. GU (*Q* values are corrected *p* values scaled by the number of comparisons, with 0 being highly significant and 1 being not significant). **(C)** Distribution of all differentially expressed transcription factors. **(D)** heatmap of the MYB and bHLH transcription factors from RU vs. GU and RL vs. RU with | log2Fold Change| ≥ 2, colors indicate log2Fpkm.

of *H. hainanensis*. Thus, we focused on anthocyanin biosynthesis. DEGs associated with anthocyanin biosynthesis pathway with | log2Fold Change| ≥ 2 in comparisons of RU vs. GU and RL vs. RU were also selected for further analysis. From this, a total of 30 DEGs (**Figure 6**) were found, of which 2 *PAL* genes, 2 *C4H* genes, and 1 *4CL* gene, which are involved in precursor biosynthesis of flavonoids, were significantly upregulated in RU. Similarly, the anthocyanidin-associated genes, *CHS*, *CHI*, *F3H*, *F3'H*, *DFR*, and *ANS*, were found to have higher expression levels in RU. In addition, two *FLS* genes which are the key genes in the conversion of dihydroflavonols to flavonols were also significantly upregulated in RU. Interestingly, some downstream *UFGT* genes involved in the transport of anthocyanidins were downregulated in RU. When the DEGs and DAMs associated with the anthocyanin biosynthesis pathway in *H. hainanensis* were analyzed separately from the total data set, we found that among the previously screened 30 DEGs, there were 17 genes that had a significant correlation to 3 metabolites, namely dihydroquercetin (mws0044), dihydrokaempferol (mws1094), and naringenin-7-o-glucoside (mws1179), in the comparison of RU vs. GU (Pearson correlation coefficient (PCC) ≥ 0.8).

DISCUSSION

Hopea hainanensis is a valuable tree that produces high-quality wood and has numerous ecological benefits, but has become increasingly rare in recent decades, which makes the study and preservation of this species even more important. In the process of investigation and research, we found that most of the mature leaves and young leaves of *H. hainanensis* in natural forests are green, while a small minority were found to have green mature leaves and young red leaves. From breeding work, it was found that this trait was fixed and not the result of phenotypic plasticity.

From previous studies, it is well known that the red pigments in most flowers, fruits, and seeds are flavonoid compounds

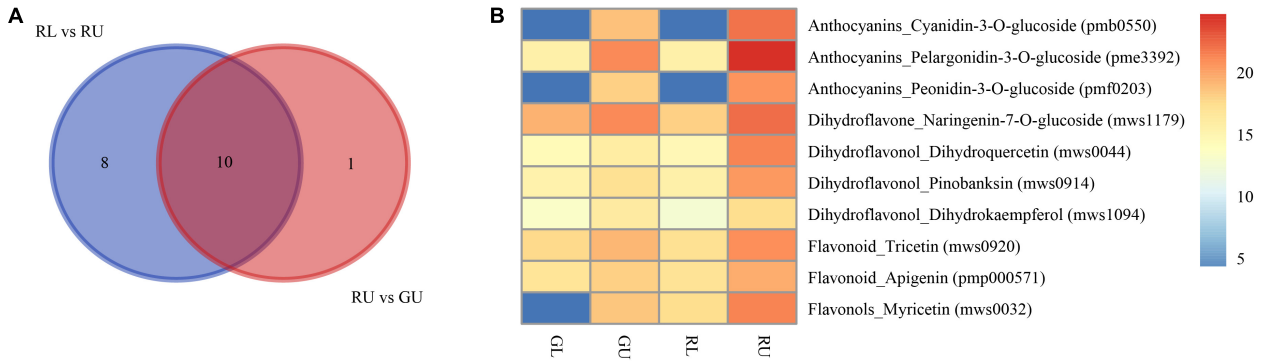


FIGURE 5 | Metabolites related to red young leaves in *H. hainanensis*. **(A)** Venn diagram indicating the DAMs enriched into phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), and anthocyanin biosynthesis (ko00942) in the comparison of RU vs. GU and RL vs. RU. **(B)** Relative abundance of red color related metabolites in RU vs. GU and RL vs. RU indicates log2Relative abundance.

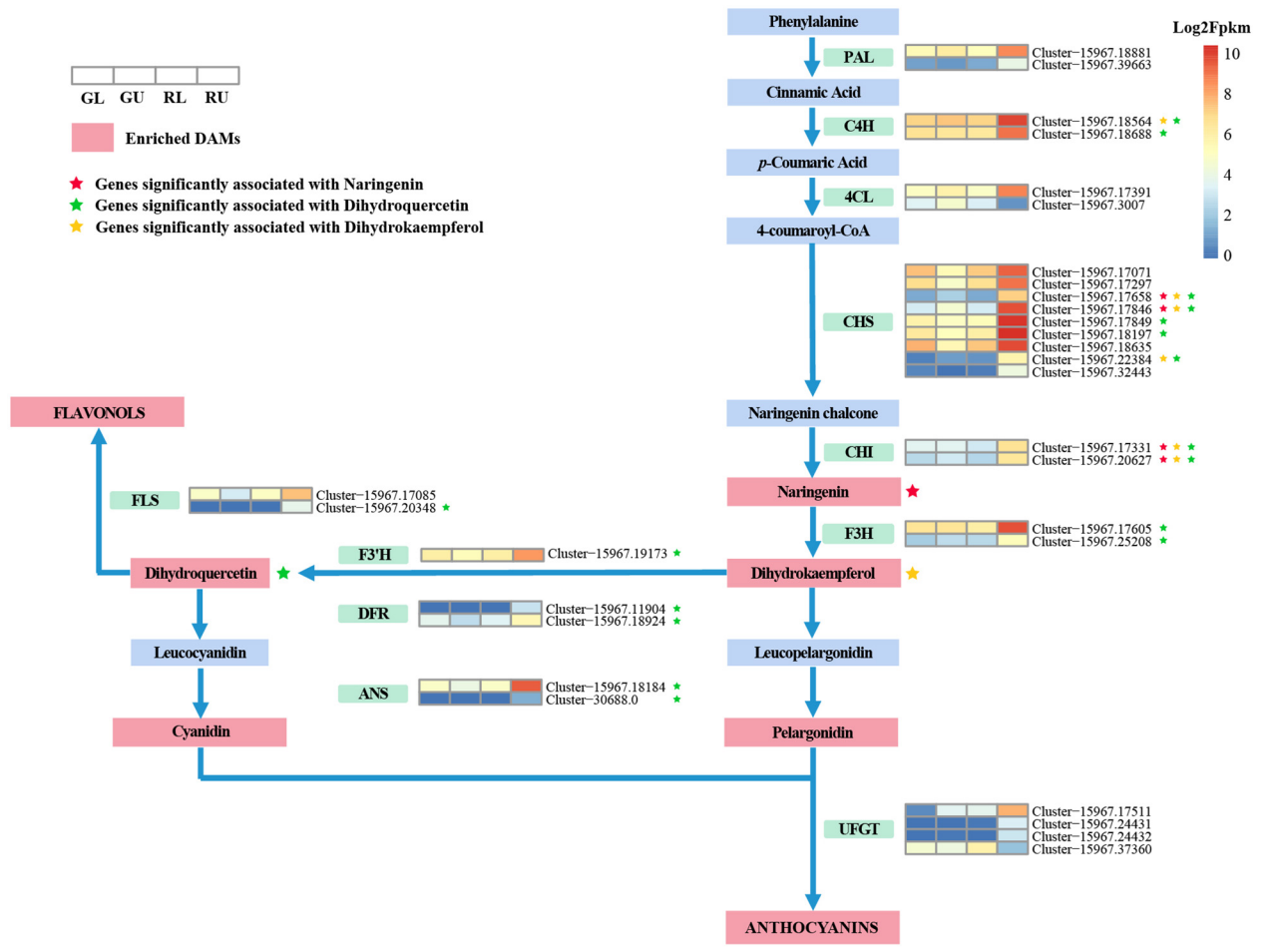


FIGURE 6 | Anthocyanin biosynthetic pathway in the young red leaves of *H. hainanensis*. Metabolites, in red boxes indicate the DAMs enriched in phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), and anthocyanin biosynthesis (ko00942) between RU vs. GU and RL vs. RU. Heatmaps represent DEGs enriched in anthocyanin biosynthesis pathway with | log2Fold Change| ≥ 2 in the comparison of RU vs. GU and RL vs. RU. Heatmaps are ordered by GL, GU, RL, and RU from left to right, with colors indicating the log2Fpkm relative expression. Colored stars indicate a significant correlation between an increase in metabolite accumulation and an increase in the expression of an associated gene. Enzyme abbreviations: *PAL*, phenylalanine ammonium lyase; *C4H*, cinnamate-4-hydroxylase; *4CL*, 4-coumaroyl-CoA synthase; *CHS*, Chalcone synthase; *CHI*, chalcone isomerase; *F3H*, flavanone 3-hydroxylase; *F3′H*, flavonoid-3′-hydroxylase; *DFR*, dihydroflavonol 4-reductase; *ANS*, anthocyanidin synthase; *UFGT*, UDP-glucose flavonoid 3-O glucosyltransferase; *FLS*, flavonol synthase.

(Falcone Ferreyra et al., 2012). In this study, the metabolome of *H. hainanensis* leaves were comprehensively analyzed, and over 700 compounds were identified. The metabolome analyses showed that pigments, such as flavonoids were significantly accumulated in the leaf forms differently than those in young green-leaved forms. This indicated that the difference in leaf color mainly resulted from the differences in the accumulation of pigments. Leaf pigments are essential compounds in the metabolic functioning of leaves. Better understanding of the pathways by which these compounds are produced helps researchers to elucidate the evolution of plant pigments and infer their function as well as provide a blueprint for making alterations to a given pathway, resulting in selective and/or increased production of desired compounds. Chalcone synthase (*CHS*) is the first step in controlling flavonoid biosynthesis (Winkel-Shirley, 2001; Chaudhary et al., 2016), and it is the first enzyme identified in the flavonoid biosynthetic pathway. *CHS* is an important regulatory gene located upstream in the flavonoid biosynthesis pathway, and its overexpression may positively affect the expression of downstream *CHI* genes that affect the production of flavonoids (Kreuzaler and Hahlbrock, 1972; Awasthi et al., 2016). Here we saw the upregulation of *CHS*, which is accompanied by the upregulation of *CHI* and its downstream counterpart naringenin. In the anthocyanin biosynthetic pathway, dihydrokaempferol is a key intermediate product that can be converted to dihydroquercetin by the enzyme *F3'H*, and/or converted to dihydromyricetin by *F3'5'H* (Falcone Ferreyra et al., 2012). Here, we found that two *F3H* genes were upregulated, and dihydrokaempferol increased. An *F3'H* gene was found to be highly expressed along with increases in the accumulation of the direct downstream product in the red young leaves. This result is similar to a study that determines the pathway by which red longan (*Dimocarpus longan*) fruits were produced. Yi et al. (2021) revealed that genes related to enzymes leading up to dihydrokaempferol were significantly upregulated in red pericarp longan fruits. While the pathways to produce red pigments might be similar between these two distantly related species, the timing for activating them is quite different as the pathway is initiated in young leaves of *H. hainanensis* and in maturing fruits of longan. Apart from these intermediate products mentioned above, their downstream genes and products were also upregulated in *H. hainanensis*. On the whole, we found *C4H*, *CHS*, *CHI*, *F3H*, *F3'H*, *DFR*, *ANS,* and *FLS* were positively correlated with differentially accumulated metabolic intermediates in the comprehensive analysis of the metabolome and transcriptome, and these genes possessed significantly higher expression levels, which is consistent with results in the purple-leaved Jujube (Li et al., 2021). These results suggest that the high expression of *C4H*, *CHS*, *CHI*, *F3H,* and *F3'H* could promote the accumulation of dihydroquercetin (mws0044) and dihydrokaempferol (mws1094). The accumulation of these two metabolites will recruit and promote the expression of downstream genes, such as *DFR*, *ANS*, and *FLS*, to further generate downstream products. In addition, we found that in the anthocyanin biosynthesis of RU vs. RL, some *UFGT* genes (cluster-15967.49721) were negatively correlated with pelargonidin-3-o-glucoside (−0.831) as the log2Fold Change of this gene in RU vs. GU was −1.83. Downregulation of some *UFGT* genes was also found in red pericarp longan fruits (Yi et al., 2021) and thought to be involved with increased anthocyanin accumulation. Furthermore, the high expression of *FLS* and the accumulation of the flavonoid apigenin (pmp000571) also revealed that flavonoids, in addition to anthocyanins, pelargonidin, and cyanidin, are also responsible for the red color of young red-leaved lines.

Flavonoid biosynthesis genes are often regulated by the interaction of transcription factors in different gene families. In monocotyledons and dicotyledons, genes involved in the anthocyanin biosynthesis pathway are differentially regulated by R2R3-MYB transcription factors, bHLH, and WD40 proteins (Grotewold, 2005; Petroni and Tonelli, 2011; Xu et al., 2015; Zhuang et al., 2019). Therefore, the combination of R2R3-MYB, bHLH, and WD40 transcription factors and their interactions (MYB-bHLH-WD40 complex) determine the activation and spatio-temporal expression of anthocyanin synthesis structural genes (Petroni and Tonelli, 2011). Unsurprisingly, our results also showed similarities to other studies in apple and grape (Walker et al., 2007; An et al., 2020), wherein the TFs, MYB and bHLH, are the key regulators of anthocyanin biosynthesis and accumulation. The MYB and bHLH transcription factors are present in all eukaryotes and are the two largest families of plant transcription factors (Feller et al., 2011). More than 20 years ago, the first transcription factor encoding proteins in the MYB domain required for anthocyanin synthesis was discovered in plants (Paz-Ares et al., 1987). Since then more studies have found that MYB transcription factors, such as some R2R3-MYBs, including AtMYB75/PAP1, AtMYB90/PAP2, AtMYB113, and AtMYB114 that control anthocyanin biosynthesis in vegetative tissues play an important role in flavonoid biosynthesis (Gonzalez et al., 2008). As for the bHLH transcription factors, the abnormal expression of *bHLH3* in mulberry fruits was shown to disrupt the balance of the flavonoid metabolic network, leading to changes in the content and proportion of anthocyanins, flavonoids, and flavonols in the different colored mulberry fruits (Li et al., 2020). In this process, MYB transcription factors that are mainly involved in the regulation of flavonoid biosynthetic genes, for example, MYB transcription factors SbY1 in *Sorghum bicolor,* may have an impact on the expression of *CHS*, *CHI,* and *DFR*, resulting in regulation of 3-deoxyflavonoid biosynthesis (Du et al., 2009). In *Solanum tuberosum*, StD is thought to regulate the *F3H*, *DFR,* and *F3'5'H* genes, and thus play a role in the regulation of anthocyanin biosynthesis (Jung et al., 2009). In our study, we also found that three MYB TFs *HhMYB66* (Cluster-15967.19622), *HhMYB91* (Cluster-15967.12496), and *HhMYB6* (Cluster-15967.32152), and one bHLH TF *HhbHLH70* (Cluster-15967.18674) were significantly in higher abundance in RU. Moreover, *CHS*, *CHI*, *F3H*, *F3'H,* and *DFR* were also upregulated, implying that these genes may be the key genes regulating flavonoid/anthocyanin synthesis in *H. hainanensis* red leaves.

By combining the results of metabolome and transcriptome data we suggest that there are two main anthocyanin biosynthesis pathways after branching from dihydrokaempferol, and the significant upregulation of genes in these pathways leads to an increased accumulation of anthocyanin and flavonoid

derivatives, which together are responsible for the young red leaves of *H. hainanensis*. Now, the red pigments responsible for young red leaves have been identified, while the follow-up work examining the simultaneously upregulated genes related to stilbenoid production and photosynthesis (found in this study) needs to be conducted to uncover potentially synergistic roles of these phytochemicals in leaf metabolism.

MATERIALS AND METHODS

Plant Materials

During the investigation and cultivation of *H. hainanensis*, we found that most *H. hainanensis* have green mature and young leaves, whereas a small number of individuals possess mature green leaves and young red leaves. After we collected seeds and raised seedlings individually (the seedlings from one tree's seeds are regarded as a line), it was found that their offspring had the same leaf-color, and this pattern was maintained after reproduction, which indicated that genetic differences controlled the leaf color trait. To investigate the metabolomic and transcriptomic differences between the red and green leaf lines, we chose one red leaf individual (referred to as Line 5 in our lab breeding nomenclature) and one green leaf pedigree (referred to as Line 10) for further analyses. The red-leafed individuals were found during the course of forest surveys. Both red and green individuals have been grown from seeds for 2 years in the outdoor common garden at the Research Institute of Tropical Forestry (Chinese Academy of Forestry).

Next, we selected three independent samples from each pedigree to represent the biological duplications (referred to as Red1, Red2, and Red3; Green1, Green2, and Green3). To optimize variation among red and green leaves, we sampled the leaves along the axis of the stem from the 2nd expanded leaf (young leaves) to the 6th unfolded leaf (mature leaves) with a total of 12 samples. Duplicated samples from young red-leaved lines (RU), green mature leaves in red lines (RL), young green leaves in green lines (GU), and green mature leaves in green lines (GL) were harvested and stored at −80°C after snap-freezing with liquid nitrogen.

UPLC-MS/MS Instrumentation and Analyses

The freeze-dried leaves were crushed using a mixer mill (MM 400, Retsch, Haan, German) with a zirconia bead for 1.5 min at 30 Hz. From this 100 mg powder was weighed and extracted overnight at 4°C with 1.2 mL 70% aqueous methanol. Following centrifugation at 12,000 rpm for 10 min, the supernatant was removed and syringe-filtered (SCAA-104, 0.22μm pore size; ANPEL, Shanghai, China)[1] before UPLC-MS/MS analysis.

The sample extracts were analyzed using a UPLC-ESI-MS/MS system (UPLC, SHIMADZU Nexera X2[2]; MS, Applied Biosystems 4500 Q TRAP[3]). The analytical conditions were as follows for UPLC: column, Agilent SB-C18 (1.8 μm, 2.1 mm*100 mm); the mobile phase consisted of solvent A, pure water with 0.1% formic acid, and solvent B, acetonitrile with 0.1% formic acid. Sample measurements were performed with a gradient program that employed the starting conditions of 95% A, 5% B. Within 9 min, a linear gradient to 5% A, 95% B was programmed, and a composition of 5% A, 95% B was kept for 1 min. Subsequently, a composition of 95% A, 5% B was adjusted within 1.10 min and kept for 2.9 min. The column oven was set to 40°C and the injection volume was 4 μL. The effluent was injected into an ESI-triple quadrupole-linear ion trap (QTRAP)-MS.

The mass spectrometry analysis followed the method of Chen et al. (2013). Linear ion trap (LIT) and triple quadrupole (QQQ) scans were acquired on a triple quadrupole-linear ion trap mass spectrometer (Q TRAP), AB4500 Q TRAP UPLC/MS/MS system, equipped with an ESI Turbo Ion-Spray interface and operating in positive and negative ion modes and controlled by Analyst 1.6.3 software (AB Sciex). The ESI source operation parameters were as follows: ion source turbo spray; source temperature 550°C; ion spray voltage (IS) 5,500 V (positive ion mode)/-4,500 V (negative ion mode); ion source gas I (GSI), gas II (GSII), and curtain gas (CUR) were set at 50, 60, and 25.0 psi, respectively; the collision gas (CAD) was set to high. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. The QQQ scans were acquired as MRM experiments with collision gas (nitrogen) set to medium. DP and CE for individual MRM transitions were done further with DP and CE optimization. A specific set of MRM transitions were monitored for each period according to the metabolites eluted within this period (Falcone Ferreyra et al., 2012).

Unsupervised PCA (principal component analysis) (Chen et al., 2009) was performed by statistics function prcomp within R[4]. The data was normalized before unsupervised PCA. The hierarchical cluster analysis (HCA) results of samples and metabolites were presented as heatmaps with dendrograms, while the Pearson correlation coefficients (PCC) between samples were calculated by the cor function in R and presented as heatmaps. Both HCA and PCC were carried out by R package pheatmap[5]. For HCA, normalized signal intensities of metabolites (unit variance scaling) are visualized as a color spectrum. Significantly accumulated metabolites between groups were determined by VIP $\geq$ 1 and absolute Log2FC (fold change) $\geq$ 1. VIP values were extracted from orthogonal partial least squares discriminant analysis (OPLS-DA) results, which also contain score plots and permutation plots, and was generated using R package MetaboAnalystR (Chong and Xia, 2018). Data were log-transformed (log2) and mean centered before OPLS-DA. To avoid overfitting, a permutation test (200 permutations) was performed. Identified metabolites were annotated using the KEGG compound database[6] (Kanehisa and Goto, 2000) and

[1] http://www.anpel.com.cn

[2] www.shimadzu.com.cn

[3] www.appliedbiosystems.com.cn

[4] www.r-project.org

[5] https://cran.r-project.org/web/packages/pheatmap/index.html

[6] http://www.kegg.jp/kegg/compound

the annotated metabolites were then mapped to the KEGG pathway database[7]. Pathways with significantly accumulated metabolites were then fed into metabolite sets enrichment analysis (MSEA), with significance determined by *p*-values from hypergeometric tests.

RNA Sequencing and Differentially Expressed Genes Analysis

Total RNA was extracted from the young and mature leaves of *H. hainanensis* red and green lines. RNA-Seq was performed by Biomarker Technologies Co., Ltd. (Beijing, China). Sequencing libraries were generated using the NEBNext®UltraTM RNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, MA, United States) following the manufacturer's recommendations. Sample-specific indices were added during library preparation such that sequencing could be multiplexed. The library preparations were sequenced on an Illumina HiSeq X platform (Illumina, Inc., San Diego, CA, United States).

After filtering and quality control of the raw sequence data, clean reads were *de novo* assembled by Trinity v2.6.6 (Grabherr et al., 2011; Haas et al., 2013), and Corset v 1.07 (Davidson and Oshlack, 2014) was used to perform hierarchical clustering of the transcripts by comparing the number of reads and expression patterns of the transcripts. The transcript sequences obtained by Trinity were used as the reference sequence, and the longest Cluster sequence obtained by Corset hierarchical clustering was used as a unigene for subsequent analysis. To conduct the gene function annotation, the unigene sequences were compared with KEGG, NR, Swiss-Prot, GO, COG/KOG, and Trembl databases using the BLAST software (Altschul et al., 1990). Then the amino acid sequences predicted from unigenes were compared with Pfam database by HMMER software (Mistry et al., 2013) to obtain unigene annotation information. Plant-transcription factor prediction was performed using iTAK software (Zheng et al., 2016), which integrates PLNTFDB and PLANTTFDB with TFs identified by hmmscan.

The transcriptome assembled by Trinity was used as a reference sequence, and the clean reads of each sample were mapped on the reference to calculate the mapping rate of each sample with bowtie2 in RSEM. DESeq2 v1.22.2 (Varet et al., 2016) was used to obtain the DEGs between any two biological conditions. After that the Benjamini-Hochberg method was used to conduct multiple hypothesis testing correction on the hypothesis test probability (*p*-value) to obtain the false discovery rate (FDR). DEGs were selected under the conditions of $|\log2\text{Fold Change}| \geq 1$ and $\text{FDR} < 0.05$. Finally, we analyzed the KEGG pathway enrichment in all DEGs and drew the corresponding network regulation pathway map.

CONCLUSION

Integrated metabolome and transcriptome analyses of young and mature leaves of red and green lines of *H. hainanensis* revealed that the accumulation of anthocyanins and flavonoids in RU was significantly higher in red samples than in the three green samples. The main color-related products in red leaves were anthocyanins and flavonoids, among which pelargonidin and cyanidin were the most important metabolites. Transcriptome results showed that the key genes in the anthocyanin pathway in RU were expressed at significantly higher levels compared to green-leaved samples, leading to the greater accumulation of red-colored metabolites. In addition to the structural genes in the anthocyanin/flavonoid pathway, the TFs MYB and bHLH were highly expressed in RU over the three green-leaved samples. Three MYBs *HhMYB66* (Cluster-15967.19622), *HhMYB91* (Cluster-15967.12496), and *HhMYB6* (Cluster-15967.32152), and one bHLH *HhbHLH70* (Cluster-15967.18674) were expressed at significantly higher levels in RU. As such these four TFs are likely essential to initiate the pathway associated with increased anthocyanin and flavonoids found in RU. While the anthocyanin pathway in *H. hainanensis* is similar to others described from distantly related plant lineages, it is also different in several aspects, such as the absence of delphinidin derivatives. The anthocyanin/flavonoid pathway described here provides a valuable resource for the study of leaf pigment evolution as well as in the development of red-leaved cultivars for use in horticultural and forestry applications.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI; PRJNA795198.

AUTHOR CONTRIBUTIONS

GH designed and supervised implementation of the studies. GL and GY were responsible for sampling. XL and QH supervised the statistical analyses, constructed the tables, and wrote the manuscript. ZZ and KL cultivated the seedlings of *H. hainanensis*. XW and ZW carried out all technical aspects and crafted the final version. LT provided suggestions on various aspects of the study and helped edit the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the "Fundamental Research Funds for the Central Non-profit Research Institution of CAF" to GH (Grant No. CAFYBB2020SZ005) and the Chinese Academy of Agricultural Sciences Elite Youth Program to ZW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.830413/full#supplementary-material

[7]http://www.kegg.jp/kegg/pathway.html

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

An, J. P., Wang, X. F., Zhang, X. W., Xu, H. F., Bi, S. Q., You, C. X., et al. (2020). An apple MYB transcription factor regulates cold tolerance and anthocyanin accumulation and undergoes MIEL1-mediated degradation. *Plant Biotechnol. J.* 18, 337–353. doi: 10.1111/pbi.13201

Awasthi, P., Mahajan, V., Jamwal, V. L., Kapoor, N., Rasool, S., Bedi, Y. S., et al. (2016). Cloning and expression analysis of chalcone synthase gene from *Coleus forskohlii*. *J. Genet.* 95, 647–657. doi: 10.1007/s12041-016-0680-8

Chaudhary, P. R., Bang, H., Jayaprakasha, G. K., and Patil, B. S. (2016). Variation in Key Flavonoid Biosynthetic Enzymes and Phytochemicals in 'Rio Red' Grapefruit (*Citrus paradisi* Macf.) during Fruit Development. *J. Agric. Food Chem.* 64, 9022–9032. doi: 10.1021/acs.jafc.6b02975

Chen, M. (2014). Chlorophyll Modifications and Their Spectral Extension in Oxygenic Photosynthesis. *Annu. Rev. Biochem.* 83, 317–340. doi: 10.1146/annurev-biochem-072711-162943

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., et al. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* 6, 1769–1780. doi: 10.1093/mp/sst080

Chen, Y. H., Zhang, R. P., Song, Y. M., He, J. M., Sun, J. H., Bai, J. F., et al. (2009). RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: finding potential biomarkers for breast cancer. *Analyst* 134, 2003–2011. doi: 10.1039/b907243h

Chong, J., and Xia, J. (2018). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* 34, 4313–4314. doi: 10.1093/bioinformatics/bty528

Davidson, N. M., and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genom. Biol.* 15:410. doi: 10.1186/s13059-014-0410-6

Demmig-Adams, B., and Iii, W. W. A. (1996). The role of xanthophyll cycle carotenoids in the protection of photosynthesis. *Trends Plant Sci.* 1, 21–26. doi: 10.1016/s1360-1385(96)80019-7

Demmig-Adams, B., Lopez-Pozo, M., Stewart, J. J., and Adams, W. W. III (2020). Zeaxanthin and Lutein: photoprotectors, Anti-Inflammatories, and Brain Food. *Molecules* 25:3607. doi: 10.3390/molecules25163607

Du, H., Zhang, L., Liu, L., Tang, X. F., Yang, W. J., Wu, Y. M., et al. (2009). Biochemical and molecular characterization of plant MYB transcription factor family. *Biochemistry* 74, 1–11. doi: 10.1134/s0006297909010015

El-Nakhel, C., Pannico, A., Graziani, G., Kyriacou, M. C., Giordano, M., Ritieni, A., et al. (2020). Variation in macronutrient content, phytochemical constitution and *in vitro* antioxidant capacity of green and red butterhead lettuce dictated by different developmental stages of harvest maturity. *Antioxidants* 9:300. doi: 10.3390/antiox9040300

Falcone Ferreyra, M. L., Rius, S. P., and Casati, P. (2012). Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Front. Plant Sci.* 3:222. doi: 10.3389/fpls.2012.00222

Feller, A., Machemer, K., Braun, E. L., and Grotewold, E. (2011). Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *Plant J.* 66, 94–116. doi: 10.1111/j.1365-313X.2010.04459.x

Ge, H. M., Yang, W. H., Zhang, J., and Tan, R. X. (2009). Antioxidant oligostilbenoids from the stem wood of *Hopea hainanensis*. *J. Agric. Food Chem.* 57, 5756–5761. doi: 10.1021/jf900756d

Ge, H. M., Zhu, C. H., Shi da, H., Zhang, L. D., Xie, D. Q., Yang, J., et al. (2008). Hopeahainol A: an acetylcholinesterase inhibitor from *Hopea hainanensis*. *Chemistry* 14, 376–381. doi: 10.1002/chem.200700960

Gonzalez, A., Zhao, M., Leavitt, J. M., and Lloyd, A. M. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* 53, 814–827. doi: 10.1111/j.1365-313X.2007.03373.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–U130. doi: 10.1038/nbt.1883

Grotewold, E. (2005). Plant metabolic diversity: a regulatory perspective. *Trends Plant Sci.* 10, 57–62. doi: 10.1016/j.tplants.2004.12.009

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Han, T., Wang, J., Ren, H., Yi, H., Zhang, Q., and Guo, Q. (2019). Changes in defense traits of young leaves in subtropical forests succession. *Plant Ecol.* 220, 305–320. doi: 10.1007/s11258-019-00916-1

Jideani, A., Silungwe, H., Takalani, T., Omolola, A., Udeh, H., and Anyasi, T. (2021). Antioxidant-rich natural fruit and vegetable products and human health. *Int. J. Food Prop.* 24, 41–67.

Jung, C. S., Griffiths, H. M., De Jong, D. M., Cheng, S., Bodis, M., Kim, T. S., et al. (2009). The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. *Theor. Appl. Genet.* 120, 45–57. doi: 10.1007/s00122-009-1158-3

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kreuzaler, F., and Hahlbrock, K. (1972). Enzymatic synthesis of aromatic compounds in higher plants: formation of naringenin (5,7,4' -trihydroxyflavanone) from p-coumaroyl coenzyme A and malonyl coenzyme A. *FEBS Lett.* 28, 69–72. doi: 10.1016/0014-5793(72)80679-3

Latchman, D. S. (1993). Transcription factors: an overview. *Int. J. Exp. Pathol.* 74, 417–422.

Li, H., Yang, Z., Zeng, Q. W., Wang, S. B., Luo, Y. W., Huang, Y., et al. (2020). Abnormal expression of bHLH3 disrupts a flavonoid homeostasis network, causing differences in pigment composition among mulberry fruits. *Hort. Res.* 7:83. doi: 10.1038/s41438-020-0302-8

Li, S., Deng, B., Tian, S., Guo, M., Liu, H., and Zhao, X. (2021). Metabolic and transcriptomic analyses reveal different metabolite biosynthesis profiles between leaf buds and mature leaves in *Ziziphus jujuba* mill. *Food Chem.* 347:129005. doi: 10.1016/j.foodchem.2021.129005

Lu, X., Zang, R., Ding, Y., Huang, J., and Xu, Y. (2020). Habitat characteristics and its effects on seedling abundance of *Hopea hainanensis*, a Wild Plant with Extremely Small Populations. *Biodivers. Sci.* 28, 289–295.

Ly, V., Nanthavong, K., Pooma, R., Hoang, V. S., Khou, E., and Newman, M. F. (2018). Hopea hainanensis. *IUCN Red List Threat. Species* 2018:e.T32357A2816074.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121. doi: 10.1093/nar/gkt263

Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P. A., and Saedler, H. (1987). The regulatory c1 locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J.* 6, 3553–3558.

Petroni, K., and Tonelli, C. (2011). Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* 181, 219–229. doi: 10.1016/j.plantsci.2011.05.009

Rahnasto-Rilla, M., Tyni, J., Huovinen, M., Jarho, E., Kulikowicz, T., Ravichandran, S., et al. (2018). Natural polyphenols as sirtuin 6 modulators. *Sci. Rep.* 8:4163. doi: 10.1038/s41598-018-22388-5

Santos-Buelga, C., Mateus, N., and De Freitas, V. (2014). Anthocyanins. Plant pigments and beyond. *J. Agric. Food Chem.* 62, 6879–6884. doi: 10.1021/jf501950s

Schwechheimer, C., and Bevan, M. (1998). The regulation of transcription factor activity in plants. *Trends Plant Sci.* 3, 378–383. doi: 10.1016/S1360-1385(98)01302-8

Varet, H., Brillet-Gueguen, L., Coppee, J. Y., and Dillies, M. A. (2016). SARTools: a DESeq2-and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS One* 11:e0157022. doi: 10.1371/journal.pone.0157022

Walker, A. R., Lee, E., Bogs, J., McDavid, D. A. J., Thomas, M. R., and Robinson, S. P. (2007). White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J.* 49, 772–785. doi: 10.1111/j.1365-313X.2006.02997.x

Waterland, N., Moon, Y., Tou, J., Kopsell, D., Kim, M. J., and Park, S. (2019). Differences in Leaf Color and Stage of Development at Harvest Influenced Phytochemical Content in Three Cultivars of Kale (*Brassica oleracea* L. and *B. napus*). *J. Agric. Sci.* 11:14.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* 126, 485–493. doi: 10.1104/pp.126.2.485

Xiaoying, W., Fangfang, H., Xianhua, G., Weiqiang, Z., Yuhui, H., Xiuyu, X., et al. (2017). Introduction Performance of *Hopea hainanensis* and *Vatica mangachapoi* in Guangdong Tree Park. *Guangdong For. Sci. Technol.* 33, 52–56.

Xu, W., Dubos, C., and Lepiniec, L. (2015). Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends Plant Sci.* 20, 176–185. doi: 10.1016/j.tplants.2014.12.001

Yi, D., Zhang, H., Lai, B., Liu, L., Pan, X., Ma, Z., et al. (2021). Integrative Analysis of the Coloring Mechanism of Red Longan Pericarp through Metabolome and Transcriptome Analyses. *J. Agric. Food Chem.* 69, 1806–1815. doi: 10.1021/acs.jafc.0c05023

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhuang, H., Lou, Q., Liu, H., Han, H., Wang, Q., Tang, Z., et al. (2019). Differential Regulation of Anthocyanins in Green and Purple Turnips Revealed by Combined De Novo Transcriptome and Metabolome Analysis. *Int. J. Mol. Sci.* 20:4387. doi: 10.3390/ijms20184387

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 March 2022
doi: 10.3389/fpls.2022.828270

Beta-Amylase and Phosphatidic Acid Involved in Recalcitrant Seed Germination of Chinese Chestnut

Yang Liu[1,2†], Yu Zhang[1†], Yi Zheng[3], Xinghua Nie[1], Yafeng Wang[1], Wenjie Yu[1], Shuchai Su[2], Qingqin Cao[1], Ling Qin[1] and Yu Xing[1]**

[1] Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, College of Plant Science and Technology, Beijing University of Agriculture, Beijing, China, [2] Key Laboratory for Silviculture and Conservation of Ministry of Education, College of Forestry, Beijing Forestry University, Beijing, China, [3] Bioinformatics Center, Beijing University of Agriculture, Beijing, China

Chinese chestnut (*Castanea mollissima*), a species with recalcitrant seeds, is an important source of nuts and forest ecosystem services. The germination rate of recalcitrant seeds is low in natural habitats and decreases under conditions of desiccation and low temperature. The germination rate of cultivated Chinese chestnut seeds is significantly higher than that of wild seeds. To explore the reasons for the higher germination rate of cultivated seeds in Chinese chestnut, 113,524 structural variants (SVs) between the wild and cultivated Chinese chestnut genomes were detected through genome comparison. Genotyping these SVs in 60 Chinese chestnut accessions identified allele frequency changes during Chinese chestnut domestication, and some SVs are overlapping genes for controlling seed germination. Transcriptome analysis revealed downregulation of the abscisic acid synthesis genes and upregulation of the beta-amylase synthesis genes in strongly selected genes of cultivated seeds. On the other hand, hormone and enzyme activity assays indicated a decrease in endogenous ABA level and an increase in beta-amylase activity in cultivated seeds. These results shed light on the higher germination rate of cultivated seeds. Moreover, phosphatidic acid synthesis genes are highly expressed in seed germination stages of wild Chinese chestnut and may play a role in recalcitrant seed germination. These findings provide new insight into the regulation of wild seed germination and promote natural regeneration and succession in forest ecosystems.

Keywords: SVs, seed germination, recalcitrant seeds, amylase, phosphatidic acid

OPEN ACCESS

Edited by:
Rong Wang,
East China Normal University, China

Reviewed by:
Xin Tong,
East China Normal University, China
Yan Chen,
Mianyang Normal University, China

***Correspondence:**
Ling Qin
qinlingbac@126.com
Yu Xing
xingyubua@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *03 December 2021*
Accepted: *02 February 2022*
Published: *25 March 2022*

Citation:
Liu Y, Zhang Y, Zheng Y, Nie X, Wang Y, Yu W, Su S, Cao Q, Qin L and Xing Y (2022) Beta-Amylase and Phosphatidic Acid Involved in Recalcitrant Seed Germination of Chinese Chestnut. Front. Plant Sci. 13:828270. doi: 10.3389/fpls.2022.828270

INTRODUCTION

Seeds act as an important vehicle by which angiosperms to disperse offspring, representing a key stage in the plant life cycle. Seeds are generally classified as "recalcitrant" and "orthodox" based on their tolerance of desiccation (Xia et al., 2014). Recalcitrant seeds die rapidly when stored under desiccation and low-temperature conditions (Walters et al., 2013). Numerous forest species, including *Castanea* (Vieitez et al., 2011), *Quercus* (Sghaier-Hammami et al., 2020), and *Aesculus* (Obroucheva et al., 2016), have recalcitrant seeds. Fagaceae species are the main trees with recalcitrant seeds used to maintain forest ecosystems and play important roles in the survival of

wild animals and forest ecological restoration (Walters et al., 2013; Wilf et al., 2019). However, most Fagaceae plant seeds have a low germination rate in the wild, and thus, increasing their germination rate is necessary for afforestation and ecological protection.

To improve the germination rate of recalcitrant seeds, their germination mechanism should be clarified. Several studies have focused on the sensitivity to dehydration and low temperature of recalcitrant seeds in terms of cell biology, physiology and molecular biology (Romero-Rodríguez et al., 2018; Li et al., 2021; Xia and Zhu, 2021), and the findings contribute to the conservation of recalcitrant seeds and the natural regeneration and ecological protection of forests. Seed germination is a complex adaptive trait of higher plants that is influenced by a large number of genes, endogenous hormones and environmental factors (Shu et al., 2016). In species with recalcitrant seeds, high levels of gibberellins (GAs) and low levels of ABA are observed in mature seeds, and the levels of auxin (IAA) and GAs increase during the seed germination process (Romero-Rodríguez et al., 2018). Moreover, exogenous ABA treatments of recalcitrant seeds results in a reduced germination rate (Wang et al., 2019). Genes related to hormone biosynthesis and signal transduction are significantly differentially expressed under desiccation (Dussert et al., 2018; Kijak and Ratajczak, 2020). The *ZEAXANTHIN EPOXIDASE* (*ZEP*), *NINE-CIS-EPOXYCAROTENOID DIOXYGENASE* (*NCED*), *PYRABACTIN RESISTANCE 1* (*PYR1*) genes related to ABA and *YUCCA, ADP-RIBOSYLATION FACTOR* (*ARF*), genes related to IAA have been found to regulate the seed desiccation tolerance acquisition in *Quercus variabilis* (Li et al., 2021). Additionally, paclobutrazol interferes with GA-related gene expression to induce desiccation tolerance in the seeds of *Citrus limon* (Marques et al., 2019). Hormones also regulate seed germination by controlling metabolism, and the amylase activity (Zaynab et al., 2021), lipid metabolism (Cui et al., 2020), glycerophospholipid metabolism (Chen et al., 2018) and reactive oxygen species (ROS) metabolism (Romero-Rodríguez et al., 2018) play key roles in recalcitrant seed germination. Compared to ungerminated seeds, germinated seeds show higher beta-amylase activity (Zaynab et al., 2021). Moreover, there is a close association between reduced seed viability and increasing phosphatidic acid (PA), and glycerophospholipid metabolism genes regulated desiccation sensitivity in recalcitrant seeds (Chen et al., 2018; Li et al., 2021). Nevertheless, a low germination rate of recalcitrant seeds in Fagaceae has not yet been addressed. Most Fagaceae plants grow in the wild, and the reproduction *via* seedlings in afforestation practices mainly relies on seeds. With the cultivation and utilization of forest trees, some cultivated species with high seed germination rates should represent an important source of seedlings for ecological restoration.

Fortunately, there are both wild and cultivated types of Chinese chestnut, which is in the *Castanea* genus. Compared with wild Chinese chestnut seeds, cultivated Chinese chestnut seeds have a higher levels of seed germination capacity under natural conditions. Chinese chestnut (*Castanea mollissima*) has a long history of commercial nut production and is widely cultivated in 26 provinces in China. As starchy nuts, the seeds (dry weight) contain 46∼64% starch, 12∼22% soluble sugars and 0.27∼0.64% lipids (Chen et al., 2017). As the germination characteristics of cultivated seeds are different from those of wild seeds of Chinese chestnut, we explored the mechanisms underlying the seed germination characteristics of Chinese chestnut to provide new insights into recalcitrant seeds.

In this study, we used omics to explore the molecular mechanism of the difference in the germination of recalcitrant seeds of wild and cultivated Chinese chestnut. Recently, wild (HBY_2) and cultivated (N11_1) Chinese chestnut genome assemblies were generated by using Pacific Biosciences single-molecule sequencing technology (Xing et al., 2019; Wang J. et al., 2020), and such Chinese chestnut genomes will provide a platform for comparative genomics analysis to characterize functional and structural features. To identify the differential gene loci related to seed germination of wild and cultivated Chinese chestnut, structural variants (SVs) between wild and cultivated plants were detected through genome comparison and genotyping, and these SV overlapping genes may control seed germination in Chinese chestnut. Comparative transcriptome analysis revealed the related genes involved in the differences in wild and cultivated seed germination. The results help lay a theoretical foundation for improving the recalcitrant seed germination and promoting natural regeneration and succession in forest ecosystems.

MATERIALS AND METHODS

Germination Assay, Endogenous Hormone and Transcriptome Sequencing of Chinese Chestnut Seeds

Mature Chinese chestnut seeds were harvested from a wild tree (HBY_2) and a tree of cultivar 'Jingshuhong' growing in an orchard in Beijing city, China. Undamaged Chinese chestnut seeds were selected for germination in darkness at 22°C. The seed germination assay included wild and cultivar groups. A total of 270 seeds were equally divided among three biological replicates for each group (**Figure 1A**). The wild and cultivated seeds were soaked in sterile water for a seed imbibition experiment until the seed weight did not increase. The relative water content was determined as the initial relative water content minus the relative water content that after desiccation. Then, the seeds were planted in a planter (50 cm × 30 cm × 40 cm) containing sterilized sand in a climate chamber with a 16 h light (25°C): 8 h dark (25°C) cycle and 60% relative humidity. The seeds were considered to have successfully germinated when the emerged radicle was at least 5 mm long and the germination rate reached 30% (Roach et al., 2010). According to the germination assay, seed embryos were sampled at 0 (S1) and 3 h of imbibition (S2) and when the seed germination rate reached 30% (S3: 96 h of imbibition for 'Jingshuhong' seeds, 216 h for wild seeds), respectively. Three biological replicates were tested for each sample. Endogenous hormone [GA_3, GA_4, ABA, IAA, zeatin riboside (ZR), isopentenyladenoside (IPA), brassinosteroids (BR) and methyl jasmonate (MeJA)] contents of these samples were

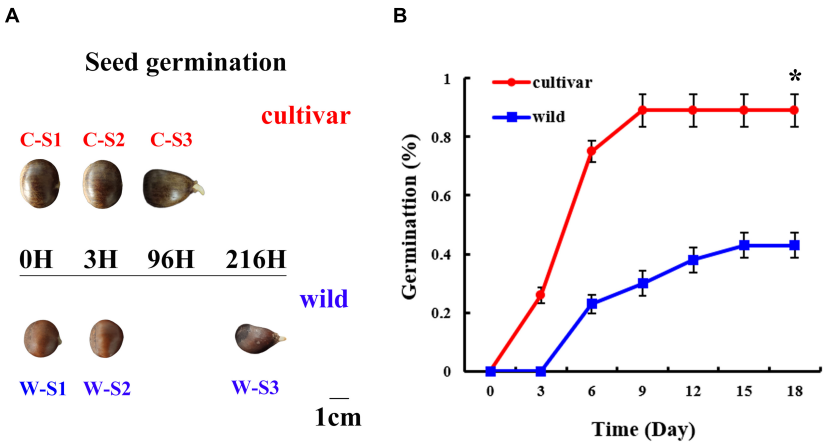


FIGURE 1 | Morphology and physiology of the seed germination process at different developmental stages in wild and cultivated Chinese chestnut. **(A)** Morphology of wild and cultivated Chinese chestnut seeds at different developmental stages of germination. S1: seed embryos after imbibition for 0 h; S2: seed embryos after imbibition for 3 h; S3: radicle emergence at 96 h for cultivated seeds and 216 h for wild seeds. **(B)** The seed germination rate of Chinese chestnut. * indicates a significant difference at $P < 0.05$.

determined by enzyme-linked immunosorbent assays (Deng et al., 2008), and each sample was measured in parallel three times. Alpha-amylase and beta-amylase activities were measured as described previously (Gimbi and Kitabatake, 2002). Then, the RNA of the samples was extracted from the seed embryos for transcriptome sequencing using a Plant RNA Kit (OMEGA, GA, United States), and libraries from the RNA samples were sequenced using the Illumina HiSeq 4000 platform (Illumina, CA, United States) at Novogene Bioinformatics Technology Co., Ltd. (Tianjing, China). Clean data were aligned to the Chinese chestnut V2 and N11_1 genome using STAR (version 2.7.1a). Counts were generated using HTSeq, and DESeq2 was used to identify differentially expressed genes (DEGs).

Genome Assembly and Chinese Chestnut Genome Annotation

Recalcitrant seed germination is a complex process regulated by various factors. To reveal the basis of the higher seed germination rate of cultivated than wild Chinese chestnut, we used comparative genome analysis to identify related genes. Wild and cultivated genomes of Chinese chestnut have been assembled, but the wild genome still lacks a high-quality chromosomal assembly and annotation version. Therefore, we assembled a chromosome-scale genome of wild Chinese chestnut. Fresh leaf tissue from Chinese chestnut (HBY_2) was used to construct Hi-C libraries, which were sequenced using the Illumina HiSeq PE150 platform (Illumina, CA, United States) at Novogene Bioinformatics Technology Co., Ltd. (Tianjing, China). We utilized Burrows–Wheeler Aligner (BWA) (version 0.7.17-r1188) (Li and Durbin, 2009) to compare the raw Hi-C reads with the draft assembled sequence (Xing et al., 2019); and low-quality reads were removed by SAMtools (version 1.9) (Li et al., 2009). Valid interaction pairs were then applied to build interaction matrices and scale up the major contigs to chromosome-scale scaffolds with LACHESIS (version 201701) (Burton et al., 2013).

For annotation analysis, different tissue types, including leaves, mixed floral buds, leaf buds, stems, female flowers, staminate catkins, nuts, burs, inner and outer shells, globular stage embryos and roots, were collected and immediately frozen in liquid nitrogen. Equal amounts of RNA from these tissues were pooled and used for single-molecule real-time (SMRT) sequencing. Full-length cDNA was synthesized using SMARTer PCR cDNA Synthesis Kit (Clontech Laboratories, CA, United States), and the library was subjected to SMRT sequencing using the PacBio Sequel platform. At the same time, we collected 44 Illumina-based RNA-seq data sets from various tissues, including male flowers, female flowers, leaves, mixed floral buds, male floral buds, nut developmental stages, embryo developmental stages, ovule developmental stages, stems and roots.

Homology alignment and *de novo* annotation were used to identify repetitive elements in the Chinese chestnut genome. For homology-based prediction, the Repbase TE library[1] was utilized to search against the Chinese chestnut genome using RepeatMasker (version 4.0) (Tarailo-Graovac and Chen, 2009) and RepeatProteinMask[2] with default parameters. For *de novo* prediction, a *de novo* repetitive element database was constructed by LTR_FINDER[3], RepeatScout[4] and RepeatModeler[5] with default parameters. A hybrid approach of *de novo* prediction, homology-based prediction, and transcriptome-based prediction approaches was applied to identify protein-coding regions and to predict genes. For homology-based prediction, Exonerate (version 2.47.3) (Slater and Birney, 2005) was employed for mapping against *C. mollissima*, with protein sequences from *Arabidopsis thaliana*, *Oryza sativa*, *Vitis vinifera*,

[1]http://www.girinst.org/repbase/
[2]http://www.repeatmasker.org/
[3]http://tlife.fudan.edu.cn/ltr_finder/
[4]http://www.repeatmasker.org/
[5]http://www.repeatmasker.org/RepeatModeler.html

Populus trichocarpa, *Malus domestica*, *Cucumis sativus*, *Juglans microcarpa* × *Juglans regia*, *Casuarina equisetifolia* and *Quercus robur*. Homologous proteins were then aligned to the Chinese chestnut genome by TBLASTN (Altschul et al., 1997), which were then aligned to the assembly by GeneWise (v2.4.1) (Birney et al., 2004) to predict gene structures. For *de novo* prediction, Augustus (version 3.0.2) (Stanke et al., 2006), GlimmerHMM (version 3.0.2) (Majoros et al., 2004), SNAP (version 11-29-2013) (Korf, 2004), GeneID (version 1.4) (Parra et al., 2000), and GENSCAN (version 1.0) (Salamov and Solovyev, 2000) were used to predict coding regions in the repeat-masked genome. For transcriptome-based prediction, Cufflinks (version 2.1.1) (Trapnell et al., 2012) was used to link transcripts from the TopHat results to gene models, and then PASA (version 2.3.3) (Campbell et al., 2006) was applied on the basis of the assembled RNA-seq unigenes. Finally, an integrated gene set from the above three methods was produced by EVidenceModeler (EVM version 1.1.1) (Haas et al., 2008).

Gene Functional Annotation

In a previous study, a standard pipeline was established for thoroughly annotating predicted protein-coding genes (Zheng et al., 2019). In brief, the predicted gene sequences were compared to the *Arabidopsis* protein (TAIR), NCBI non-redundant (nr), and UniProt (Swiss-Prot and TrEMBL) databases using the BLASTP command of DIAMOND (Buchfink et al., 2015), with an *E*-value cutoff of $1e^{-5}$. Furthermore, InterProScan was used to identify functional domains in all protein sequences by comparing them to the InterPro database (Mitchell et al., 2019). To assign GO terms for each protein-coding gene, the BLASTP results from the nr database and the discovered InterPro domains were input into the Blast2GO pipeline (Conesa and Götz, 2008). The BLASTP leads obtained from searches against the TAIR and UniProt databases were input into the AHRD program[6] to acquire accurate, succinct and useful gene functional summaries. The KEGG pathways encoded by each of the *Castanea* genomes were also predicted by KEGG Automatic Annotation Server[7]. The iTAK program was employed to identify TFs and TRs from the predicted protein-coding genes and to classify them into different families (Zheng et al., 2016).

Comparative Genomic Analysis and Detection of Structural Variants Between the V2 and N11_1 Genome

Changes in SV allele frequency across distinct Chinese chestnut populations are a consequence of domestication processes such as favorable trait selection and introgression from ancestral groups. We analyzed SV allele frequency changes from wild to cultivated materials for domestication to discover SVs under selection during Chinese chestnut domestication and breeding. The relation pipelines for SV detection and genotyping in Chinese chestnut were based on the study of Wang X. et al. (2020). SVs of Chinese chestnut genomes V2 and N11_1 were identified by Minimap with the parameter "-ax asm5" (Li, 2018). The resulting alignments were analyzed using Assemblytics (v1.1) (Nattestad and Schatz, 2016) for SV identification. SV calling was based on genome comparison, and then filtering by Illumina read mapping.

Structural Variants and SNP Genotyping in the Chinese Chestnut Population

A total of 60 accessions were collected, including 28 wild and 32 cultivated samples of Chinese chestnut in China (**Supplementary Data 1**). Young leaves were immediately frozen in liquid nitrogen and stored at −80°C; genomic DNA was extracted from young leaves using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) and used to construct sequencing libraries. Paired-end sequencing libraries with an insert size of approximately 450 bp and a read length of 150 bp were sequenced using an Illumina HiSeq 2500 sequencer. The raw reads were filtered to obtain high-quality sequences by Trimmomatic (v0.39) (Bolger et al., 2014). The high-quality reference SVs identified between the V2 and N11_1 genome were genotyped in the 60 Chinese chestnut accessions based on the protocols of Gao et al. (2019) and Wang X. et al. (2020). Clean Illumina reads for each accession were aligned to the V2 and N11_1 genome by BWA-MEM (v0.7.17) (Li and Durbin, 2009), with no more than 3% mismatches allowed. SVs in each accession were classified as V2 (same as V2), N11_1 (same as N11_1), and heterozygous (carrying both V2 and N11_1 alleles) genotypes. The allele frequency of certain SVs was computed for each group, and the significance of the frequency differences between two compared groups was established using Fisher's exact test. Raw *P* values of SVs were adjusted using the false discovery rate (FDR) approach, and SVs with corrected *P* values < 0.05 were considered as those under selection.

In addition, we employed Illumina reads matched to the V2 genome for SNP calling. Variant calling was performed independently with Sentieon (Freed et al., 2017) with default parameters. GATK (v4.1.1.0) was used for hard filtering with the settings parameters "QD $<$ 2.0 || FS $>$ 60.0 || MQ $<$ 40.0 || MQRankSum $<$ −12.5 || ReadPosRankSum $<$ −8.0" (McKenna et al., 2010). SNPs with at least 70% genotyped accessions, with a minor allele frequency (MAF) of no less than 0.03 and overlapping with known SNPs from the V2 and N11_1 genome alignment were preserved.

Population Genetic Analysis

IQ-TREE (v2.0.3) with maximum likelihood was used to build phylogenetic trees for the Chinese chestnut accessions using the full SNPs at fourfold degenerated sites (4DTV) in 1,000 bootstrap runs with *Castanea henryi* as the outgroup (Sun et al., 2020). Principal component analysis (PCA) was performed using the program Plink (v1.90) (Purcell et al., 2007). VCFtools (Danecek et al., 2011) was used to calculate F_{ST} and π across the genome with a 15 kb window based on the SNPs. XP-CLR analyses were implemented to detect selective sweeps based on SNPs across chestnut populations (Chen et al., 2010). Correlation coefficients (r^2) were calculated for all pairs of SNPs to measure linkage

[6]https://github.com/groupschoof/AHRD

[7]https://www.genome.jp/tools/kaas/

disequilibrium (LD) decay using PopLDdecay (v3.41) (Zhang et al., 2019) with default parameters.

RESULTS

Physiological Features of Seed Germination in Wild and Cultivated Chinese Chestnut

As based on observation of the seed germination process, 0–12 h (0–12H) represented a rapid initial penetration phase; and 12–48 h (12–48H) represented a plateau phase in terms of water content (**Supplementary Figure 1**). There was a significant difference in the water content of seeds, with averages water contents of 46 and 52% for wild and cultivated Chinese chestnut seeds before the water uptake phase, respectively (**Supplementary Figure 1**). Radicles started to emerge after imbibition for 96 h (4 days) and 216 h (9 days) in wild and cultivated Chinese chestnut seeds, respectively (**Figure 1A**). There were significant changes in the seed germination rate, with total germination rates of 43 and 89% for wild and cultivated Chinese chestnut seeds, respectively (**Figure 1B**). The cultivated and wild seed germination rates reached a maximum at 9 and 15 days, respectively.

Genomic Structural Variants Between Wild and Cultivated Genomes

We assembled a chromosome-scale genome of wild Chinese chestnut, and a total of 2,652,199 read pairs (90.55 Gb clean reads) were acquired by Hi-C sequencing, offering over 115-fold coverage (**Supplementary Tables 1, 2**). Then, to improve gene prediction accuracy, high-quality transcripts of 11 mixed

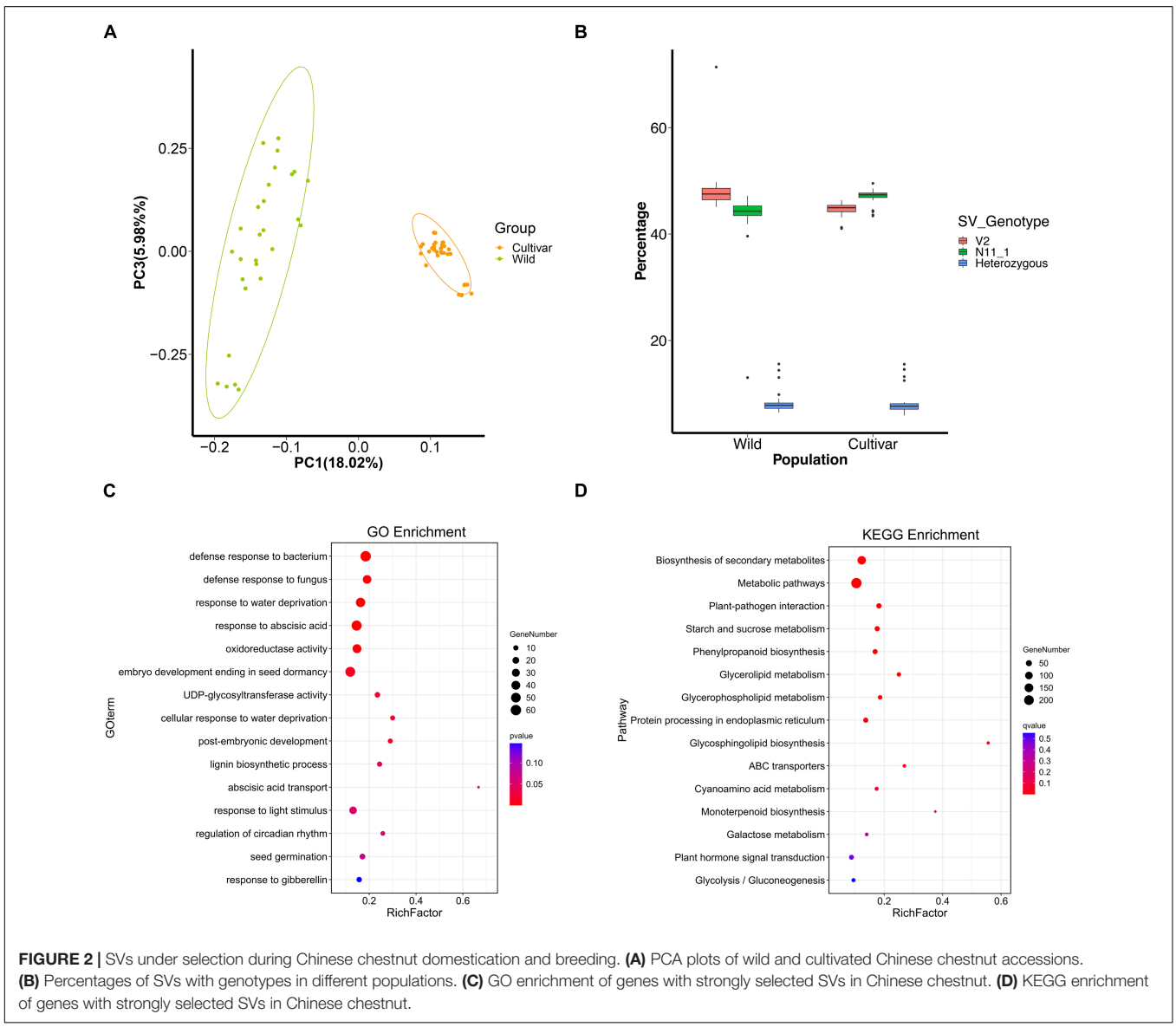


FIGURE 2 | SVs under selection during Chinese chestnut domestication and breeding. **(A)** PCA plots of wild and cultivated Chinese chestnut accessions. **(B)** Percentages of SVs with genotypes in different populations. **(C)** GO enrichment of genes with strongly selected SVs in Chinese chestnut. **(D)** KEGG enrichment of genes with strongly selected SVs in Chinese chestnut.

tissue types were sequenced using a single-molecule real-time sequencer from Pacific Biosciences (**Supplementary Table 3**). Finally, a total of 33,991 protein-coding genes were aligned to the wild Chinese chestnut genome (V2) (**Supplementary Figure 2** and **Supplementary Tables 4, 5**). The Chinese chestnut V2 genome assembly can be accessed at *Castanea* Genome Database (CGD[8]). Indels were discovered through direct comparison of the genomes of wild (V2) and cultivated (N11_1) Chinese chestnut with mapping of PacBio long reads. A total of 113,524 indels were detected, ranging in size from 10 to 20,094 bp, including intergenic, intronic, upstream, downstream, and exonic areas (**Supplementary Data 2** and **Supplementary Table 6**). Most of the indels are short, with 87.7% being less than 100 bp and only 0.6% exceeding 10 kb (**Supplementary Figure 3**). To further identify the SVs under strong environmental and artificial selection, 113,524 indels were genotyped in the Chinese chestnut population to investigate allele frequency changes, including 28 wild and 32 cultivated accessions (**Supplementary Data 1**). The short read sequencing data yielded 796 Gb of high-quality clean data, and the sequencing depths ranged from 10× to 42× in the wild and cultivated populations of Chinese chestnut. Genotyping the SVs in two reference genomes using Illumina short reads further supported their high specificity (**Supplementary Table 6**). Phylogenetic analysis and PCA clearly separated the wild and cultivated groups of Chinese chestnut (**Figure 2A** and **Supplementary Figure 4**). LD values were also calculated among wild and cultivated populations, and higher LD was found in the cultivated population (**Supplementary Figure 5**).

[8]www.castaneadb.net

Selection of Structural Variants for Seed Germination

Structural variant loci with the V2 alleles were common in the wild population, accounting for 48.46% of the SVs in each accession, whereas 43.48% of SV loci showed the homozygous N11_1 genotype (**Figure 2B**). The allele frequencies of 4,892 SVs were significantly differentially changed between the wild and cultivated populations of Chinese chestnut (**Supplementary Data 4**), and these indels might affect 4,673 genes in the V2 genome. These genes are significantly involved in biosynthesis of secondary metabolites, plant-pathogen interaction, starch and sucrose metabolism, and glycerophospholipid metabolism according to KEGG enrichment function (**Figure 2D**), and response to abscisic acid, response to water deprivation, and seed germination according to GO

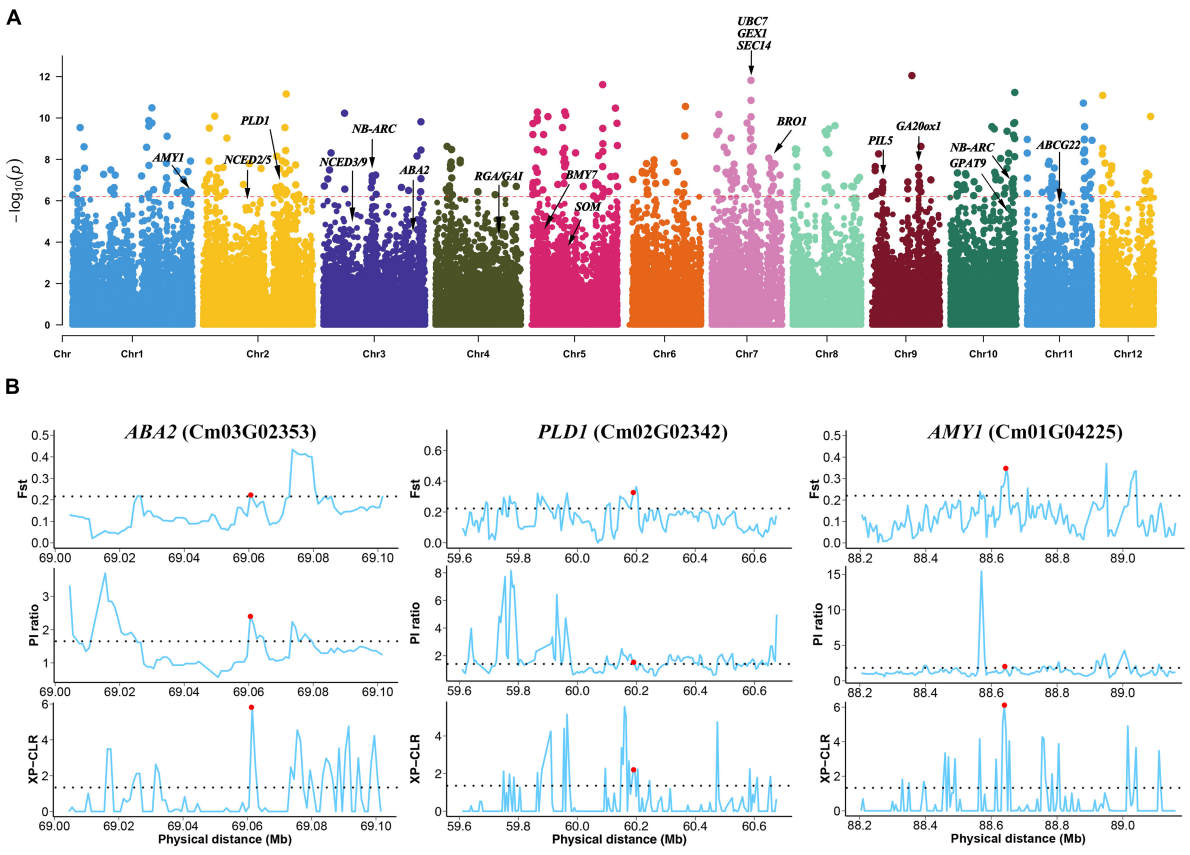


FIGURE 3 | Strong selection of genes between wild and cultivated Chinese chestnut. **(A)** Genome-wide distribution of selective sweeps in Chinese chestnut. **(B)** F_{ST}, π, and XP-CLR values across the genomic regions of the *ABA2*, *PLD1* and *AMY1* genes. The dashed horizontal line represents the selection threshold (top 5% of the genome). Red dots denote the genes that are connected.

enrichment function (**Figure 2C** and **Supplementary Data 3, 5**). These genes were also selected at the sequence level using F_{ST}, nucleotide diversity (π), and the cross-population composite likelihood ratio (XP-CLR) (**Figure 3**).

Interestingly, some genes involved in plant hormone signal transduction and starch and sucrose metabolism were detected, and these genes may influence the differences in seed germination between wild and cultivated Chinese chestnut. The regions included a 12 bp insertion in the exon of *PHYTOCHROME INTERACTING FACTOR 3-LIKE 5* (*PIL5*, Cm09G00427), and an 18 bp deletion in the intergenic region of *SOMNUS* (*SOM*, Cm05G01513), *ABA DEFICIENT 2* (*ABA2*, Cm03G02353) for ABA biosynthesis and *GIBBERELLIN 20 OXIDASE 1* (*GA20ox1*, Cm09G01563) for GA biosynthesis in plant hormone signal transduction and *ALPHA-AMYLASE-LIKE 1* (*AMY1*, Cm01G04225) and *BETA-AMYLASE 7* (*BMY7*, Cm05G00315) for starch degradation *via* alpha-amylase and beta-amylase activities (**Figure 3B**). Additionally, some genes involved in glycerophospholipid metabolism were detected in wild and cultivated Chinese chestnut, including a 17 bp deletion in phospholipase D (*PLD1*, Cm02G02342) and a 39 bp deletion in *GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE 9* (*GPAT9*, Cm10G02185), which are involved in activating the synthesis of phosphatidic acid. The *ARABIDOPSIS THALIANA ATP-BINDING CASSETTE G22* (*ABCG22*, Cm11G01020) gene, an ABC transporter gene, is involved in drought susceptibility in seed germination. Moreover, the 1.61 Mb genomic region with a high selection score contains four genes on chromosome 7 (**Figure 3A**), including *UBIQUITIN CARRIER PROTEIN 7* (*UBC7*, Cm07G01553) and *SECRETION 14* (*SEC14*, Cm07G01557), which are involved in plant responses to multiple stress conditions (Ascencio-Ibáñez et al., 2008; Feng et al., 2020). *GAMETE EXPRESSED PROTEIN 1* (*GEX1*, Cm07G01555) contributes to gametophyte development (Alandete-Saez et al., 2011), suggesting strong selection of this region during domestication. In addition, there are several regions of genes involved in plant pathogen interactions, including a 33 bp deletion in the promoter of *NB-ARC* (Cm10G02154) and a 17 bp deletion in *NB-ARC* (Cm03G01169), were detected. These genes may contribute to differences in disease resistance and responses to stress conditions between wild and cultivated Chinese chestnut. The results demonstrated a common

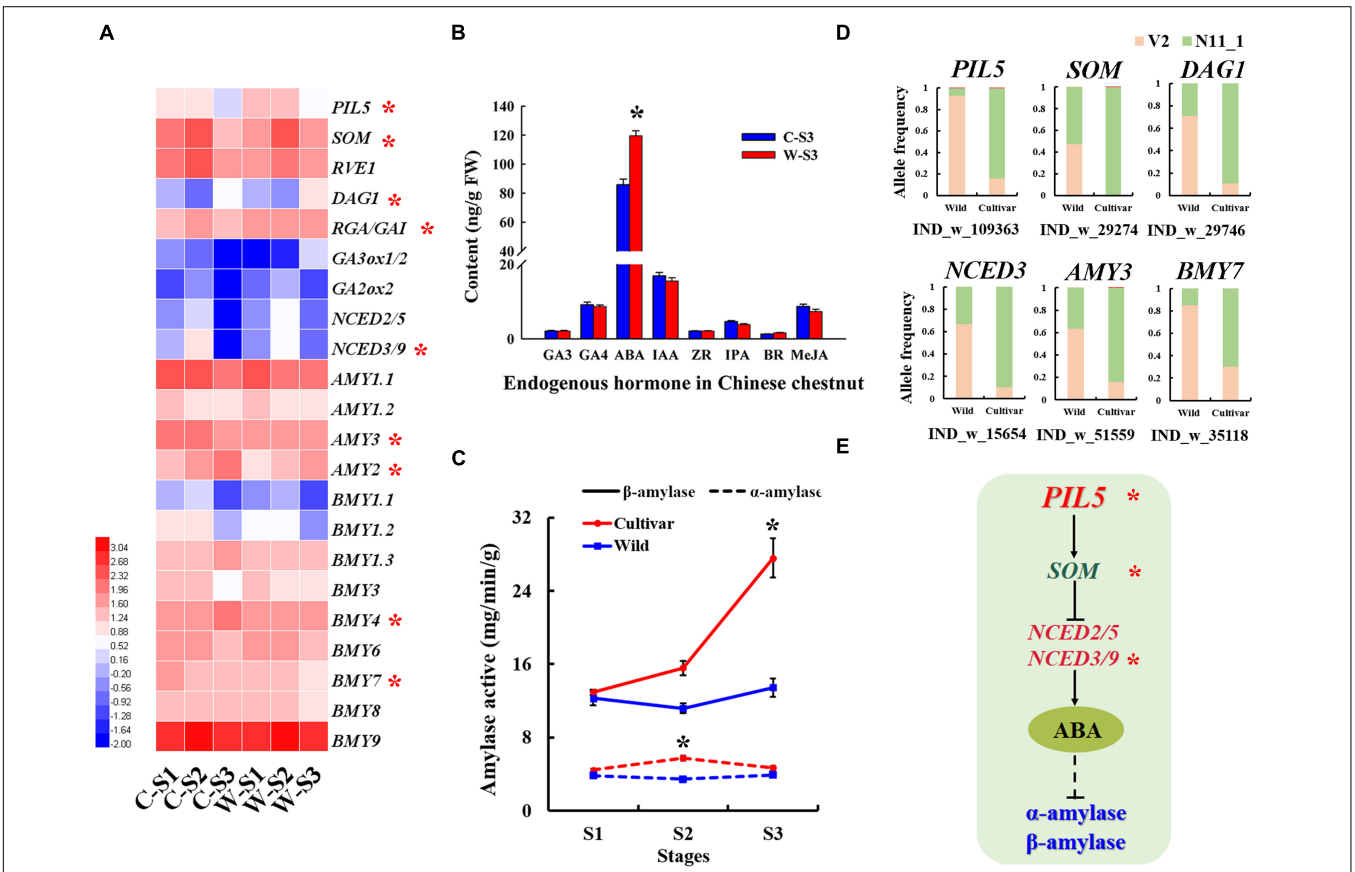


FIGURE 4 | Selected amylase genes associated with the germination of Chinese chestnut seeds. **(A)** Seed germination-related gene expression profiles of Chinese chestnut, where * indicates a significant difference at $P < 0.05$. **(B)** Levels of endogenous hormones in Chinese chestnut at seed germination stage S3, where * indicates a significant difference at $P < 0.05$. **(C)** Alpha-amylase and beta-amylase activities in seed germination stages of wild and cultivated Chinese chestnut seeds; * indicates a significant difference at $P < 0.05$. **(D)** The allele frequencies of selected SVs in wild and cultivated Chinese chestnut. **(E)** The model of seed germination regulation by ABA synthesis pathway genes under strong selection in Chinese chestnut; * indicates the significance of the key genes.

selection preference for the N11_1 allele in Chinese chestnut domestication and improvement.

Selected Structural Variants Affecting Expression of Seed Germination Genes in Wild and Cultivated Chinese Chestnut

To further verify the strongly artificially selected genes and identify DEGs associated with seed germination in wild and cultivated Chinese chestnut, stage S1–S3 samples were used for transcriptomic sequencing analysis based on seed germination phases. There were 5,562, 6,741 and 9,105 DEGs in stages S1, S2 and S3 of cultivated Chinese chestnut seed germination, respectively. According to KEGG pathway analysis, these DEGs are significantly associated with the biosynthesis of secondary metabolites, metabolic pathways, starch and sucrose metabolism, and plant hormone signal transduction (**Supplementary Figure 6**). These genes may be related to recalcitrant seed germination in Chinese chestnut.

In general, seed germination was influenced by alpha-amylase and beta-amylase activity. Notably, the *AMY* and *BMY* genes exhibited allele frequency change patterns indicative of strong selection in the wild and cultivated populations of Chinese chestnut. For example, a 599 bp deletion in the *AMY3* gene promoter had a frequency of 63.64% in the wild population and 16.13% in the cultivated population in the V2 genome, and the 11 bp deletion in the *BMY7* gene had a frequency of 85.19% in the wild population and 30.00% in the cultivated population in the V2 genome (**Figure 4D**). Furthermore, the *AMY1*, *AMY2,* and *AMY3* genes were significantly more highly expressed during the S1–S2 stages of cultivated seed germination (**Figure 4A**) and may be involved in the differential of the alpha-amylase activity in the seed imbibition stage of wild and cultivated Chinese chestnut (**Figure 4C**). In the radicle emergence stage (S3), the beta-amylase activity of cultivated seeds was significantly higher than that of wild seeds, and there was no significant difference in alpha-amylase activity between cultivar and wild seeds (**Figure 4C**). The *BMY4* and *BMY7* genes were strongly selected in the wild and cultivar genomes and significantly highly expressed in the radicle emergence stage (S3) of the cultivated seeds; it may be involved in the differential of beta-amylase activity between of wild and cultivated seeds in the S3 stage (**Figures 4A,C**). Additionally, according to the SV divergence across the cultivated and wild genomes, GA and ABA signal transduction genes may be involved in seed germination differences. The *NCED2/5* and *NCED3/9* genes were strongly selected and showed specific allele frequency change patterns (**Figure 4D**). We also found that several genes of the ABA metabolic pathway to be significantly upregulated expression in stage S3 of wild Chinese chestnut, including *ABA4*, *ABA2*, *NCED2/5*, *NCED3/9* and *ABSCISIC ALDEHYDE OXIDASE 3* (*AAO3*) (**Supplementary Figure 8**). In particular, *NCED2/5* and *NCED3/9* were activated in wild seed germination stage S3. The *PIL5* gene was also detected in the SV divergence analysis of the genome and may be involved in seed germination. In terms of gene expression, the *PIL5* gene was significantly more highly expressed in wild seed germination stages, which directly activates transcription of the *SOM*, GIBBERELLIC ACID INSENSITIVE/*REPRESSOR OF GA* (*GAI/RGA*) and *DOF AFFECTING GERMINATION 1* (*DAG1*) genes to regulate the expression of genes related to GA and ABA biosynthesis, and was significantly more highly expressed in the wild seed germination stages (**Supplementary Figure 9**). Moreover, expression of the GA biosynthetic genes *GA3ox1/GA3ox2* was suppressed in stages S1–S2 and activated in stage S3 of wild seed germination, and *GA2ox2* was activated in wild seed germination stages S1–S3 (**Figure 4A** and **Supplementary Figure 7**). In addition, the content of the endogenous hormone ABA was significantly

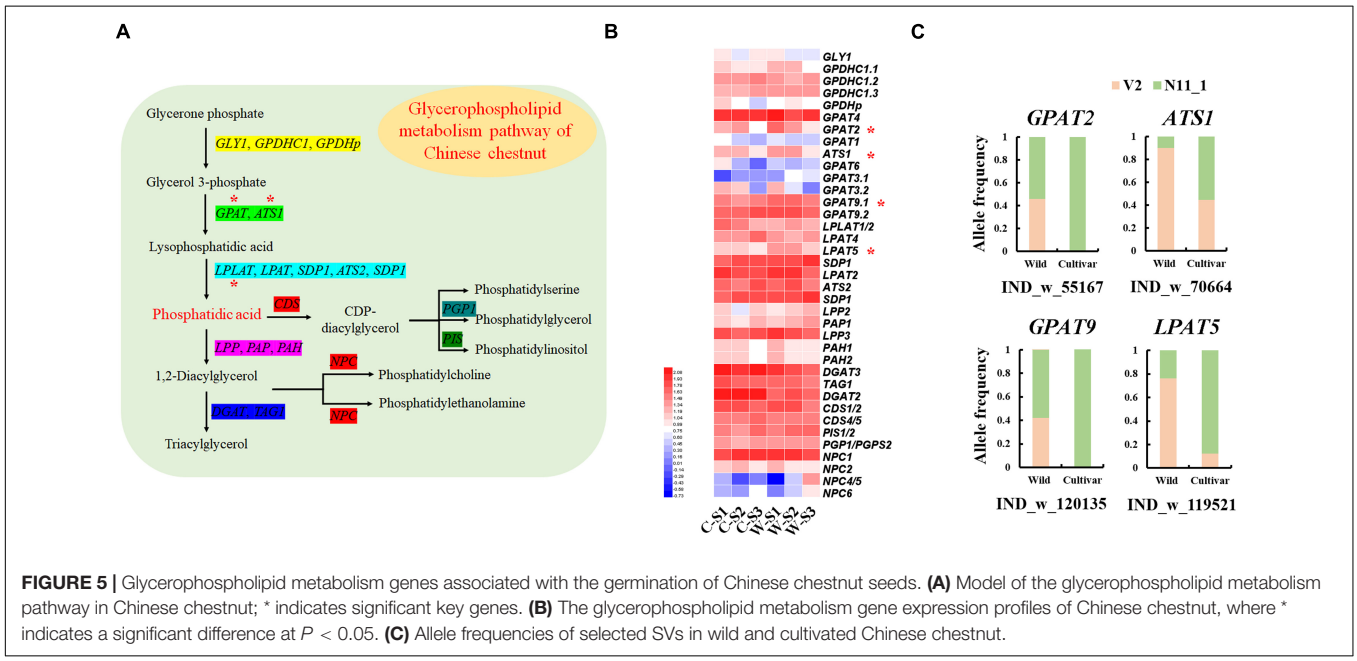


FIGURE 5 | Glycerophospholipid metabolism genes associated with the germination of Chinese chestnut seeds. **(A)** Model of the glycerophospholipid metabolism pathway in Chinese chestnut; * indicates significant key genes. **(B)** The glycerophospholipid metabolism gene expression profiles of Chinese chestnut, where * indicates a significant difference at $P < 0.05$. **(C)** Allele frequencies of selected SVs in wild and cultivated Chinese chestnut.

higher in wild seeds than in cultivated seeds (**Figure 4B**), consistent with agreement with the role of *NCED2/5* and *NCED3/9* in the activation of this direct regulation of ABA levels. There were no significant changes in the contents of other endogenous hormone in the cultivated and wild seed germination stages (**Figure 4B**).

Although the lipid content of Chinese chestnut seeds was less than 1%, we found that glycerophospholipid metabolism genes were strongly selected in the wild and cultivated populations (**Figure 2D**). The *GLYCEROL-3-PHOSPHATE SN-2-ACYLTRANSFERASE 2* (*GPAT2*), *GPAT9*, *ACYLTRANSFERASE 1* (*ATS1*) and *LYSOPHOSPHATIDYL ACYLTRANSFERASE 5* (*LPAT5*) genes showed specific allele frequency change patterns. For example, a 45 bp deletion in the *LPAT9* gene had a frequency of 76.19% in the wild group and 12.50% in the cultivated group (**Figure 5C**). These genes, which are involved in the synthesis of phosphatidic acid (**Figure 5A**), were significantly highly expressed during wild seed germination (**Figure 5B**). Additionally, the other genes involved in phosphatidic acid synthesis showed high expression during wild seed germination, although they were not strongly selected in the genomes.

DISCUSSION

Seed germination is crucial for plant development and breeding programs. This process is affected by environmental conditions and genetic structure, and plants require large amounts of energy, such as starch, proteins, and lipids during germination (Zhao et al., 2018; Zaynab et al., 2021). The starchy nuts of Chinese chestnut have a dry weight of 46∼64% (Chen et al., 2017), and the germination characteristics of these recalcitrant seeds differ between cultivated and wild trees. In general, seeds are unable to survive drying and chilling, as they rapidly lose their germination capacity and viability during storage (Pammenter and Berjak, 1999), and wild seeds have a strong dormancy capacity and low germination rates (Nakamura et al., 2017; Wang et al., 2018). Various factors are responsible for the low rate of seed germination, among which amylase has been recognized as an important factor in regulating seed germination (Damaris et al., 2019). In the present study, cultivated seeds displayed a stronger germination capacity than wild seeds. Some strongly selected alpha-amylase and beta-amylase genes were identified by comparing the genomes V2 and N11_1. *BMY4* and *BMY7* were significantly more highly expressed during the germination stages of cultivated seeds, and the activity of beta-amylase was significantly higher than that of alpha-amylase during the seed germination stage. These findings implied that beta-amylase may be play a key role in the high rate of seed germination processes of cultivated Chinese chestnut. Similarly, beta-amylase activity has been also verified to be an important factor in recalcitrant seed germination (Zaynab et al., 2021). Moreover, the amylase gene encodes an enzyme involved in starch degradation and is regulated by ABA and GA (Vanstraelen and Benková, 2012). ABA biosynthetic genes were strongly selected, with suppressed expression to decrease ABA levels in cultivated Chinese chestnut seeds. *PIL5*, a key negative regulator of seed germination, activates the expression of *SOM* by binding directly to its promoter, partially regulating expression of ABA and GA metabolic genes (Oh et al., 2004; Kim et al., 2008). Among the regulatory genes associated with Chinese chestnut seed germination (**Figure 4E**), high expression of the *PIL5* gene is also involved in the differential germination between wild and cultivated seeds.

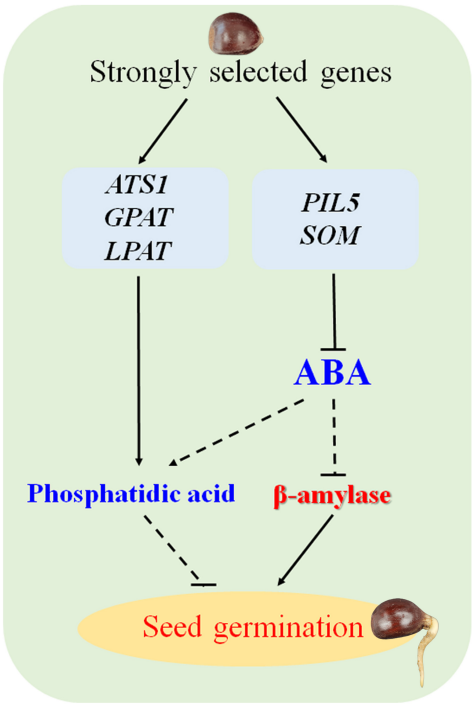


FIGURE 6 | Model of the seed germination mechanism regulated by strongly selected genes in Chinese chestnut.

Lipid degradation and conversion to sugars are the key factors for seed germination in oil palm (Cui et al., 2020). Membrane lipid analysis in relation to recalcitrant seed desiccation tolerance has a close relationship with a reduced seed germination rate (Chen et al., 2018), and phosphatidic acid synthesis genes are involved in the dehydration sensitivity of recalcitrant seed germination in cork oak (Li et al., 2021). Moreover, phosphatidic acid results in a response to ABA during seed germination (Katagiri et al., 2005). Regardless, it remains unclear whether the lipids are involved in the germination of starchy seed species. Although the lipid content of Chinese chestnut is only 0.27∼0.64% (Chen et al., 2017), there are some strongly selected genes involved in phospholipid metabolism and phosphatidic acid synthesis in cultivated and wild genomes, and these genes were upregulated in wild seed germination stages. This indicates that phospholipids may be involved in the germination of Chinese chestnut seeds.

The germination rate of the highly recalcitrant seeds of cultivated the Chinese chestnut was influenced by several factors, including natural and artificial selection. Finally, a possible working model for proposed for the regulatory mechanism of germination in recalcitrant seeds of Chinese chestnut is proposed

(**Figure 6**). These findings will contribute to improving the germination rate of recalcitrant seeds and provide insight into recalcitrant seed germination.

DATA AVAILABILITY STATEMENT

The sequencing datasets presented in this study can be found in online repositories. The Chinese chestnut genome assembly can be accessed at Castanea Genome Database (www.castaneadb.net). The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

YX and LQ designed the project. YL and YX wrote and revised the manuscript. YuZ, YiZ, and YL analyzed the data and constructed the database. YL, YuZ, YW, WY, and XN collected the samples. SS and QC participated in the manuscript. All authors revised and approved the manuscript.

FUNDING

This work was supported by grants from the National Key Research & Development Program of China (2018YFD1000605) and the National Natural Science Foundation of China (31870671).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.828270/full#supplementary-material

Supplementary Data 4 | GO and KEGG enrichment results of were strongly selected genes.

Supplementary Data 5 | The allele frequencies of SVs in Chinese chestnut.

REFERENCES

Alandete-Saez, M., Ron, M., Leiboff, S., and McCormick, S. (2011). *Arabidopsis thaliana GEX1* has dual functions in gametophyte development and early embryogenesis. *Plant J.* 68, 620–632. doi: 10.1111/j.1365-313X.2011.04713.x

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Ascencio-Ibáñez, J. T., Sozzani, R., Lee, T. J., Chu, T. M., Wolfinger, R. D., Cella, R., et al. (2008). Global analysis of *Arabidopsis* gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiol.* 148, 436–454. doi: 10.1104/pp.108.121038

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M., and Buell, C. R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* 7:327. doi: 10.1186/1471-2164-7-327

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Chen, H., Yu, X., Zhang, X., Yang, L., Huang, X., Zhang, J., et al. (2018). Phospholipase Dα1-mediated phosphatidic acid change is a key determinant of desiccation-induced viability loss in seeds. *Plant Cell Environ.* 41, 50–63. doi: 10.1111/pce.12925

Chen, L., Lu, D., Wang, T., Li, Z., Zhao, Y., Jiang, Y., et al. (2017). Identification and expression analysis of starch branching enzymes involved in starch synthesis during the development of chestnut (*Castanea mollissima* Blume) cotyledons. *PLoS One* 12:e0177792. doi: 10.1371/journal.pone.0177792

Conesa, A., and Götz, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008:619832. doi: 10.1155/2008/619832

Cui, J., Lamade, E., and Tcherkez, G. (2020). Seed germination in oil palm (*Elaeis guineensis* Jacq.): a review of metabolic pathways and control mechanisms. *Int. J. Mol. Sci.* 21:4227. doi: 10.3390/ijms21124227

Damaris, R. N., Lin, Z., Yang, P., and He, D. (2019). The rice Alpha-amylase, conserved regulator of seed maturation and germination. *Int. J. Mol. Sci.* 20:450. doi: 10.3390/ijms20020450

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Deng, A., Tan, W., He, S., Liu, W., Nan, T., Li, Z., et al. (2008). Monoclonal antibody-based enzyme linked immunosorbent assay for the analysis of jasmonates in plants. *J. Integr. Plant Biol.* 50, 1046–1052. doi: 10.1111/j.1744-7909.2008.00715.x

Dussert, S., Serret, J., Bastos-Siqueira, A., Morcillo, F., Déchamp, E., Rofidal, V., et al. (2018). Integrative analysis of the late maturation programme and desiccation tolerance mechanisms in intermediate coffee seeds. *J. Exp. Bot.* 69, 1583–1597. doi: 10.1093/jxb/erx492

Feng, H., Wang, S., Dong, D., Zhou, R., and Wang, H. (2020). *Arabidopsis* Ubiquitin-Conjugating Enzymes *UBC7*, *UBC13*, and *UBC14* are required in plant responses to multiple stress conditions. *Plants (Basel)* 9:723. doi: 10.3390/plants9060723

Freed, D., Aldana, R., Weber, J., and Edwards, J. (2017). The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv [preprint]* doi: 10.1101/115717

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51, 1044–1051. doi: 10.1038/s41588-019-0410-2

Gimbi, D. M., and Kitabatake, N. (2002). Changes in alpha-and beta-amylase activities during seed germination of African finger millet. *Int. J. Food Sci. Nutr.* 53, 481–488. doi: 10.1080/09637480220164361

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Katagiri, T., Ishiyama, K., Kato, T., Tabata, S., Kobayashi, M., and Shinozaki, K. (2005). An important role of phosphatidic acid in ABA signaling during germination in *Arabidopsis thaliana*. *Plant J.* 43, 107–117. doi: 10.1111/j.1365-313X.2005.02431.x

Kijak, H., and Ratajczak, E. (2020). What do we know about the genetic basis of seed desiccation tolerance and longevity? *Int. J. Mol. Sci.* 21:3612. doi: 10.3390/ijms21103612

Kim, D. H., Yamaguchi, S., Lim, S., Oh, E., Park, J., Hanada, A., et al. (2008). SOMNUS, a CCCH-type zinc finger protein in *Arabidopsis*, negatively regulates light-dependent seed germination downstream of *PIL5*. *Plant Cell.* 20, 1260–1277. doi: 10.1105/tpc.108.058859

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinform.* 5:59. doi: 10.1186/1471-2105-5-59

Li, D., Li, Y., Qian, J., Liu, X., Xu, H., Zhang, G., et al. (2021). Comparative transcriptome analysis revealed candidate genes potentially related to desiccation sensitivity of recalcitrant *Quercus variabilis* seeds. *Front. Plant Sci.* 12:717563. doi: 10.3389/fpls.2021.717563

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Marques, A., Nijveen, H., Somi, C., Ligterink, W., and Hilhorst, H. (2019). Induction of desiccation tolerance in desiccation sensitive *Citrus limon* seeds. *J. Integr. Plant Biol.* 61, 624–638. doi: 10.1111/jipb.12788

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi: 10.1093/nar/gky1100

Nakamura, S., Pourkheirandish, M., Morishige, H., Sameri, M., Sato, K., and Komatsuda, T. (2017). Quantitative trait loci and maternal effects affecting the strong grain dormancy of wild barley (*Hordeum vulgare* ssp. spontaneum). *Front. Plant Sci.* 8:1840. doi: 10.3389/fpls.2017.01840

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023. doi: 10.1093/bioinformatics/btw369

Obroucheva, N., Sinkevich, I., and Lityagina, S. (2016). Physiological aspects of seed recalcitrance: a case study on the tree *Aesculus hippocastanum*. *Tree Physiol.* 36, 1127–1150. doi: 10.1093/treephys/tpw037

Oh, E., Kim, J., Park, E., Kim, J. I., Kang, C., and Choi, G. (2004). PIL5, a phytochrome-interacting basic helix-loop-helix protein, is a key negative regulator of seed germination in *Arabidopsis thaliana*. *Plant Cell.* 16, 3045–3058. doi: 10.1105/tpc.104.025163

Pammenter, N., and Berjak, P. (1999). A review of recalcitrant seed physiology in relation to desiccation-tolerance mechanisms. *Seed Sci. Res.* 9, 13–37. doi: 10.1017/S0960258599000033

Parra, G., Blanco, E., and Guigó, R. (2000). GeneID in Drosophila. *Genome Res.* 10, 511–515. doi: 10.1101/gr.10.4.511

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Roach, T., Beckett, R. P., Minibayeva, F. V., Colville, L., Whitaker, C., Chen, H., et al. (2010). Extracellular superoxide production, viability and redox poise in response to desiccation in recalcitrant *Castanea sativa* seeds. *Plant Cell Environ.* 33, 59–75. doi: 10.1111/j.1365-3040.2009.02053.x

Romero-Rodríguez, M. C., Archidona-Yuste, A., Abril, N., Gil-Serrano, A. M., Meijón, M., and Jorrín-Novo, J. V. (2018). Germination and early seedling development in *Quercus ilex* recalcitrant and non-dormant seeds: targeted transcriptional, hormonal, and sugar analysis. *Front. Plant Sci.* 9:1508. doi: 10.3389/fpls.2018.01508

Salamov, A. A., and Solovyev, V. V. (2000). *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res.* 10, 516–522. doi: 10.1101/gr.10.4.516

Sghaier-Hammami, B., Hammami, S. B. M., Baazaoui, N., Gómez-Díaz, C., and Jorrín-Novo, J. V. (2020). Dissecting the seed maturation and germination processes in the non-orthodox *Quercus* ilex species based on protein signatures as revealed by 2-DE coupled to MALDI-TOF/TOF proteomics strategy. *Int. J. Mol. Sci.* 21:4870. doi: 10.3390/ijms21144870

Shu, K., Liu, X. D., Xie, Q., and He, Z. H. (2016). Two faces of one seed: hormonal regulation of dormancy and germination. *Mol. Plant* 9, 34–45. doi: 10.1016/j.molp.2015.08.010

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6:31. doi: 10.1186/1471-2105-6-31

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinform.* 7:62. doi: 10.1186/1471-2105-7-62

Sun, Y., Lu, Z., Zhu, X., and Ma, H. (2020). Genomic basis of homoploid hybrid speciation within chestnut trees. *Nat. Commun.* 11:3375. doi: 10.1038/s41467-020-17111-w

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* 25, 4.10.1–4.10.14. doi: 10.1002/0471250953.bi0410s25

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Vanstraelen, M., and Benková, E. (2012). Hormonal interactions in the regulation of plant development. *Annu. Rev. Cell Dev. Biol.* 28, 463–487. doi: 10.1146/annurev-cellbio-101011-155741

Vieitez, A. M., San-José, M. C., and Corredoira, E. (2011). Cryopreservation of zygotic embryonic axes and somatic embryos of European chestnut. *Methods Mol. Biol.* 710, 201–213. doi: 10.1007/978-1-61737-988-8_15

Walters, C., Berjak, P., Pammenter, N., Kennedy, K., and Raven, P. (2013). Plant science. Preservation of recalcitrant seeds. *Science* 339, 915–916. doi: 10.1126/science.1230935

Wang, J., Tian, S., Sun, X., Cheng, X., Duan, N., Tao, J., et al. (2020). Construction of pseudomolecules for the Chinese chestnut (*Castanea mollissima*) genome. *G3 (Bethesda)* 10, 3565–3574. doi: 10.1534/g3.120.401532

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., et al. (2020). Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat. Commun.* 11:5817. doi: 10.1038/s41467-020-19682-0

Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., et al. (2018). Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* 50, 1435–1441. doi: 10.1038/s41588-018-0229-2

Wang, Y., Htwe, Y. M., Li, J., Shi, P., Zhang, D., Zhao, Z., et al. (2019). Integrative omics analysis on phytohormones involved in oil palm seed germination. *BMC Plant Biol.* 19:363. doi: 10.1186/s12870-019-1970-0

Wilf, P., Nixon, K. C., Gandolfo, M. A., and Cúneo, N. R. (2019). Eocene Fagaceae from Patagonia and Gondwanan legacy in Asian rainforests. *Science* 364:eaaw5139. doi: 10.1126/science.aaw5139

Xia, K., Hill, L. M., Li, D. Z., and Walters, C. (2014). Factors affecting stress tolerance in recalcitrant embryonic axes from seeds of four *Quercus* (Fagaceae) species native to the USA or China. *Ann. Bot.* 114, 1747–1759. doi: 10.1093/aob/mcu193

Xia, K., and Zhu, Z. Q. (2021). Characterization of physiological traits during development of the recalcitrant seeds of *Quercus serrata*. *Plant Biol.* 23, 1000–1005. doi: 10.1111/plb.13309

Xing, Y., Liu, Y., Zhang, Q., Nie, X., Sun, Y., Zhang, Z., et al. (2019). Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). *Gigascience.* 8:giz112. doi: 10.1093/gigascience/giz112

Zaynab, M., Pan, D., Fatima, M., Sharif, Y., Chen, S., and Chen, W. (2021). Proteomics analysis of *Cyclobalanopsis gilva* provides new insights of low seed germination. *Biochimie* 180, 68–78. doi: 10.1016/j.biochi.2020.10.008

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhao, M., Zhang, H., Yan, H., Qiu, L., and Baskin, C. C. (2018). Mobilization and role of starch, protein, and fat reserves during seed germination of

six wild grassland species. *Front. Plant Sci.* 9:234. doi: 10.3389/fpls.2018.00234

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S., et al. (2019). Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res.* 47, D1128–D1136. doi: 10.1093/nar/gky944

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fpls.2022.832559

Genomic Data Reveals Profound Genetic Structure and Multiple Glacial Refugia in *Lonicera oblata* (Caprifoliaceae), a Threatened Montane Shrub Endemic to North China

Xian-Yun Mu[1], Yuan-Mi Wu[1], Xue-Li Shen[1], Ling Tong[1], Feng-Wei Lei[1], Xiao-Fei Xia[2] and Yu Ning[3]*

[1] Laboratory of Systematic Evolution and Biogeography of Woody Plants, School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China, [2] Beijing Museum of Natural History, Beijing, China, [3] Institute of Wetland Research, Chinese Academy of Forestry, Beijing, China

OPEN ACCESS

Edited by:
Rong Wang,
East China Normal University, China

Reviewed by:
Kai Jiang,
Shanghai Chenshan Plant Science Research Center, Chenshan Botanical Garden (CAS), China
Guangda Tang,
South China Agricultural University, China

***Correspondence:**
Xian-Yun Mu
xymu85@bjfu.edu.cn
orcid.org/0000-0001-5434-3875

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *10 December 2021*
Accepted: *21 March 2022*
Published: *09 May 2022*

Citation:
Mu X-Y, Wu Y-M, Shen X-L, Tong L, Lei F-W, Xia X-F and Ning Y (2022) Genomic Data Reveals Profound Genetic Structure and Multiple Glacial Refugia in Lonicera oblata (Caprifoliaceae), a Threatened Montane Shrub Endemic to North China. Front. Plant Sci. 13:832559. doi: 10.3389/fpls.2022.832559

Characterizing genetic diversity and structure and identifying conservation units are both crucial for the conservation and management of threatened species. The development of high-throughput sequencing technology provides exciting opportunities for conservation genetics. Here, we employed the powerful SuperGBS method to identify 33, 758 high-quality single-nucleotide polymorphisms (SNP) from 134 individuals of a critically endangered montane shrub endemic to North China, *Lonicera oblata*. A low level of genetic diversity and a high degree of genetic differentiation among populations were observed based on the SNP data. Both principal component and phylogenetic analyses detected seven clusters, which correspond exactly to the seven geographic populations. Under the optimal $K = 7$, Admixture suggested the combination of the two small and geographically neighboring populations in the Taihang Mountains, Dongling Mountains, and Lijiazhuang, while the division of the big population of Jiankou Great Wall in the Yan Mountains into two clusters. High population genetic diversity and a large number of private alleles were detected in the four large populations, while low diversity and non-private alleles were observed for the remaining three small populations, implying the importance of these large populations as conservation units in priority. Demographic history inference suggested two drastic contractions of population size events that occurred after the Middle Pleistocene Transition and the Last Glacial Maximum, respectively. Combining our previous ecological niche modeling results with the present genomic data, there was a possible presence of glacial refugia in the Taihang and Yan Mountains, North China. This study provides valuable data for the conservation and management of *L. oblata* and broadens the understanding of the high biodiversity in the Taihang and Yan Mountains.

Keywords: conservation units, genetic diversity, *Lonicera oblata*, glacial refugia, Taihang Mountains, threatened species, Yan Mountains

INTRODUCTION

The Earth's biodiversity consists of approximately 9 million types of organisms (e.g., plants, animals, protists, and fungi), while biodiversity losses have exerted a profound impact on the ecology and society as a whole (Chapin et al., 2000; Cardinale et al., 2012). A total of 10% of tree species (> 8,000) on the earth are threatened with extinction (Cavender et al., 2015). Endemic species are important members of biodiversity hotspots and are of great value to biodiversity conservation (Myers et al., 2000). They are often characterized by narrow geographic ranges and specialized niche requirements, and they tend to have a small population size (Işik and Kani, 2011). Due to the sensitivity to internal factors (e.g., genetic bottleneck, inbreeding depression, and genetic drift) and/or external factors (e.g., human disturbance and environmental randomness), species with small populations are more vulnerable, which may lead to reduced fitness of certain individuals (Charlesworth and Willis, 2009), or even extinction. Moreover, low genetic variation and gene flow are frequently detected in rare plants compared to common plants (Cole, 2003). Understanding the genetic structure of endangered woody species can provide valuable information for their conservation and management.

The Sino-Japanese floristic region holds one of the oldest floras in the North Hemisphere with high species richness (Chen et al., 2018b; Lu et al., 2018). Climatic oscillations and geological events have greatly influenced the genetic pattern and distributional range of many plant species, particularly during the Quaternary glacial–interglacial cycles. Compared to North America and Europe, the fauna and flora in Asia were reported to be less affected by glaciation events. Northern China was less affected by massive ice sheets (Qiu et al., 2011), and an arid belt was suggested to have occurred in this region (Guo et al., 2008). Composing the main part of the Sino-Japanese floristic region, the northern China flora exhibits unique genetic patterns of forestry species. Southern warm-temperate forests mixed with northern cool-temperate forests occur in this region, indicating complicated genetic diversity and genetic differentiation patterns (Ye et al., 2017). It is commonly reported based on limited molecular markers that during the northward expansion of species, there would be a gradual decline in population genetic diversity [e.g., *Acer mono* Maxim. (Guo et al., 2014; Liu et al., 2014) and *Quercus mongolia* Fisch. ex Ledeb. (Zeng et al., 2015)]. In addition, single/multiple refugia in northeast China and high genetic diversity in northern populations are also observed [e.g., *Juglans mandshurica* Maxim. (Bai et al., 2010), *Eleutherococcus senticosus* (Rupr. and Maxim.) Maxim. (Wang et al., 2016), and *Schisandra sinensis* C. Bailley (Ye et al., 2019)]. North China is an important geographic and floristic region of northern China; it contains both numerous rugged mountains and the North China Basin and provides a north–south migration corridor for wildlife. Compared to the substantial progress made by previous literature focusing on the species with a broad geographic range, the genetic patterns of woody plants endemic to North China remain to be clearly understood, especially in the era of population genomics.

Two great mountains are located in North China, namely, the NNE trending Taihang Mountains (THMs) and the EW trending Yan Mountains (YMs), which harbor the highest seed plant diversity in northern China, as well as a high endemism rate (Wang et al., 1995). The THMs lie between the Ordos-Shanxi plateau on the west and the North China Basin on the east, with a north latitude from 34° to 40° (Wang and Li, 2008). The THMs are also a transitional region located between the second staircase (altitude 1,000–2,000 m) and the third staircase (altitude < 500 m) in the three-step topography of China. The intense uplift during the Late Pliocene to Pleistocene (Wu et al., 1999), along with heterogeneous local geographic and environmental changes (Guo et al., 2008; Wang et al., 2018), provided a unique habitat for several species endemic to this region, such as *Opisthopappus taihangensis* (Ling) Shih and *Taihangia rupestris* Yuet Li. The YMs lie in the north of the North China Basin with an east longitude from 115° to 119°. The west end of the YMs is connected to the northeast end of THMs. Most mountains in the THMs and YMs are extremely steep, particularly in the THMs. These two great mountains meet in the northwest of Beijing, and the valley between Changping District on the westside and Yanqing District on the northside are considered as the division. Uncovering the genetic composition of species endemic to the THMs and YMs may help us understand the genetic pattern and preservation of biodiversity in North China and can also allow us to further explore the role of these mountains in the survival and dispersal of endemic species.

High-throughput sequencing (HTS) technology can rapidly detect thousands of single-nucleotide polymorphisms (SNPs) in a cost- and time-efficient manner on a genome-wide scale. Compared to Sanger sequencing method, which generates limited loci, the large data generated from HTS can provide robust phylogeny, reducing the potential incomplete lineage sorting, and enable us to address a wide range of evolutionary questions. Furthermore, a small sample size per population (e.g., a minimum of two individuals) can also generate big data in the population genomic study based on the HTS approach (Nazareno et al., 2017). Thus, it is an effective tool for the assessment of population genetic diversity compared to traditional genetic markers (Glover et al., 2010; Li et al., 2020). Restriction site-associated DNA (RAD) is one of the most effective genotyping-by-sequencing (GBS) methods that allow for extensive SNP discovery on the genome level (Davey et al., 2011; Sonah et al., 2013; Puritz et al., 2014). It has been widely applied in population genetics for numerous model and non-model species (e.g., Wang et al., 2013; Luo et al., 2019; Feng et al., 2020; Xiong et al., 2020; Boukteb et al., 2021; Qiao et al., 2021). In particular, the UGbS-Flex (also denoted as SuperGBS) is a powerful tool for population genetic research with the ability to generate long reads (300–700 bp) and maximize SNP callings irrespective of the species ploidy level, breeding system, and reference genome availability (Qi et al., 2018; Cheng et al., 2020).

Lonicera oblata K. S. Hao ex P. S. Hsu and H. J. Wang (Caprifoliaceae) is a critically endangered montane shrub endemic to North China (Zhu et al., 2019; Wu et al., 2021), and it is ranked second in the list of national key protected wild plants in China.[1] Results from our 13-year field investigation identified

[1]http://www.forestry.gov.cn/main/5461/20210908/162515850572900.html

seven highly fragmented populations that typically grow in the THMs and YMs (**Figure 1**). The majority of individuals occurred in habitats of stony and steep cliffs in open forests with an elevation of ca. 1,100 m. Although the red berries of *L. oblata* are frequently observed in the wild, no seedlings are reported. The habitats of several populations are exposed to human activities (e.g., tourism, logging, and grazing), and some individuals are injured directly by cutoff. The highly threatened condition of *L. oblata* may be attributed to the small number and size of populations scattered in such a vast geographic range, unique but extremely fragmented habitats, strong human disturbances, and sensitivity to climate change (Wu et al., 2021). The highly isolated distributional pattern and mountain-top preferred habitat of *L. oblata* is similar to species in the sky islands in southwest China (He and Jiang, 2014), which may suffer both morphological and genetic variation because of local geographic vicariance and population isolation. With the exception of its endangerment status, little is known about the biological and genetic characteristics of *L. oblata*. Thus, research on the genetic diversity and population structure of *L. oblata* at the genome level is imperative for the adequate determination of conservation strategies for policy-makers.

In this study, we employed the novel approach of RAD sequencing, SuperGBS (Qi et al., 2018), on the threatened montane shrub *L. oblata* endemic to the two great mountains, the THMs and the YMs, in North China to (i) assess the population structure and genetic diversity across the full range of the species; (ii) infer the genetic signatures of potential refugia during the Quaternary climate oscillations in the THMs and YMs; and (iii) identify conservation units (CU) and provide suggestions for the conservation of this species. Our results provide both a theoretical basis and a valuable genetic database for the conservation planning of *L. oblata* and contribute to improving our understanding of the preservation and evolution of biodiversity in North China.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

A total of 134 accessions from all seven natural populations covering the entire geographic range of *L. oblata* were sampled. Based on the number of individuals, the seven populations can be classified into four large populations [Heduling (HDL, N = 24), Wutai Mountains (WTS, N = 34), Jiming Mountains (JMS, N = 20), and Jiankou Great Wall (JK, N = 36)] located at the species marginal distribution region and three small populations [Bijia Mountains (BJS, N = 6), Dongling Mountains (DLS, N = 6), and Lijiazhuang (LJZ, N = 8)] located at the central region (**Figure 1**). The only individual found in Beijing Songshan National Nature Reserve, which is near the JK population in Beijing, was not included in this study. Among them, JK and JMS are distributed in the YMs, while the remaining populations are located in the THMs (**Figure 1** and **Supplementary Table 1**). A minimum distance of 30 m from one individual to another was set during the sampling. Leaves were dried in silica gel in the field and stored at −80°C until further use. Voucher specimens were preserved in the herbarium at Beijing Forestry University. Genomic DNA was extracted from the above leaf tissue using the DNAsecure Plant Kit (Tiangen Biotech, Beijing, China) following the manufacturer's protocol. DNA quantity and quality were determined by Agarose Gel and NanoDrop.

Preparation of Libraries and Sequencing

Genotyping-by-sequencing was performed following the method described by Qi et al. (2018) using the enzyme combination of *Pst*I/*Msp*I. Briefly, 250 ng of DNA from each sample was double-digested with *Pst*I and *Msp*I, followed by the barcode adapter ligation at the *Pst*I site and a common Y-adapter at the *Msp*I site. Unligated adapters were removed by the recovery system of the improved magnetic bead. Recovered fragments with a length of 300–700 bp were PCR-amplified and tested for density using Qubit2.0 to ensure density values greater than 5 ng/μl. The libraries were subsequently sequenced using the Illumina HiSeq X Ten, PE150 platform at OE Biotech Co., Ltd., Qingdao, China.

Single-Nucleotide Polymorphisms Calling

Raw reads were split by barcode using Stacks version 2.4 (Catchen et al., 2013) with the option "process_radtags –renz-1 -r -s 0 –retain_header –adapter_mm 1 –adapter_1 –adapter_2 –b." The quality of passed reads was checked and filtered using Fastp version 0.20.0 (Chen et al., 2018a). Restriction sites, base quality values < 20, and the last 5 bp of the raw reads were likely to contain errors and were thus removed to obtain clean reads with the option "–n_base_limit 5 –cut_window_size 4 –cut_mean_quality 20 –length_required 75 –qualified_quality_phred 15." As the whole genome information of *L. oblata* was not yet reported at the time of this study, a *de novo* assemble of the GBS reference genome was generated in Stacks version 2.4 with the option "ustacks -M 3, cstacks -n 3, sstacks, tsv2bam, gstacks, populations" following Qi et al. (2018). Clean reads were aligned against the obtained GBS reference genome using Bowtie 2 version 2.3.4.1 with default parameters (Langmead and Salzberg, 2012), and subsequently genotyped and screened using GATK version 3.8–1 (McKenna et al., 2010) for SNP and INDEL predictions. Finally, the obtained SNPs were filtered using VCFtools version 0.1.16 (Danecek et al., 2011) with option "–maf 0.05 –max-missing 0.8 –minDP 4 –min-alleles 2 –max-alleles 2."

Genetic Diversity, Population Structure, and Phylogenetic Analysis

The number of private alleles, expected and observed heterozygosity (H_E and H_O), polymorphism information content (*PIC*), nucleotide diversity (π), the effective number of alleles (N_E), and Wright's F statistics [fixation index (F_{ST}) and inbreeding coefficient (F_{IS})] were calculated using the module "population" in Stacks version 2.4, VCFtools version 0.1.16, and genepop version 1.1.4 (Rousset, 2008). Reynold's genetic distance (DR) among populations was determined from F_{ST} as follows:

$$DR = -ln(1 - F_{ST}).$$

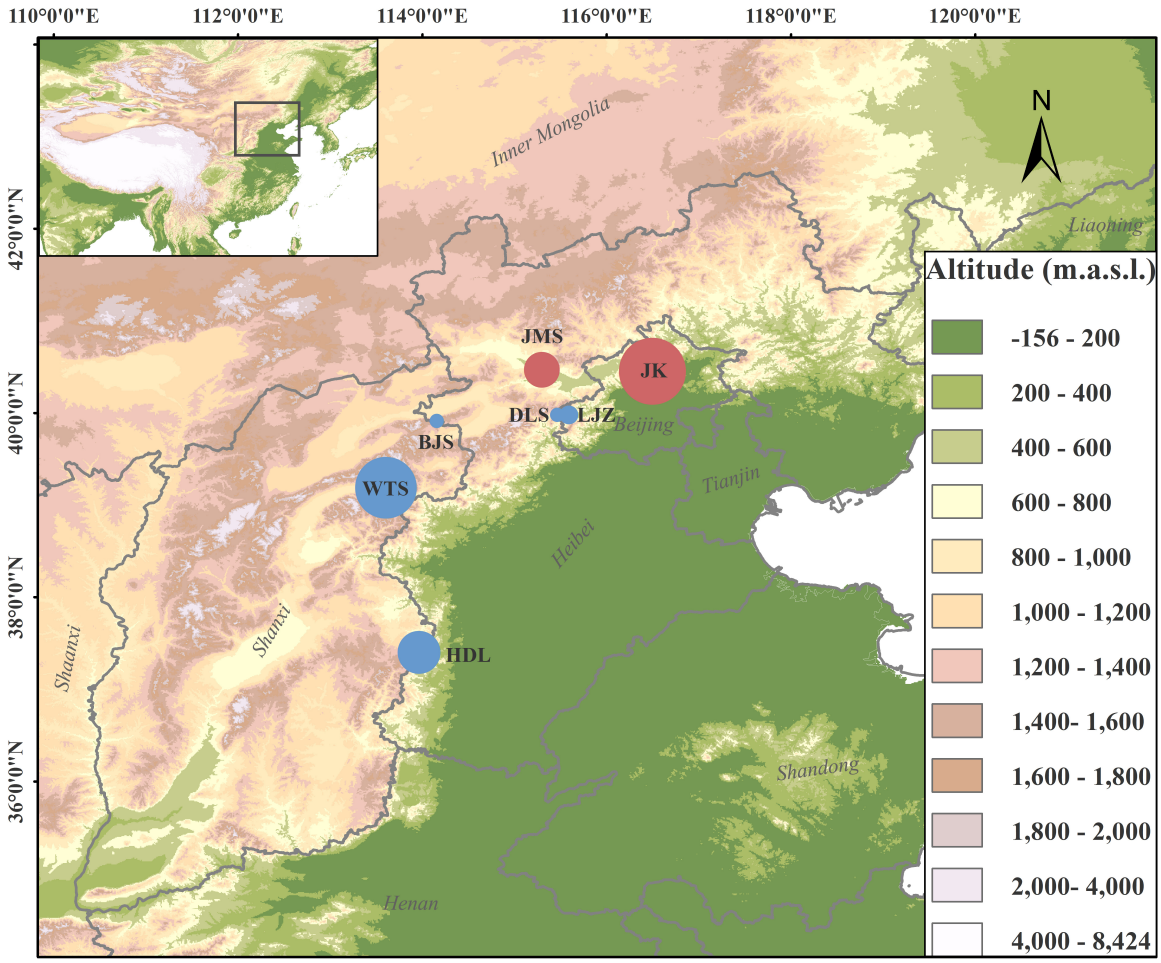


FIGURE 1 | Geographic distributions of sampled *Lonicera oblata* populations. The two red circles on the northside represent populations from Yan Mountains, and the five blue circles represent populations from Taihang Mountains. The circle size corresponds to the population sample size.

Evolutionary clusters were identified using Admixture version 1.3.0 (Alexander et al., 2009) with the default parameters. The predefined genetic clusters (K) ranged between 2 and 10, each of which was repeated 10 times. The optimal value of K was determined using cross-validation (CV) error, which has the minimum error in Admixture. Principal component analysis (PCA) was performed by GCTA version 1.26.0 (Yang et al., 2011) to explore the genetic structure of the species. A maximum likelihood-based phylogenetic tree was performed using IQ-TREE 2 (Minh et al., 2020) to clarify the genetic relationships among populations with the coalescent SNP dataset, and node support values (ML_{BS}) were calculated with 1,000 ultrafast bootstrap replicates.

Demographic History Inference

The demographic history of *L. oblata* was inferred using Stairway Plot 2 (Liu and Fu, 2020). The mutation rate was set as 7.7e-9 per site per year following Pu et al. (2020). The generation time was set as seven, based on our observations from the cultivated individual since its seed stage in 2016. The tallest individual is about 0.5 m in height, and it may bloom in the spring of 2022. Though the RAD sequencing did not cover the full genome, the obtained SNP are most likely a part of the latter, which may reflect the trend of its historical demography. Two representative populations, namely, the southmost large population HDL and the small LJZ located at central were employed for demographic history inference. All samples were treated as one group, a one-dimensional site frequency spectrum (1D-SFS) was constructed using ANGSD software (Korneliussen et al., 2014), and bootstrap iterations of 200 were implemented. The result was visualized in R.

RESULTS

Single-Nucleotide Polymorphisms Discovery

A total of 625, 652, 430 raw reads (91.93 Gb) were generated, resulting in 594, 664, 390 clean reads (86.82 Gb) and an average of 0.65 Gb per sample (**Supplementary Table 2**). The average

sequencing depth of all samples was 47.33 ×, with a coverage range of 87.82–96.86%. A total of 33,758 SNPs were obtained and employed for downstream analysis.

Population Genetic Diversity and Differentiation

Table 1 reports the descriptive statistics of genetic diversity (H_E, H_O, *PIC*, π, and N_E) for the seven populations of *L. oblata*. *PIC* values in all seven populations were very low (<0.19), with a value of 0.2364 at the species level. Relatively higher values of genetic diversity were observed among the four large populations (JK and JMS from YMs, WTS, and HDL from THMs) at the marginal regions compared to the three small populations located at the center (DLS, LJZ, and BJS). The northwest population WTS exhibited the highest values of H_E (0.2291), H_O (0.2054), *PIC* (0.1853), π (0.2328), and N_E (1.3826) (**Table 1**), while the lowest values of H_E (0.1477), *PIC* (0.1172), π (0.1625), and N_E (1.2563) were determined for BJS, one of the smallest populations. Furthermore, a large number of private alleles were detected in the four large populations (N = 1581 in JK, N = 256 in JMS, N = 724 in WTS, and N = 1,489 in HDL), while none were detected in the three small populations.

A high level of genetic differentiation (F_{ST} = 0.3245) was observed at the species level, along with a low inbreeding coefficient (F_{IS} = 0.0986). The highest genetic differentiation was detected between the most northeast population JK in the YMs and the most southwest population HDL in the THMs (F_{ST} = 0.3963), while the value was minimized between the two closest (geographically) neighbored populations located at the center of its distributional range, LJZ and DLS (F_{ST} = 0.1859, **Table 2**). The two YM populations, JK and JMS, exhibited large genetic variations (F_{ST} = 0.3099). The southernmost population HDL differed greatly from the other THM populations (F_{ST} of 0.2626–0.3805). Notably, WTS, one of the four largest populations on the northwest side, exhibited less variation with its neighbors, namely, BJS, DLS, LJZ, and HDL, with F_{ST} values varying from 0.2320 to 0.2723 (**Table 2**). The higher the F_{ST} between two populations, the longer the DR (**Table 2**), which may also suggest a lower gene flow.

Population Structure and Phylogenetic Analysis

The genetic structure of *L. oblata* was investigated by Admixture (**Figure 2A**). The CV analysis suggested an optimal K value of 7 (**Supplementary Figure 1**). However, the seven suggested clusters did not fully correspond to their seven natural populations. Admixture suggested that LJZ and DLS formed one cluster, and it identified JK to have a fine-scale genetic structure and divided it into two clusters (**Figure 2A**). The first principal component (PC1) of the PCA, which explained 15.79% of all genetic variance, was able to differentiate the seven geographic groups (**Figure 2B**). The four large natural populations were clearly identified, and the three small populations were plotted close to each other (**Figure 2B**). The SNP phylogeny revealed a strong phylogeographic pattern (**Figure 2C**), consisting of the YM and THM clades. Seven distinct lineages were detected, corresponding exactly to the seven natural populations with full clade supports (ML_{BS} = 100) (**Figure 2C** and **Supplementary Figure 2**).

Demographic History of *Lonicera oblata*

The variations of effective population size dating from one million years ago were inferred using Stairway Plot 2 (**Figure 3**). The two populations of *L. oblata*, namely, HDL and LJZ, both underwent drastic decline after the Middle Pleistocene Transition (ca. 1.2–0.7 million years ago, Clark et al., 2006). Although expanded later, the population size contracted dramatically again after the Last Glacial Maximum (LGM, ca. 22 ka years ago). A recent population expansion following the LGM was suggested for HDL, while the population size of LJZ was decreasing.

DISCUSSION

Genetic Composition and Variation

The genome-wide data generated by HTS are relatively easier to obtain than traditional markers, and the high number of informative loci provide us not only the solid basis for accurately characterizing the population structure (Nazareno et al., 2017) but also the exciting opportunity to address a wide range of factors for conservation genetics (Médail and Baumel, 2018; Li et al., 2020). We investigated the population genetic diversity of *L. oblata*, a critically endangered montane shrub endemic to North China, based on nuclear genomic data from the SuperGBS approach. The results reveal a low genetic diversity and strong genetic structure. *PIC* is an important index for the evaluation of genetic diversity and can be classified into three scales, such as $0 < PIC < 0.25$, $0.25 < PIC < 0.5$, and $PIC > 0.5$. The higher the *PIC* value, the higher the polymorphism information in the population (Botstein et al., 1980). In this study, low *PIC* (< 0.25) values were detected at the population and species level, suggesting a low degree of genetic diversity within *L. oblata*. H_E values determined here were markedly lower than other *Lonicera* species (e.g., H_E = 0.2863 in this study vs. H_E = 0.78 in *L. maackii* (Rupr.) Maxim.) (Barriball et al., 2015), as well as other endangered plants species [e.g., H_E = 0.3482 in *Tetraena mongolica* Maxim. (Cheng et al., 2020), H_E = 0.364 in *Firmiana danxiaensis* H. H. Hsue and H. S. Kiu (Chen et al., 2014; **Table 1**)]. Though only eight individuals were sampled, the LJZ population possesses a relatively high level of genetic diversity among the three small populations (JMS, DLS, and LJZ, **Table 1**). There are lots of mountains with altitudes of ca. 1,200 m in LJZ, whose ridges are connected. The vertical slopes between the top and the hillside in most mountains may decrease human interference and provide the opportunity for frequent genetic exchange within the population. Note that the genetic diversity of the central populations (BJS, DLS, and LJZ) was averagely lower than that of the four marginal populations (JK, JMS, WTS, and HDL) in *L. oblata* (**Table 1**), so the scenario of declining genetic diversity during the northward expansion of the species is not supported in this study. In contrast to the low genetic diversity, we detected pronounced genetic structure within *L. oblata* (F_{ST} = 0.3245), implying a long history of independent evolution and less gene flow among populations.

TABLE 1 | The statistics of the number of samples (*N*), expected heterozygosity (H_E), observed heterozygosity (H_O), polymorphism information content (*PIC*), nucleotide diversity (π), efficient allelic number (N_E), and the number of private alleles among populations (Pops).

Mountains	Pops	*N*	H_E	H_O	*PIC*	π	N_E	Private alleles
Yan Mountains	JK	36	0.2015	0.1787	0.1618	0.2046	1.3421	1,581
	JMS	20	0.1928	0.1828	0.1550	0.1982	1.3270	256
Taihang Mountains	DLS	6	0.1541	0.1664	0.1232	0.1694	1.2645	0
	LJZ	8	0.1888	0.1857	0.1524	0.2027	1.3165	0
	BJS	6	0.1477	0.1932	0.1172	0.1625	1.2563	0
	WTS	34	0.2291	0.2054	0.1853	0.2328	1.3826	724
	HDL	24	0.2044	0.1794	0.1642	0.2091	1.3469	1,489
ALL		134	0.2863	0.1863	0.2364	0.2874	1.4554	4,050

The population genetic composition is generally a function of both the species-specific biological characteristics and the corresponding ecological factors. The low genetic diversity and high genetic differentiation detected for *L. oblata* may be attributed to the following factors. (i) The extremely low numbers both at the population and individual level. The only seven uneven populations are highly fragmented and scattered at the top of mountains in a broad geographic range, and the limestone-preferred habitat of the species has been disturbed by long-term anthropogenic activities throughout history. The strong genetic differentiation among populations of *L. oblata* implies that the phenomenon of sky islands, which is significant in southwest China, may also occur in the THMs and the YMs in North China. (ii) The combined side effect of its reproductive system and fruit dispersal characteristics. A mixed mating system of biased outcross and partial self-fertilization is observed in *L. oblata*, and insects are necessary for pollination (Wu et al., 2022), while pollen limitation occurs during its flowering period. The harsh climate conditions during its flowering period (strong wind, sudden drop in temperature, snow, and/or dusty wind in the early spring) greatly interfere with the activities of effective pollinators and the pollination success of *L. oblata*. The highly fragmented populations and persistent strong winds at the top of hills and cliffs further inhibit pollination. The red fleshy fruits of *L. oblata* are attractive to birds, thus providing the possibility of long-distance dispersal of seeds and genetic exchange among populations. However, only a small-bodied bird *Zosterops* was recorded to feed on the fruits of *L. oblata* during our field investigation in Beijing in July. *Zosterops* live in a distance of ca. < 2 km in thick forests during the breeding season (May to July). Extending the distributional range of *L. oblata* over the contribution of birds like *Zosterops* is difficult. (iii) Seedlings barely survive in the field, which further decreases the genetic exchange between populations. Seedlings were difficult to find in the community during our field investigations, and an obvious inverted triangle population pyramid was observed (unpublished data). Hence, multiple negative effects may contribute to the strong genetic structure of *L. oblata*, whose destiny may be further challenged by future climate changes (Wu et al., 2021).

Population Structure and Glacial Refugia

The genetic structure of a population provides further information on the evolution of a species. In this study, a significant phylogeographic structure was identified from the structure, principal component, and phylogeny analyses, suggesting the division of seven groups (**Figure 2**). However, the seven clusters identified in Admixture did not correspond to those in the latter two analyses. In particular, the Admixture analysis placed DLS and LJZ into a single cluster, while the eastmost large population JK was divided into two clusters (**Figure 2A**). When $K = 5$, samples of JK did not differentiate intrapopulations, and the three small populations, namely, BJS, DLS, and LJZ, compose one cluster (**Supplementary Figure 3**). Such a scenario suggests an Admixture identity of these three small populations distributed in the species' central region. The presence of two clusters in JK under the scenario $K = 7$ is difficult to explain. The six samples from one of the JK clusters (JK20–JK25) were randomly selected in the field just like the other samples. There were no distinct characteristics (e.g., elevation, slope direction, and canopy density) identified for the six samples compared to the remaining samples. JK is strongly exposed to tourism activity compared to the other six populations. The Jiankou Great Wall, a grand section of the Great Wall, attracts thousands of tourists each year. Many *L. oblata* individuals grow right on the Great Wall and are threatened by tourists during their climb. Some deaths of individuals due to the direct damage from tourists were witnessed during our fieldwork. Future work will include fine-scale sampling and landscape genomic study in order to understand the fine-scale genetic structure within JK.

The PCA and phylogenetic analysis revealed the seven identified groups to fully correspond to the natural populations

TABLE 2 | Genetic differentiation coefficient (F_{ST}) and Reynolds genetic distance (DR) among populations.

Pops	JK	JMS	LJZ	DLS	BJS	WTS	HDL
JK	–	0.3709	0.3683	0.4385	0.4974	0.3895	0.5047
JMS	0.3099	–	0.2497	0.3245	0.4460	0.3455	0.4861
LJZ	0.3081	0.2210	–	0.2057	0.3692	0.2657	0.4214
DLS	0.3550	0.2771	0.1859	–	0.4763	0.3179	0.4788
BJS	0.3919	0.3598	0.3087	0.3789	–	0.2640	0.4740
WTS	0.3226	0.2921	0.2333	0.2723	0.2320	–	0.3046
HDL	0.3963	0.3850	0.3439	0.3805	0.3775	0.2626	–

The lower triangle presents interpopulation F_{ST}, and the upper triangle presents the DR.

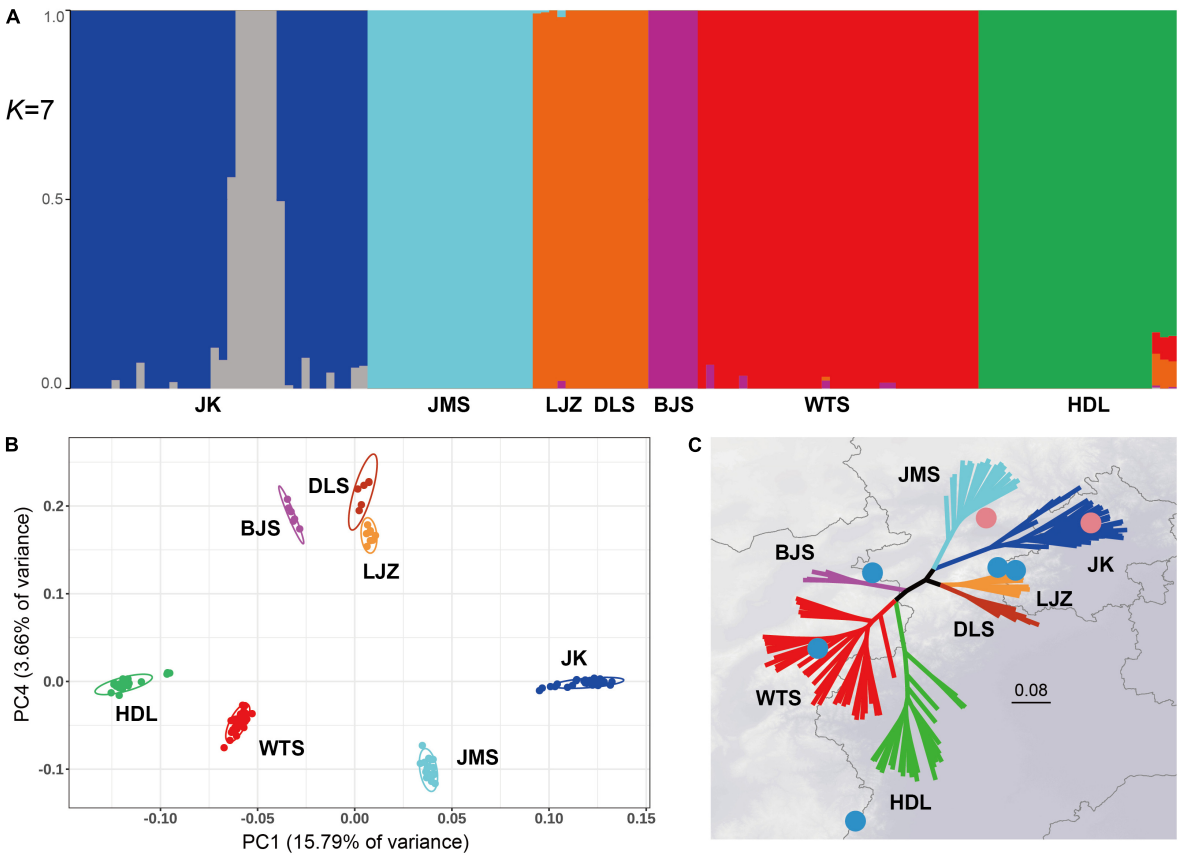


FIGURE 2 | Genetic structure and phylogenetic relationships of the seven *Lonicera oblata* populations. **(A)** Admixture proportions of genetic clusters for each individual of the seven populations. The scenario of K = 7 is the best value according to cross-validation analysis. **(B)** Principal component analysis plot for the 134 *L. oblata* individuals based on the first two principal components. **(C)** A maximum-likelihood tree based on 33,758 single-nucleotide polymorphisms (SNPs) of the nuclear genome, with seven fully supported lineages (ML_{BS} = 100), which exactly correspond to species' natural populations. The two pink circles on the northside represent populations from Yan Mountains, and the five blue circles represent populations from Taihang Mountains.

of the species. Clade support values do not generally reach high levels in the study of population genetic diversity, and the sample(s) of one population may sometimes be resolved in the lineage of another population, such as cushion willow (Chen et al., 2019), grape (Liang et al., 2019), *Prunus* (Liang et al., 2019), and *Q. aquifolioides* Rehder and E. H. Wilson (Du et al., 2017). In this study, we obtained a well-resolved phylogeny of *L. oblata* at the population level. The phylogenetic tree consisted of two key clades, namely, the THMs and YMs, and the seven natural populations were all resolved as monophyly with full clade supports on the molecular tree (ML_{BS} = 100) (**Figure 2C**). Both the strong phylogeographic pattern and the fully supported clade values at the population level suggest the seven natural populations to be evolutionary significant units following the definition of Moritz (1994). Strong phylogeographic characteristics combined with great genetic variation among populations (**Table 2**) also imply these populations be defined as management units following Avise (2000). Among them, the four large populations (JK and JMS in the YMs and WTS and HDL in the THMs), which have plenty of private alleles, are of greater importance as primary conservation units. The high level of genetic differentiation and well-resolved phylogeny in *L. oblata* may also suggest that long-term vicariance exists among these highly fragmented populations, as well as a low gene flow.

The NNE trending THMs and EW trending YMs act as dispersal corridors not only between the northeast cool-temperate and southern warm-temperate forests but also among fragmented populations of endangered species in China. These two great mountains in North China may also provide potential refugia for species during the global climate changes of the Quaternary (Zeng et al., 2011; Hou et al., 2020; Lin et al., 2021). Our results demonstrate many of the extant populations of *L. oblata* are genetically isolated. A high level of genetic diversity and plentiful private alleles were detected in the four large populations located in the marginal region of the *L. oblata* distribution range (N = 1,581 in JK, N = 256 in JMS, N = 724 in WTS, and N = 1,489 in HDL, **Table 1**). In contrast, low genetic diversity and no private alleles were detected in the populations located in the central region (LJZ, DLS, and BJS). Our ecological niche modeling analysis (ENM) of *L. oblata* suggests a wide historical distribution range, covering its current range since

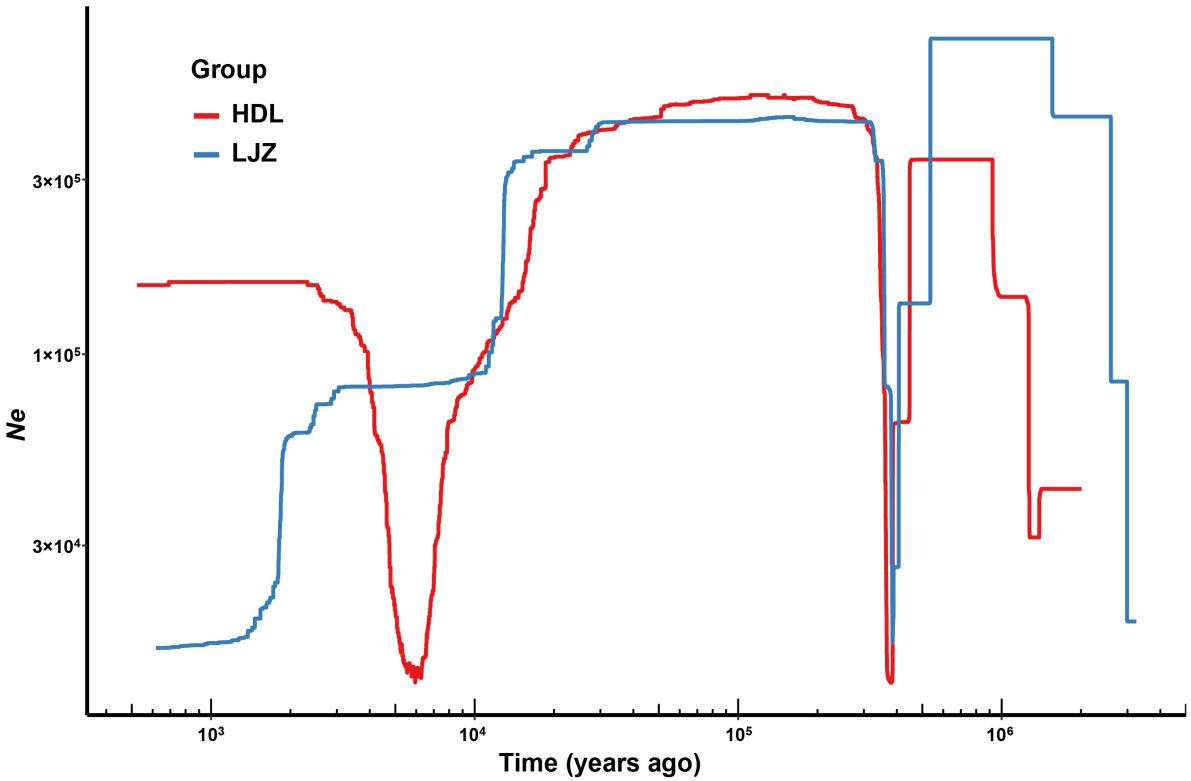


FIGURE 3 | Demographic history of two populations of *Lonicera oblata*, namely, Heduling (HDL) and Lijiazhuang (LJZ) inferred by Stairway Plot 2. The x-axis indicates the time before the present, and the y-axis represents the historical effective population size.

the LGM, followed by a dramatic decline to the present (Wu et al., 2021). Affected by the intensive uplift of both the Qinghai-Tibetan Plateau during the Late Cenozoic (Li et al., 2001) and THMs and YMs during the Late Pliocene to Pleistocene (Wu et al., 1999), which intensified monsoon while enhanced aridity in the Asian interior and repeated glacial events during the Quaternary, historical climate and geographic characteristics of northern China varied greatly. The demographic history of *L. oblata* may also have suffered from these factors, particularly the LGM, and its population size decreased sharply (**Figure 3**). This might also explain that no private allele was detected in the three small current populations, namely, BJS, DLS, and LJZ. Taken together, the genomic data and ENM suggest the existence of LGM refugia for *L. oblata* in the THMs and YMs, and the hypothesis of *in situ* survival is supported. The two greatest mountains in North China, namely the THMs and the YMs, may have acted as "Noah's Ark" for numerous plant lineages during the LGM, providing vital refugia for the postglacial preservation of biodiversity in North China.

Threats and Conservation Suggestions for *Lonicera oblata*

The endangerment of species may be a result of both internal and/or external factors. Our field investigation suggests the possible worsening of the *L. oblata* endangerment situation. Although it occupies a relatively large distribution range in North China, only seven populations and a total of ca. 1,000 individuals of the species were recorded. Its limestone-specific habitat may greatly restrict its expansion, and the limited populations are highly fragmented and affected by human activities (e.g., agriculture, tourism, and logging). Furthermore, pollen limitation and harsh climate conditions during its flowering period restrain its survival and reproduction (Wu et al., 2022). Our results reveal a low genetic diversity and high genetic differentiation, which may suggest a reduction in the species fitness. Moreover, potentially suitable regions for the expansion of the species are limited (Wu et al., 2021). The dry Chinese Loess Plateau to the west of the THMs, the cold and dry desert in the Yin Mountains to the north of the YMs, and the cold temperature in Northeast China all constrain the future expansion of *L. oblata* severely. Due to the great conservation achievements of forestry and ecology across China in recent decades, the majority of previously bare mountaintops are now densely occupied by woody plants. This may also act as a challenge to the expansion of *L. oblata*, which favors an open habitat. Thus, this critically endangered species endemic to North China is currently facing the plight of "nowhere to go" (Nogués-Bravo et al., 2007; Loarie et al., 2009).

The exploration of population genetic patterns can provide vital information for the conservation and management of threatened species. Considering the aforementioned threatening

factors, several conservation suggestions are proposed here: (i) forceful *in situ* conservation should be provided at both the population and the individual level; (ii) JK, JMS, HDL, and WTS populations should be given priority among the seven conservation units, including the construction of mini-reserves for the former three that are not distributed in national parks or nature reserves, and forceful conservation actions should be performed; (iii) the collection and long-term preservation of seeds from multi-populations should be accomplished, with the timely determination of seed germination and artificial propagation techniques; and (iv) as the intersection regions between the north of the THMs and west of the YMs are predicted to provide stable climate conditions in the future (Wu et al., 2021), this area should be given the priority to experiments of reintroduction and *ex situ* conservation.

CONCLUSION

In this study, we performed a genome-wide population genetic investigation on *L. oblata*, a highly threatened montane shrub endemic to North China, whose natural distribution range lies in the contact zone between southern warm-temperate and northern cold-temperate forests. Based on the SuperGBS method, we determined a low level of genetic diversity, a high degree of genetic differentiation, and a strong phylogeographic structure. Sharp population size contraction was suggested, which occurred after the middle Pleistocene Transition and the LGM, respectively. The results from the current genomic data combined with our previous ENM analysis suggest the existence of LGM refugia in the THMs and YMs. Affected by multiple factors relating to geography, climate, and biological and ecological characteristics, the future of *L. oblata* is challenged by the "nowhere to go" scenario. Four key conservation units were identified for *L. oblata*, and several conservation suggestions were provided. Further work characterizing the whole genome of *L. oblata* and a fine-scale landscape genomic study should be performed especially on the JK population with the aim to understand how geographic and ecological factors shape its genetic structure and evolutionary history. Our population genomic study provides valuable information for the conservation and management of *L. oblata* and contributes to the further understanding of the role of the THMs and YMs in preserving the biodiversity in North China.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, PRJNA787048.

AUTHOR CONTRIBUTIONS

X-YM conceptualized, supervised the study, analyzed the data, wrote, and revised the draft. Y-MW, X-LS, LT, F-WL, X-FX, and YN contributed to the fieldwork and collected the materials. Y-MW contributed to the species distribution maps. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Beijing Natural Science Foundation (grant no. 5192012), the National Natural Science Foundation of China (grant nos. 32070235 and 31800348), and the Plant Germplasm Investigation Program from the Ministry of Agriculture and Rural Affairs of China (grant no. 13200346).

ACKNOWLEDGMENTS

We thank Nan Yang at Beijing Baihuashan National Nature Reserve, Wan-Jie Jiang at Beijing Songshan National Nature Reserve, and Ying-Chun Xie at Yangyuan No. 1 Middle School in Hebei Province for their help during our fieldwork.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.832559/full#supplementary-material

Supplementary Table S1 | Population information.

Supplementary Table S2 | Sequencing information.

Supplementary Figure S1 | The optimal *K* value.

Supplementary Figure S2 | Phylogenetic tree of 134 individuals based on SNP data matrix.

Supplementary Figure S3 | Structure analysis of *K* = 2–10.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Avise, J. C. (2000). *Phylogeography — the History and Formation of Species*. Cambridge, MA: Harvard University Press.

Bai, W. L., Liao, W. J., and Zhang, D. Y. (2010). Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytol.* 188, 892–901. doi: 10.1111/j.1469-8137.2010.03407.x

Barriball, K., McNutt, E. J., Gorchov, D. L., and Rocha, O. J. (2015). Inferring invasion patterns of *Lonicera maackii* (Rupr) Herder (Caprifoliaceae) from the genetic structure of 41 naturalized populations in a recently invaded area. *Biol. Invasions* 17, 2387–2402. doi: 10.1007/s10530-015-0882-7

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Boukteb, A., Sakaguchi, S., Ichihashi, Y., Kharrat, M., Nagano, A. J., Shirasu, K., et al. (2021). Analysis of genetic diversity and population structure of *Orobanche foetida* populations from *Tunisia* using RADseq. *Front. Plant Sci.* 12:618245. doi: 10.3389/fpls.2021.618245

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., et al. (2012). Biodiversity loss and its impact on humanity. *Nature* 486, 59–67. doi: 10.1038/nature11148

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Cavender, N., Westwood, M., Bechtoldt, C., Donnelly, G., Oldfield, S., Gardner, M., et al. (2015). Strengthening the conservation value of ex situ tree collections. *Oryx* 49, 416–424. doi: 10.1017/S0030605314000866

Chapin, F. S. III, Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., et al. (2000). Consequences of changing biodiversity. *Nature* 405, 234–242. doi: 10.1038/35012241

Charlesworth, D., and Willis, J. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.* 10, 783–796. doi: 10.1038/nrg2664

Chen, J. H., Huang, Y., Brachi, B., Yun, Q. Z., Zhang, W., Lu, W., et al. (2019). Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. *Nat. Commun.* 10:5230. doi: 10.1038/s41467-019-13128-y

Chen, S., Li, M., Hou, R., Liao, W., Zhou, R., and Fan, Q. (2014). Low genetic diversity and weak population differentiation in *Firmiana danxiaensis*, a tree species endemic to Danxia landform in northern Guangdong, China. *Biochem. Syst. Ecol.* 55, 66–72. doi: 10.1016/j.bse.2014.02.029

Chen, Y. S., Deng, T., Zhou, Z., and Sun, H. (2018b). Is the East Asian flora ancient or not? *Natl. Sci. Rev.* 5, 920–932. doi: 10.1093/nsr/nwx156

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018a). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Cheng, J., Kao, H. X., and Dong, S. B. (2020). Population genetic structure and gene flow of rare and endangered *Tetraena mongolica* Maxim. revealed by reduced representation sequencing. *BMC Plant Biol.* 20:391. doi: 10.1186/s12870-020-02594-y

Clark, P. U., Archer, D., Pollard, D., Blum, J. D., Rial, J. A., Brovkin, V., et al. (2006). The middle Pleistocene transition: characteristics, mechanisms, and implications for long-term changes in atmospheric pCO2. *Quat. Sci. Rev.* 25, 3150–3184. doi: 10.1016/j.quascirev.2006.07.008

Cole, C. T. (2003). Genetic variation in rare and common plants. *Ann. Rev. Ecol. Evol. Syst.* 34, 213–237. doi: 10.1146/annurev.ecolsys.34.030102.151717

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

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

Du, F. K., Hou, M., Wang, W., Mao, K., and Hampe, A. (2017). Phylogeography of *Quercus aquifolioides* provides novel insights into the Neogene history of a major global hotspot of plant diversity in south-west China. *J. Biogeogr.* 44, 294–307. doi: 10.1111/jbi.12836

Feng, J., Zhao, S., Li, M., Zhang, C., Qu, H., Li, Q., et al. (2020). Genome-wide genetic diversity detection and population structure analysis in sweetpotato (*Ipomoea batatas*) using RAD-seq. *Genomics* 112, 1978–1987. doi: 10.1016/j.ygeno.2019.11.010

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyheim, B., and Skaala, Ø (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genet.* 11:2. doi: 10.1186/1471-2156-11-2

Guo, X. D., Wang, H. F., Bao, L., Wang, T. M., Bai, W. L., Ye, J. W., et al. (2014). Evolutionary history of a widespread tree species Acer mono in East Asia. *Ecol. Evol.* 4, 4332–4345. doi: 10.1002/ece3.1278

Guo, Z. T., Sun, B., Zhang, Z. S., Peng, S. Z., Xiao, G. Q., Ge, J. Y., et al. (2008). A major reorganization of Asian climate by the early Miocene. *Clim. Past* 4, 153–174. doi: 10.5194/cp-4-153-2008

He, K., and Jiang, X. L. (2014). Sky islands of southwest China. I. An overview of phylogeographic patterns. *Chin. Sci. Bull.* 59, 585–597. doi: 10.1007/s11434-013-0089-1

Hou, H., Ye, H., Wang, Z., Wu, J., Gao, Y., Han, W., et al. (2020). Demographic history and genetic differentiation of an endemic and endangered *Ulmus lamellosa* (Ulmus). *BMC Plant Biol.* 20:526. doi: 10.1186/s12870-020-02723-7

Işik, K., and Kani (2011). Rare and endemic species: why are they prone to extinction? *Turk. J. Bot.* 35, 411–417. doi: 10.3906/bot-1012-90

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, J., Milne, R. I., Ru, D., Miao, J., Tao, W., Zhang, L., et al. (2020). Allopatric divergence and hybridization within *Cupressus chengiana* (Cupressaceae), a threatened conifer in the northern Hengduan Mountains of western China. *Mol. Ecol.* 29, 1250–1266. doi: 10.1111/mec.15407

Li, J. J., Fang, X. M., Pan, B. T., Zhao, Z. J., and Song, Y. G. (2001). Late Cenozoic intensive uplift of Qinghai-Xizang Plateau and its impacts on environments in surrounding area. *Quat. Sci.* 21, 381–391. doi: 10.3321/j.issn:1001-7410.2001.05.001

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., et al. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nat. Commun.* 10:1190. doi: 10.1038/s41467-019-09135-8

Lin, N., Landis, J. B., Sun, Y. X., Huang, X. H., Zhang, X., Liu, Q., et al. (2021). Demographic history and local adaptation of *Myripnois dioica* (Asteraceae) provide insight on plant evolution in northern China flora. *Ecol. Evol.* 11, 8000–8013. doi: 10.1002/ece3.7628

Liu, C., Tsuda, Y., Shen, H., Hu, L., Saito, Y., and Ide, Y. (2014). Genetic structure and hierarchical population divergence history of Acer mono var. mono in south and northeast China. *PLoS One* 9:e87187. doi: 10.1371/journal.pone.0087187

Liu, X. M., and Fu, Y. X. (2020). Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* 21:280. doi: 10.1186/s13059-020-02196-9

Loarie, S. R., Duffy, P. B., Hamilton, H., Asner, G. P., Field, C. B., and Ackerly, D. D. (2009). The velocity of climate change. *Nature* 462, 1052–1055. doi: 10.1038/nature08649

Lu, L. M., Mao, L. F., Yang, T., Ye, J. F., Liu, B., Li, H. L., et al. (2018). Evolutionary history of the angiosperm flora of China. *Nature* 554, 234–238. doi: 10.1038/nature25485

Luo, Z., Brock, J., Dyer, J. M., Kutchan, T., Schachtman, D., Augustin, M., et al. (2019). Genetic diversity and population structure of a *Camelina sativa* spring panel. *Front. Plant Sci.* 10:184. doi: 10.3389/fpls.2019.00184

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Médail, F., and Baumel, A. (2018). Using phylogeography to define conservation priorities: the case of narrow endemic plants in the Mediterranean Basin hotspot. *Biol. Conserv.* 224, 258–266. doi: 10.1016/j.biocon.2018.05.028

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Moritz, C. (1994). Applications of mitochondrial DNA analysis in conservation: a critical review. *Mol. Ecol.* 3, 401–411. doi: 10.1111/j.1365-294X.1994.tb00080.x

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi: 10.1038/35002501

Nazareno, A. G., Bemmels, J. B., Dick, C., and Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol. Ecol.* 17, 1136–1147. doi: 10.1111/1755-0998.12654

Nogués-Bravo, D., Araújo, M. B., Errea, M. P., and Martínez-Rica, J. P. (2007). Exposure of global mountain systems to climate warming during the 21st Century. *Glob. Environ. Chang.* 17, 420–428. doi: 10.1016/j.gloenvcha.2006.11.007

Pu, X. D., Li, Z., Tian, Y., Gao, R. R., Hao, L. J., Hu, Y. T., et al. (2020). The honeysuckle genome provides insight into the molecular mechanism of

carotenoid metabolism underlying dynamic flower coloration. *New Phytol.* 227:930e43. doi: 10.1111/nph.16552

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., and Bird, C. E. (2014). Demystifying the RAD fad. *Mol. Ecol.* 23, 5937–5942. doi: 10.1111/mec.12965

Qi, P., Gimode, D., Saha, D., Schröder, S., Chakraborty, D., Wang, X., et al. (2018). UGbS-Flex, a novel bioinformatics pipeline for imputation-free SNP discovery in polyploids without a reference genome: finger millet as a case study. *BMC Plant Biol.* 18:117. doi: 10.1186/s12870-018-1316-3

Qiao, Y., Guo, F., Huo, N., Zhan, L., Sun, J., Zuo, X., et al. (2021). Genotyping-by-sequencing to determine the genetic structure of a Tibetan medicinal plant *Swertia mussotii* Franch. *Genet. Resour. Crop Evol.* 68, 469–484. doi: 10.1007/s10722-020-00993-6

Qiu, Y., Fu, C., and 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. doi: 10.1016/j.ympev.2011.01.012

Rousset, F. (2008). Genepop' 007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLos One* 8:e54603. doi: 10.1371/journal.pone.0054603

Wang, H., Li, J. M., and Wu, T. W. (2018). Characteristics and genesis of geoheritage resources of Taihang Mountain. *Acta Sci. Nat. Univ. Pekin.* 54, 546–554. doi: 10.13209/j.0479-8023.2017.098

Wang, H. S., Zhang, Y. L., Huang, J. S., Wu, Z. F., Zhao, S. L., Wang, H. S., et al. (1995). A floristic study on the seed plants in the North China region. *Acta Bot. Yunnanica* 7, 32–54.

Wang, N., Thomson, M., Bodles, W. J. A., Crawford, R. M. M., Hunt, H. V., Featherstone, A. W., et al. (2013). Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol. Ecol.* 22, 3098–3111. doi: 10.1111/mec.12131

Wang, S. H., Bao, L., Wang, T. M., Wang, H. F., and Ge, J. P. (2016). Contrasting genetic patterns between two coexisting *Eleutherococcus* species in northern China. *Ecol. Evol.* 6, 3311–3324. doi: 10.1002/ece3.2118

Wang, Y., and Li, H. (2008). Initial formation and Mesozoic tectonic exhumation of an intracontinental tectonic belt of the northern part of the Taihang Mountain Belt, Eastern Asia. *J. Geol.* 116, 155–172. doi: 10.1086/529153

Wu, C., Zhang, X., and Ma, Y. (1999). The Taihang and Yan mountains rose mainly in Quaternary. *North China Earthquake Sci.* 17, 1–7. doi: 10.1007/s11103-011-9753-5

Wu, Y. M., Shen, X. L., Tong, L., Lei, F. W., Mu, X. Y., and Zhang, Z. X. (2021). Impact of past and future climate change on the potential distribution of an endangered montane shrub *Lonicera oblata* and its conservation implications. *Forests* 12:125. doi: 10.3390/f12020125

Wu, Y. M., Shen, X. L., Tong, L., Lei, F. W., Xia, X. F., Mu, X. Y., et al. (2022). Reproductive biology of an endangered lithophytic shrub and implications for its conservation. *BMC Plant Biol.* 22:80. doi: 10.1186/s12870-022-03466-3

Xiong, S., Zhao, Y., Chen, Y., Gao, M., Wu, L., and Wang, Y. (2020). Genetic diversity and population structure of *Quercus fabri* Hance in China revealed by genotyping-by-sequencing. *Ecol. Evol.* 10, 8949–8958. doi: 10.1002/ece3.6598

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Ye, J. W., Yuan, Y. G., Cai, L., and Wang, X. J. (2017). Research progress of phylogeographic studies of plant species in temperate coniferous and broadleaf mixed forests in Northeastern China. *Biodivers. Sci.* 25, 1339–1349. doi: 10.17520/biods.2017265

Ye, J. W., Zhang, Z. K., Wang, H. F., Bao, L., and Ge, J. P. (2019). Phylogeography of *Schisandra chinensis* (Magnoliaceae) reveal multiple refugia with ample gene flow in northeast China. *Front. Plant Sci.* 10:199. doi: 10.3389/fpls.2019.00199

Zeng, Y. F., Liao, W. J., Petit, R. J., and Zhang, D. Y. (2011). Geographic variation in the structure of oak hybrid zones provides insights into the dynamics of speciation. *Mol. Ecol.* 20, 4995–5011. doi: 10.1111/j.1365-294X.2011.05354.x

Zeng, Y. F., Wang, W. T., Liao, W. J., Wang, H. F., and Zhang, D. Y. (2015). Multiple glacial refugia for cool-temperate deciduous trees in northern East Asia: the Mongolian Oak as a case study. *Mol. Ecol.* 24, 5676–5691. doi: 10.1111/mec.13408

Zhu, Y. X., Wu, Y. M., Shen, X. L., Tong, L., Xia, X. F., Mu, X. Y., et al. (2019). The complete chloroplast genome of *Lonicera oblata*, a critically endangered species endemic to North China. *Mitochondrial DNA B* 4, 2337–2338. doi: 10.1080/23802359.2019.1629344

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 17 May 2022
doi: 10.3389/fpls.2022.879822

High-Quality Genome Assembly of *Olea europaea* subsp. *cuspidata* Provides Insights Into Its Resistance to Fungal Diseases in the Summer Rain Belt in East Asia

Li Wang[1†], Jianguo Zhang[1,2], Dan Peng[3], Yang Tian[1], Dandan Zhao[1], Wanning Ni[1], Jinhua Long[1], Jinhua Li[1], Yanfei Zeng[1,2], Zhiqiang Wu[3], Yiyun Tang[4] and Zhaoshan Wang[1,2*†]*

[1] Key Laboratory of Silviculture of the State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China, [2] Collaborative Innovation Center of Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China, [3] Kunpeng Institute of Modern Agriculture at Foshan, Foshan, China, [4] Ecological Restoration and Industrial Development Workstation, Nujiang State Forestry and Grassland Bureau of Yunnan Province, Kunming, China

The olive tree (*Olea europaea* L.) is the most iconic fruit crop of the Mediterranean Basin. Since the plant was introduced to China in the 1960s, the summer rain climate makes it susceptible to pathogens, leading to some olive diseases. *Olea europaea* L. subsp. *cuspidata* is natively distributed in the Yunnan province of China. It has a smaller fruit size, lower oil content, and higher resistance compared to subsp. *europaea*, which makes subsp. *cuspidata* a critical germplasm resource to be investigated. Here, a high-quality genome of subsp. *cuspidata* with 1.38 Gb in size was assembled and anchored onto 23 pseudochromosomes with a mounting rate of 85.57%. It represents 96.6% completeness [benchmarking universal single-copy orthologs (BUSCO)] with a contig N50 of 14.72 Mb and a scaffold N50 of 52.68 Mb, which shows a significant improvement compared with other olive genomes assembled. The evaluation of the genome assembly showed that 92.31% of resequencing reads and an average of 96.52% of assembled transcripts could be aligned to the assembled genome. We found that a positively selected gene, *evm.model.Chr16.1133*, was shared with the results of transcriptome analysis. This gene belongs to the susceptible gene and negatively regulates the disease resistance process. Furthermore, we identified the *Cercospora* genus which causes the leaf spot disease in the infected leaves. The high-quality chromosome-level genomic information presented here may facilitate the conservation and utilization of germplasm resources of this subspecies and provide an essential genetic basis for further research into the differences in oil content and resistance between subsp. *cuspidata* and *europaea*.

Keywords: genome assembly, *Olea europaea*, susceptibility gene, demographic history, nature selection

OPEN ACCESS

Edited by:
Fang Du,
Beijing Forestry University, China

Reviewed by:
Cheng Sun,
Institute of Apiculture Research (CAAS), China
Liang Tang,
Hainan University, China

***Correspondence:**
Zhaoshan Wang
w@caf.ac.cn
Jianguo Zhang
zhangjg@caf.ac.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *20 February 2022*
Accepted: *14 March 2022*
Published: *17 May 2022*

Citation:
Wang L, Zhang J, Peng D, Tian Y, Zhao D, Ni W, Long J, Li J, Zeng Y, Wu Z, Tang Y and Wang Z (2022) High-Quality Genome Assembly of Olea europaea subsp. cuspidata Provides Insights Into Its Resistance to Fungal Diseases in the Summer Rain Belt in East Asia. Front. Plant Sci. 13:879822. doi: 10.3389/fpls.2022.879822

INTRODUCTION

The olive tree (*Olea europaea* L.) is the most iconic fruit crop of the Mediterranean Basin owing to its ecological, economical, and cultural significance. It constitutes a cornerstone of Mediterranean culture by its multiple past and present uses and omnipresence in traditional agrosystems (Gros-Balthazard et al., 2019). Virgin olive oil, the main product from olive trees and the principal component of the so-called Mediterranean diet, is recognized as a green health care cooking oil and is known as "liquid gold" for its high nutritional benefits, outstanding medical treatment and health care function, and exceptional organoleptic properties (Donaire et al., 2011). The olive plant was formally introduced into China in the 1960s and was mainly cultivated in subtropical areas (Han and He, 2007). In the Mediterranean region, the climate is hot and dry in summer and moderate and rainy in winter, and the sandy soil is neutral and alkaline. In China's cultivation region, however, the climate is hot and rainy in summer, cold and dry in winter, and the soil is mostly acidic with a heavy texture (Wang et al., 2000).

Although more than 200 cultivars are now grown in China, most of them exhibit climate and soil incompatibility, accompanied by the emergence of some olive diseases caused by fungi and bacteria (Han and He, 2007), owing to the rainy and highly humid conditions which are conducive to the occurrence and development of diseases (Moral and Trapero, 2009). Some diseases are widespread in many olive plantations, such as *Cercospora cladosporioides* Sacc., *Cycloconium oleaginum Cast*, olive anthracnose, and leaf spot diseases. These diseases are considered to be important factors leading to the reduction of olive fruit yield and oil content. Leaf spot disease is prevalent in the Sichuan and Hubei provinces of China, where high rainfall from August to September leads to easy infection by pathogens that caused the withering and falling of leaves, resulting in decreased production and flowering in the next year. Besides, in some Mediterranean regions, it has also been found that the infection of olive by pathogenic fungi inflicts serious economic losses on olive-related industries (López-Escudero and Mercado-Blanco, 2011). These all indicate that improving olive resistance is important for the olive industry and is one of the most important aspects of olive breeding. Thus, finding a new germplasm resource with functional genes to adapt to the climate of East Asia to cultivate new olive varieties with resistance through hybridization and with the existing olive varieties is an important way to promote the development of the olive industry.

Up to now, three versions olive reference assembly have been released, including two olive cultivars of *Olea europaea* L. subsp. *europaea* var. *europaea* cv. 'Farga' (Cruz et al., 2016) and *Olea europaea* L. subsp. *europaea* cv. '*Arbequina*' (Rao et al., 2021), and one oleaster of *Olea europaea* L. *sylvestris* (Unver et al., 2017), which generated genomes of 1.31 G, 1.30 G, and 1.48 G, with contig N50 values of 52.35 kb, 4.67 Mb, and 25.49 kb, respectively. Among these assembly versions, the contigs of "*Arbequina*" are almost completely anchored into 23 chromosomes by Hi-C which improved the olive genome assembly. All three samples belong to the Mediterranean climate zone. In fact, the olives are divided into six subspecies, including subsp. *europaea* (unique to the Mediterranean basin), subsp. *cuspidata*, subsp. *maroccana*, subsp. *laperrinei*, subsp. *cerasiformis* and subsp. *Guanchica* (Hannachi et al., 2009). Among them, subsp. *cuspidata* is known as native to a widespread area in southeast Europe and northeast Africa through southwest Asia to the Nujiang River Basin of Yunnan province in China (Green, 2002). Compared with subsp. *europaea*, it has a smaller fruit size and lower oil content but has better disease resistance and soil adaptability in the East Asian climate. Thus, subsp. *cuspidata* has been widely introduced to olive cultivation areas in China and is used as rootstock or a hybrid male parent to improve olive adaptability (Ye et al., 1981). Previous research shows that using subsp. *cuspidata* as rootstock grafting olive has not only increased survival rate and growth rate but also enhanced the adaptability of olive (Shi et al., 1991). In addition, an olive progeny issued from the cross *Olea europaea* L. subsp. *europaea* cv. "*Frantoio*" × subsp. *cuspidata*, is significantly superior to the parental species both in soil adaptability and disease resistance (Ma et al., 2014). Hybridization between the subspecies *europaea* and *cuspidata* has also been documented in other countries (Besnard et al., 2001). So far, resistance studies on olives mainly focus on the breeding of resistance varieties, only few related studies on identifying resistance genes have been reported.

The excellent resistance of subsp. *cuspidata* to pathogens may be ecologically owing to its long-term adaptation to the high temperature and highly humid environment in the Yunnan province of China; thus, it is a very promising germplasm for investigating resistance genes that can be used to enhance vitality and the ability of olive to resist the invasion of pathogens. Assembling the genome of this subspecies and comparing it with that of subsp. *europaea* will facilitate the conservation and utilization of germplasm resources of this subspecies, as well as further uncover the molecular basis of adaptive evolution and oil synthesis mechanisms and improve its marker-assisted breeding, etc.

In this study, we applied a combined strategy involving PacBio HiFi sequencing and Hi-C technologies to generate a chromosome-level assembly and then performed the population dynamics analysis, phylogenetic relationships, gene family expansion and contraction, whole-genome replication, unique genes analysis, positive selection, and transcriptome analysis. We found some positive selection genes were correlated with the term of response to stimulus, suggesting the relevant genes were under selection pressure after species differentiation that may be related to the environmental adaptation of subsp. *cuspidata*. We used specific genes to perform GO analysis and found some biological processes associated with oil synthesis. We also sampled the infected and healthy leaves of two cultivars to perform transcriptomic analysis and identified the *Cercospora* genus that may be causing leaf spot disease on the infected leaves of the two olive cultivars. The genes associated with resistance were identified in subsp. *cuspidata*, which can be an instance to investigate the genes against leaf spot disease between subsp. *europaea* and *cuspidata*, and is of great significance for improving the resistance of olives in the future. Given the significant differences between subsp. *cuspidata* and *europaea* in resistance and oil

content, the chromosomal genome assembly constructed here is greatly conducive to the research of oil production and resistance mechanisms, which is instructive to the molecular breeding, phylogenetic, adaptability, and evolutionary biology research of olives.

MATERIALS AND METHODS

Plant Materials

We sampled subsp. *cuspidata* individuals from the Yunnan province of China. A voucher specimen was deposited in the herbarium of the Forestry Research Institute of the Chinese Academy of Forestry. Young leaves were used for Illumina sequencing, PacBio HiFi sequencing, and the construction of Hi-C libraries. Four different tissues (stem, root, leaf, and fruit) were collected for RNA-seq analysis in order to assist genome assembly and annotation. In addition, we collected both the infected and healthy leaves of two cultivars (including "*Arbequina*" and "*Arbosana*") in the olive plantation in the Hubei province of China. Three replicates of infected and healthy leaves were separately taken for each cultivar and were used for RNA extraction and transcriptome analysis. The construction of the Hi-C libraries was provided by Novogene Co., Ltd. while other sequencing services were provided by Berry Genomics Co., Ltd. (Beijing, China).

Genome Sequencing and Transcriptome Sequencing

Short-insert-size (∼350 bp) libraries were constructed according to Illumina's standard protocol and paired-end reads (2 × 150 bp) were sequenced using an Illumina HiSeq X Ten platform (Illumina Inc., San Diego, CA, United States). A 60 Kb DNA SMRTbell library was constructed and a circular consensus sequencing (CCS) was performed on the PacBio platform (HiFi) (Pacific Biosciences Inc., Menlo Park, CA, United States). Hi-C libraries (two-cell) were constructed with the restriction endonuclease DPNll and sequenced on the Illumina HiSeq X Ten platform.

PacBio HiFi long reads were used as a backbone scaffold in genome assembly using hifiasm (version 0.14-r312) that provides better assemblies than other available tools (Cheng et al., 2021). The Illumina short reads were used to investigate the genome characteristics (such as genome size and heterozygosity) before assembly and for assembly quality evaluation. The Hi-C reads were used to anchor the contig-level assembly into the final chromosome-level genome assembly (Burton et al., 2013). To obtain the uniquely mapped read pairs, the raw data were aligned with the assembled genome using BWA-MEM (version 0.7.17-r1188) (Li and Durbin, 2009). The valid Hi-C data were evaluated using HiC-Pro based on uniquely mapped read pairs (Servant et al., 2015).

All of the RNA-seq libraries were constructed using a VAHTS mRNA-seq v2 Library Prep Kit with an average insert fragment size of ∼350 bp, and sequenced on an Illumina Novaseq 6000 platform with a paired-end model.

Quality Control of Sequencing Data

All sequencing data were filtered to eliminate low-quality bases and duplicated reads using different strategies based on the platforms used. For the Illumina Hi-Seq data, including genomic short-reads and RNA-seq reads, the PCR duplications of read pairs generated during the library construction process were first deleted. Then, adaptors were removed from the sequencing reads, and read pairs with more than 20% low-quality bases were deleted using Trimmomatic v0.33 (Bolger et al., 2014). If any read had more than 10% unknown bases, the read pair was excluded from further analysis (Chen et al., 2019). For Hi-C sequencing data, the same method used for Illumina Hi-Seq short-insert reads was adopted for filtering and then 3D was used for additional filtering. For PacBio HiFi long reads, subreads were directly filtered and corrected by the pbccs pipeline with default parameters[1].

Estimation of the Genome Size and Heterozygosity

Prior to the HiFi reads library-building sequencing, the investigation of the genome size and heterozygosity of subsp. *cuspidata* was carried out. The quality-filtered short fragments from the Illumina data were subjected to 21-mer frequency distribution analysis using Jellyfish v.2.2.10 (Marçais and Kingsford, 2011). We then performed genome analysis using GenomeScope2[2] based on the results of Jellyfish. Ultimately, we obtained the genome information of subsp. *cuspidata* (**Supplementary Figure 1**), including genome size, heterozygosity, and repetitive sequence proportions.

Genome Assembly

After filtering and correcting, the resulted HiFi CCS reads were subjected to hifiasm (v0.14-r312) for *de novo* assembly with default parameters[3], and the redundant haplotigs were removed using Purge Haplotigs (Roach et al., 2018). The haploid contigs were scaffolded using the 3D *de novo* assembly (3D-DNA) software (Dudchenko et al., 2017). Briefly, the Hi-C reads were aligned to the draft genome assembly using Juicer; a 3D-DNA analysis was conducted to generate a candidate assembly; the candidate assembly was reviewed using Juicebox v1.9.8 Assembly Tools (JBAT) (Durand et al., 2016), and then corrected artificially on the basis of candidate assembly. Benchmarking Universal Single-Copy Orthologs (BUSCO) (v3.0.2) (Simão et al., 2015) program with eudicotyledons_odb10 database was used to assess the completeness of the genome and gene annotation. Furthermore, the filtered short reads generated from Illumina and the assembled transcripts were mapped against our assembly using BWA-MEM algorithm and HISAT2 (v2.1.0) (Kim et al., 2015), respectively.

Repetitive Element Annotations

We employed the EDTA genome annotation pipeline (Ou et al., 2019) to annotate transposable elements (TEs) in the

[1]https://github.com/PacificBiosciences/ccs

[2]http://qb.cshl.edu/genomescope/

[3]https://github.com/chhylp123/hifiasm

subsp. *cuspidata* genome, including retrotransposons and DNA transposons, in which long tandem repeats (LTRs) and long interspersed nuclear elements (LINEs) belonged to the former, while terminal inverted repeats (TIRs) and helitrons belonged to the latter, and were detected by RepeatModeler. A *de novo* repeat library was produced to identify repeat sequences using RepeatMasker (v4.0.7) (Tempel, 2012) and Repbase (Bao et al., 2015) according to the recommended parameter values.

Gene Prediction and Functional Annotations

We mapped the RNA-seq data from the roots, stems, leaves, and fruits to the genome for predicting genes using the HISAT2 (v2.1.0) - StringTie (v1.3.5) pipeline and assembled the transcripts *de novo* by Trinity (Grabherr et al., 2011). Then, these transcripts were used to create transcript-based predictions with the PASA (v2.4.1) pipeline (Haas et al., 2003). The coding regions of the transcripts were annotated using a Transdecoder[4]. We also carried out homolog predictions. In such predictions, the protein sequences of *O. europaea* var. *sylvestris*, "*Arbequina*," *Juglans regia*, *Sesamum indicum*, *Solanum tuberosum,* and *Vitis vinifera* species were mapped to the genome using Exonerate v2.2.0. GlimmerHMM (v3.0.4) (Majoros et al., 2004). SNAP (Johnson et al., 2008) and AUGUSTUS (v3.3.3) (Stanke et al., 2006) were trained with genes from the PASA results and used for *de novo* gene prediction. We merged the gene models from these sources using EVidenceModeler (v1.1.1) (Haas et al., 2008). To find functional clues for the protein-coding genes of subsp. *cuspidata*, the predicted protein sequences were compared with those in several public databases [GO, EuKaryotic Orthologous Groups (KOG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, Pfam databases, and Nr databases].

Phylogenetic and Gene Family Analysis

Except for subsp. *cuspidata*, we chose one olive cultivar ("*Arbequina*") and one oleaster (*O. europaea* var. *sylvestris*). In addition, we selected another 11 plant relative species, including *S. indicum*, *S. tuberosum*, *Eucalyptus grandis*, *Glycine max*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Jatropha curcas*, *V. vinifera*, *Pistacia vera*, *Helianthus annuus*, and *Oryza sativa*, with *Oryza sativa* as outgroup. The protein sequences of all these species were downloaded from the NCBI. We first filtered these protein sequences with lengths of less than 100 bp to improve the alignment quality. OrthoFinder (v2.5.2) (Emms and Kelly, 2019) was then used to identify single-copy homologous genes and classify the protein sequences into families of 14 species with the key parameters "-M msa -S diamond -T raxml-ng," where -M is the method for gene tree inference, -S is the alignment method, and -T is the tree inference method used. We inferred the phylogenetic relationship tree among 14 species and assessed the branch support with 100 bootstrap replicates using RAxML (Stamatakis, 2014). The divergence time was calculated using MCMCtree from the PAML package (Yang, 2007). In addition, the known divergence time between *P. trichocarpa* and *J. curcas* (77 Mya, CI:70–86 Mya) from the public resource TIMETREE[5] was provided as calibration points in the analysis.

CAFE (v3.1) was used to analyze the expansion and contraction of the gene families (Han et al., 2013). We obtained the evolutionary tree and gene family clustering that were used to estimate the number of gene families of the ancestors in each phylogenetic tree branch, thereby predicting gene family contraction and expansion. The gene families with particularly large gene copy number variation were eliminated to decline parameter prediction errors using python script *cafetutorial_clade_and_size_filter.py*. The specific information of expansion and contraction gene families for the 14 species were finally obtained by applying the script *cafetutorial_report_analysis.py*, with these results used for later analyses. In addition, we uploaded the obtained gene family information to the OrthoVenn2 website for analysis[6]. Based on the gene families specific to subsp. *cuspidata* and "*Arbequina*" obtained from the above steps, we performed a functional enrichment analysis of GO terms using Fisher's exact test[7] to determine if any functional gene classes were overexpressed.

Positive Selection Analysis

By comparing the protein sequences of subsp. *cuspidata* and "*Arbequina*," we performed positive selection analysis using CODEML module in PAML, which can reveal the direction and strength of natural selection acting on the protein by estimating the non-synonymous and synonymous rates (d_N and d_S) between two protein sequences and infer the positive selection of protein-coding genes. Prior to the CODEML program, the coding sequence of "*Arbequina*" with a length greater than 100 bp was first used to create a BLAST database using Makeblastdb, and then the protein sequence of subsp. *cuspidata* was used to align to the database for a screening of orthologous genes between the two species using Blastp with the e value of 1e-5. After obtaining the file with a.homolog suffix that included all of the co-orthologs, the name of the two-way optimal paired sequence was obtained with ParaAT, which is the input format of PAML. The synonymous and non-synonymous substitution rates and positive selection in sequences were estimated and detected using CODEML, and some of the variables within the control file were configured before the CODEML run. We set "icode = 0" to specify the universal genetic code, furthermore, we set "fix_omega = 0" and "fix_kappa = 0" to ensure that the parameters of the ω and the transition/transversion ratio were estimated separately *via* maximum likelihood. Since a comparison is made between the two subspecies, we only need to set the null model to find the gene with an omega ($\omega = d_N/d_S$) value greater than 1, representing positive selection.

Whole-Genome Duplication and Synteny Analysis

Oleaster, subsp. *cuspidate* and "*Arbequina*" were selected to perform whole-genome duplication (WGD) analysis

[4]https://github.com/TransDecoder/TransDecoder

[5]http://www.timetree.org

[6]https://orthovenn2.bioinfotoolkits.net/cluster-venn

[7]https://www.omicshare.com/tools/Home/Soft/gogsea

by calculating fourfold synonymous (degenerative) third-codon transversion (4DTv) values and distributions of synonymous substitutions per synonymous site (Ks) within and between each species. The 4DTv rates of collinear gene pairs were calculated based on fourfold degenerate sites following the YN substitution model. *K*s values of the collinear orthologous gene pairs were calculated using KaKs_Calculator (v2.0) (Wang et al., 2010) with default parameters. The CIRCOS module of the TBtools (Chen et al., 2020) software was used to visualize the assembled chromosomes of the genome, gene density, GC content, repeat content, and gene synteny on individual pseudochromosomes. The nucmer (4.0.0beta2) program in MUMmer4 (Marais et al., 2018) was used to determine whether similar gene pairs were adjacent on the chromosome between subsp. *cuspidata* and "*Arbequina*," ultimately obtaining all the genes in the synteny block.

Demographic History Reconstruction

To estimate the population size history and split time of subsp. *cuspidata* and "*Arbequina*," we utilized the resequencing date from one subsp. *cuspidata* and one "*Arbequina*" individual to perform SMC++ (Terhorst et al., 2016), which is capable of analyzing unphased genomes. The sequencing data of subsp. *cuspidata* were obtained from the genome survey analysis data in this study, and the sequencing data of "*Arbequina*" were downloaded from the Genome Warehouse in the National Genomics Data Center (NGDC) with the BioProject accession number PRJCA003222. We first estimated each population marginally using an estimate. Then, we created datasets containing the joint frequency spectrum for both populations. Finally, we refined the marginal estimates into an estimate of the joint demography using split. A generation time of 20 years (Diez et al., 2015) and a mutation rate of 7.77e-09 mutations per nucleotide per generation (Xie et al., 2016; Julca et al., 2020) were used to convert the scaled times and population sizes into real times and sizes.

Identification of the Fungal Category

In order to identify the fungal species that caused the leaf spot of the two cultivars' infected leaves, the unmapped reads of all infected leaves in "*Arbequina*" and "*Arbosana*" were extracted to perform *de novo* genome assembly. The clean Fastq data of infected leaves were first mapped to the assembled genome and olive chloroplast and mitochondrial sequences with HISAT2. The unmapped reads were then extracted using samtools with the key parameters "-b -h -f 4," and performed *de novo* assembly using Trinity (v2.1.1). After this, we downloaded the Nr database from NCBI and extracted the fungi subset using TaxonKit with the parameter of "-j 8 –ids 4751," in which, "–ids 4751" represents the subset of fungi. The subset was used to create a BLAST database using Makeblastdb, and then the assembled sequences were aligned to the fungi database using Blastp with the e value of 1e-5.

TABLE 1 | Statistics of assembled subsp. *cuspidata* genome.

Term	Contig size (bp)	Contig number	Scaffold size (bp)	Scaffold number
N90	350,652	257	885,961	52
N80	2,313,398	94	34,033,801	21
N70	6,370,930	55	40,664,807	17
N60	11,521,701	40	44,152,604	14
N50	14,716,965	30	52,676,021	11
Max length (bp)	38,043,138		90,127,509	
Total size (bp)	1,379,115,243		1,379,304,243	
Total number	3,073		2,695	
Average length	448,784.65		511,801.20	

Differential Gene Expression Analysis

"*Arbequina*" and "*Arbosana*" are the most widely cultivated in plantations due to their high production (Centeno et al., 2019). We thus collected the infected and healthy leaves from these two cultivars in September for differential gene expression analysis, because olives were susceptible at this time. The transcriptome clean Fastq data from infected and healthy leaves were mapped to the assembled genome with HISAT2. The alignments were used for transcript assembly using StringTie, which assembles the genes for each data set separately and estimates the expression levels of each gene and isoform. All the gene structures found in any of the samples were merged together with the key parameter of "stringtie -merge," and then, all the transcripts and abundances were obtained using Ballgown (Pertea et al., 2016). The result of transcript quantification obtained from Ballgown was converted to the count matrices of genes and transcripts with the command of "python2 prepDE.py -i ballgown," in which the script *prepDE.py* was downloaded from http://ccb.jhu.edu/software/stringtie/dl/prepDE.py. Finally, differential gene analysis was performed with the count data using DEseq2 package in R, which provides methods to test for differential expression by using negative binomial generalized linear models (Love et al., 2014). We separately grouped these data into two groups of healthy and infected leaves for each cultivar and screened differentially expressed genes (DEGs) using DESeq2 with an adjusted p-value < 0.05 and the absolute value of a log2(FC) > 1 (Love et al., 2014), which were also used for GO analysis. Differential expression genes were further classified as upregulated and downregulated based on their log fold change (FC) values. Genes with an FC value greater than zero were considered upregulated, while those with less than zero were thought to be downregulated. Further, we calculated the FPKM (Fragments per Kilobase Million) values using Ballgown to validate the expression of each gene in infected and healthy leaves of two cultivars. Genes were considered low expressed if they had an FPKM value between.125 and 1, medium expressed if they had a value between 1 and 10, and highly expressed if the value was above 10 (Hackett et al., 2012). We also sampled three replicates of healthy leaves of subsp. *cuspidata* to compute the FPKM values to understand the expression of differential genes in subsp. *cuspidata*.

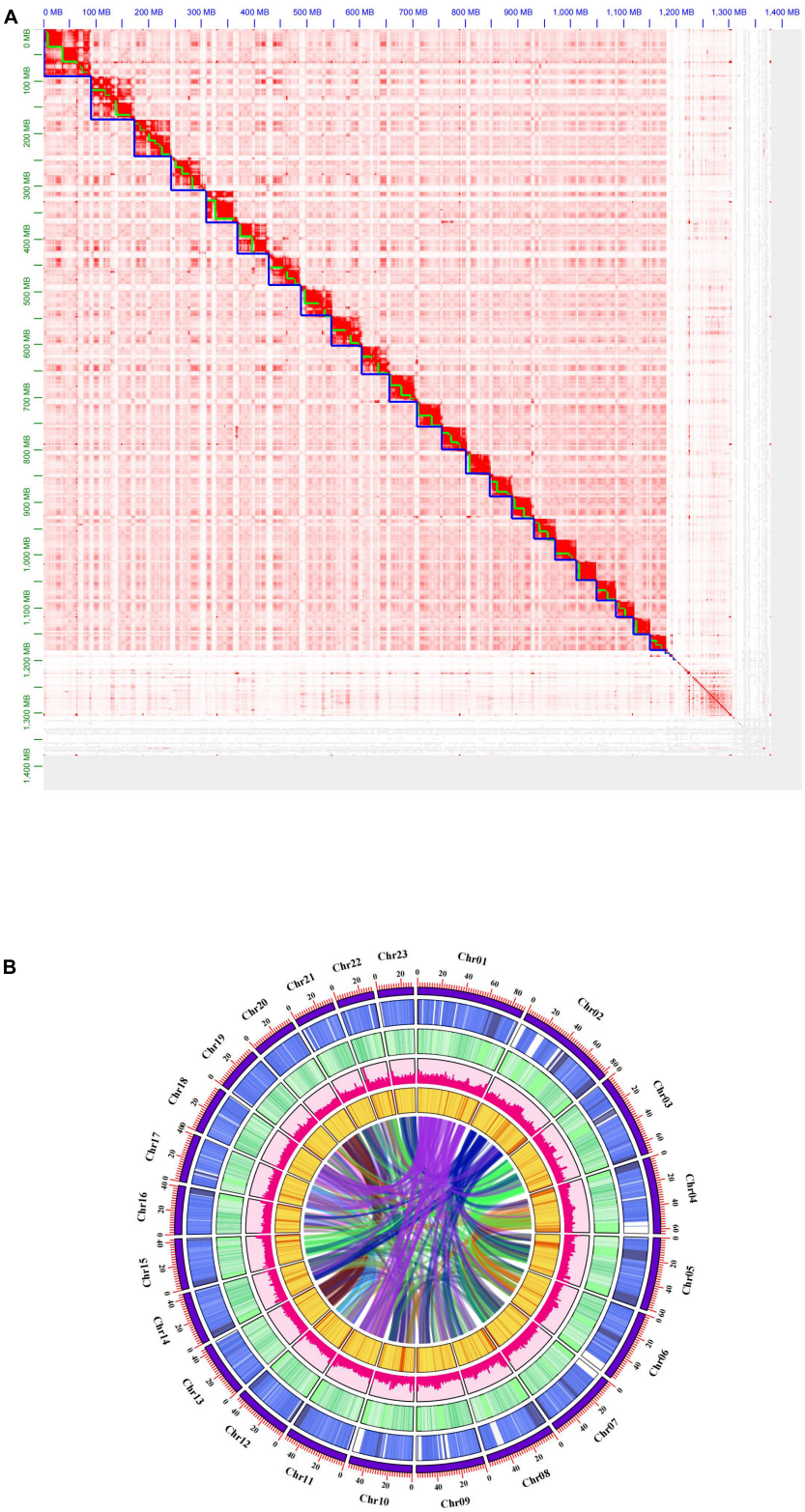


FIGURE 1 | Genome-wide Hi-C interaction heatmap and Genomic landscape. **(A)** Hi-C interaction heat map between 23 chromosomes for the subsp. *cuspidata* genome. **(B)** Genomic landscape of subsp. *cuspidata* chromosomes. Visualize the genome assembly chromosome, gene density, GC content, repeat content, SNP density, and gene collinearity on a single pseudochromosome from the outer ring to the inside.

TABLE 2 | Statistics of chromosomal level assembly of subsp. *cuspidata*.

Chr ID	Length (bp)	Chr ID	Length (bp)	Chr ID	Length (bp)
Chr1	90,127,509	Chr9	57,282,915	Chr17	40,664,807
Chr2	83,097,257	Chr10	52,971,700	Chr18	39,899,167
Chr3	70,287,963	Chr11	52,676,021	Chr19	37,263,953
Chr4	64,129,678	Chr12	47,592,757	Chr20	37,211,276
Chr5	61,350,988	Chr13	45,546,967	Chr21	34,033,801
Chr6	59,983,315	Chr14	44,152,604	Chr22	31,166,573
Chr7	58,685,853	Chr15	42,848,148	Chr23	29,903,841
Chr8	58,506,042	Chr16	40,951,526		
Total chromosome level contig length			1,180,334,661		
Total contig length			1,379,304,243		
Chromosome length/Total length			85.57%		

TABLE 3 | Completeness assessment of subsp. *cuspidata* genome by BUSCO.

Library	eudicotyledons_odb10
Complete BUSCOs (C)	2048
Complete and single-copy BUSCOs (S)	1717
Complete and duplicated BUSCOs (D)	331
Fragmented BUSCOs (F)	24
N50Missing BUSCOs (M)	49
Total BUSCO groups searched	2121
Summary (Complete BUSCOs/Total BUSCOs)	96.6%

TABLE 4 | Statistics of TE annotated repeat sequences in subsp. *cuspidata* genome.

Class	Sub-Class	Type	Length (bp)	Percent (%)
Retrotransposons	**LTR**	Ty1/Copia	137,408,274	9.96%
		Ty3/Gypsy	205,089,955	14.87%
		unknown	63,995,280	4.64%
	Non-LTR	LINE	1,905,667	0.14%
		unknown	423,876	0.03%
DNA transposons	**TIR**	CACTA	18,631,143	1.35%
		Mutator	347,273,079	25.18%
		PIF/Harbinger	19,703,347	1.43%
		Tc1/Mariner	2,351,832	0.17%
		hAT	28,621,867	2.08%
	Non-TIR	helitron	48,941,818	3.55%
	Total		960,043,533	69.61%

RESULTS

De novo Assembly of the subsp. *cuspidata* Genome

We obtained ∼253.5 Gb clean Fastq data for the Illumina short reads. To resolve any difficulties that may arise during the genome assembly process, the Kmer-based method was used to perform genome survey analysis to estimate the genome size and heterozygosity of the subsp. *cuspidata* genome using Illumina short reads. We counted the number of each 21-mer with Jellyfish, and the frequency distribution was plotted in **Supplementary Figure 1**. The subsp. *cuspidata* genome size was then estimated to be 1.18 Gb with 0.36% heterozygosity, and the coverage is ∼34.7-fold relative to the actual assembly results. To obtain a high-quality genome assembly, a total of ∼44.72 Gb of PacBio HiFi long reads (reads: 3,294,182, average N50: ∼14.85 Kb) were generated and subjected to hifiasm for *de novo* genome assembly. After assembly and deduplication, the consensus sequences resulted in a contig level assembly of 1.38 Gb spanning 3,073 contigs, with a contig N50 of 14.7 Mb and the longest contig of 38.04 Mb (**Table 1**). We obtained ∼450 Gb of Hi-C Fastq clean data with the effect rate of 34.61%, and used it for chromosome construction using 3D *de novo* assembly. A total of 1.18 Gb sequences spanning 2,695 scaffolds were finally anchored onto 23 pseudochromosomes (**Figure 1**), with a scaffold N50 of 52.68 Mb and the longest scaffold of 90.13 Mb (**Table 1**). The mounting rate was 85.57% (**Table 2**), and the average GC content was 0.36. The BUSCO results showed that more than 2,048 (96.6%) genes were completely recalled, of which 81% were single-copy and 15.6% originated from duplication (**Table 3**). A total of 879,715 transcripts were acquired, with an average of 96.52% reads located in the assembled genome (**Supplementary Table 1**). The mapping rate of resequencing reads exceeded 92.31% of the whole genome.

Repetitive Sequences, Gene Prediction, and Functional Annotations

We annotated all repetitive sequences to further characterize the genome of subsp. *cuspidata* by integrating *de novo* and homology-based approaches. We predicted 69.61% of the genome as transposable elements. DNA transposons were the most abundant characterized elements, in which, TIRs accounted for 30.2% and non-TIRs accounted for 3.6%. In retrotransposons, LTRs accounted for 29.5% and non-LTRs accounted for 0.17% (**Table 4**).

A total of 46,904 protein-coding genes were predicted in the current assembly, and then we implemented the gene function annotation using GO, KEGG, KOG, SwissProt, Pfam annotation, and Nr annotation databases. From this analysis, most of the predicted genes were functionally annotated in these databases (**Table 5**).

Genome Evolution, Phylogeny, and Synteny Analysis

A total of 65,396 gene families were obtained in all species, namely, subsp. *cuspidata*, "*Arbequina*," var. *sylvestris*, *S. indicum*, *S. tuberosum*, *E. grandis*, *G. max*, *A. thaliana*, *P. trichocarpa*,

TABLE 5 | Statistics of functional annotation of protein-coding genes in subsp. *cuspidata* genome.

Database	Annotated gene number	Percent (%)
GO	26,012	57.60
KEGG	8,327	18.44
KOG	8,941	19.80
SwissProt	33,018	73.12
Pfam annotation	32,739	72.50
Nr annotation	45,146	99.98

J. curcas, *V. vinifera*, *P. vera*, *H. annuus*, *O. sativa*. We reconstructed a phylogenetic tree based on a concatenated sequence alignment of all single-copy genes which are shared by these species and estimated their divergence time. All the relationships were well supported with > 90% bootstrap values (**Figure 2**). As expected, oleaster and "*Arbequina*" were grouped together, and the splice time between them occurred approximately 3.48 (1.94, 5.14) million years ago (Mya), subsp. *cuspidata* diverged from them about 6.5 (4.21, 9.29) Mya, while olive diverged from *S. indicum* about 61.54 (41.02, 81.44) Mya.

The population demographic history inferred with SMC++ software showed evidence for a considerable and continuous decline in both population sizes. The population of subsp. *cuspidata* started approximately 13 Mya (**Figure 3**), closing to the high central plateau of the Qinghai-Tibet Plateau timeframe (∼10–13 Mya) (Zhang et al., 2010). The splice time between subsp. *cuspidata* and "*Arbequina*" was approximately 5.5 Mya, which was generally consistent with the timing of the phylogenetic tree.

Whole-genome duplication (WGD) is seen as an important factor with a significant effect on plant genome evolution (Mcgrath and Lynch, 2012). To further understand the genomic evolution of subsp. *cuspidata*, "*Arbequina*" and oleaster, we performed WGD analysis; the collinearity of inter- and intra-olive genomes provided evidence of these three species' WGD events (**Figure 4**). By determining the distribution of 4DTv and *Ks* values, we detected one main peak within subsp. *cuspidata* (the peak of 4DTV: ∼0.092, *Ks*: ∼0.389), "*Arbequina*" (4Dtv: 0.091, *Ks*: ∼0.271), and oleaster (4Dtv: ∼0.085, *Ks*: ∼0.221), indicating that all three species had experienced one WGD event, which was similar to the result of previous research (Rao et al., 2021). Following that, species divergence occurred. The divergence of subsp. *cuspidata* - oleaster occurred at a peak of *Ks* ∼0.137, followed by subsp. *cuspidata* – "*Arbequina*" (*Ks*, ∼0.135) and "*Arbequina*" - oleaster (*Ks*, ∼0.013) divergence.

Synteny analysis revealed a high linear relationship between subsp. *cuspidata* and "*Arbequina*." A total of 43,711 genes in subsp. *cuspidata* were found to have synteny with "*Arbequina*." The synteny between chromosomes was partially dislocated (**Figure 5**), which may have been caused by two reasons: First, the "*Arbequina*" adopted the sequencing technology of Oxford Nanopore, whose error rate was as high as ∼40%, much higher than PacBio HIFI (lower than 1%) (Laver et al., 2015; Ye and Ma, 2016). Second, the genome of "*Arbequina*" was assembled by merging the results of the three different software (including Canu, Wtdgb, and SMARTdenovo), which may have introduced further errors.

Comparative Genomics Analysis

We compared six oil species that aimed to search for genes associated with oil production. A total of 10,813 gene families were shared by these six species, and 681 gene families were unique in subsp. *cuspidata*, 394 in "*Arbequina*," 656 in *O. europaea* var. *sylvestris*, 477 in *S. indicum*, 2,853 in *H. annuus*, and 1,932 in *G. max* (**Figure 6**). These specific gene families of subsp. *cuspidata* and "*Arbequina*" were then separately annotated to GO terms. In "*Arbequina*," unique genes were grouped into annotations of nine biological processes, eight cellular components, and nine molecular functions (**Supplementary Figure 2**). In the biological process group, we obtained 384 biological process descriptions, of which 51 were significantly expressed ($P < 0.05$) containing 81 genes (**Supplementary Table 2**). Interestingly, we found some significant expression processes associated with lipid biosynthetic, including the metabolic and/or catabolic process of *S*-glycoside and glycosinolate and the biosynthetic and metabolic process of the acetyl-CoA. Lipid is one of the major carbon storage compounds (Li et al., 2010), while glycogen is one of the major metabolites for carbon storage in many plants (Govindprasad et al., 2017). The acetyl-CoA is the most abundant short-chain acylCoA in olive fruit (Sanchez-Ortiz et al., 2012) and serves as a precursor for fatty acid synthesis (Salas et al., 2013; Priore et al., 2014). Thus, both glycogen and acetyl-CoA play an important role in fatty acid synthesis. This result suggests the important position of genes associated with oil synthesis in genes specific to "*Arbequin*." Similarly, in subsp. *cuspidata*, unique genes were grouped into annotations of 9 biological processes, 8 cellular components, and 10 molecular functions (**Supplementary Figure 3**). In the biological process group, 389 biological process descriptions were obtained, of which 132 were significant and contained 1,415 genes. We also found some significant expression of the progress related to lipid synthesis, such as the biosynthetic and/or metabolic process of glycosyl compound, carbohydrate derivative, aromatic compound, organic cyclic compound, cellular lipid, and trehalose (**Supplementary Table 3**).

We conducted a positive selection analysis between "*Arbequina*" and subsp. *cuspidata*. A total of 38,158 single copy orthologous genes were compared and 2,777 genes accounting for 7.28% were finally identified under positive selection ($d_N/d_S > 1$) in subsp. *cuspidata*. GO enrichment analyses show that these genes were categorized into 37 functional groups, including 17 biological processes, 9 cellular components, and 11 molecular function annotations (**Supplementary Figure 4**). Significantly, we found a term of response to stimulus (GO: 0050896) with 66 genes in biological process, such as response to water, response to inorganic substance, response to endogenous stimulus, response to biotic stimulus, defense response (**Supplementary Table 4**), suggesting the relevant genes were under selection pressure after species differentiation that may be related to the environmental adaptation of subsp. *cuspidata*.

Identification of the Fungal Genus

Compared with healthy leaves, the symptom of the infected leaves is pathogen-induced spot (**Figure 7**). To identify the fungal species, we extracted the unmapped sequences from all infected leaves for *de novo* genome assembly and aligned them to the constructed fungal library. We found the fungi in genus *Cercospora* presented in all six alignment results and with the highest identity, including *Cercospora beticola*, *Cercospora zeina,* and *Cercospora kikuchii*, they were well supported with > 40% identity (**Supplementary Tables 5**, **6**). As we expected, three fungi were causing foliar diseases. *Cercospora* is known to be one of the main groups of plant pathogenic fungi, which can cause necrotic

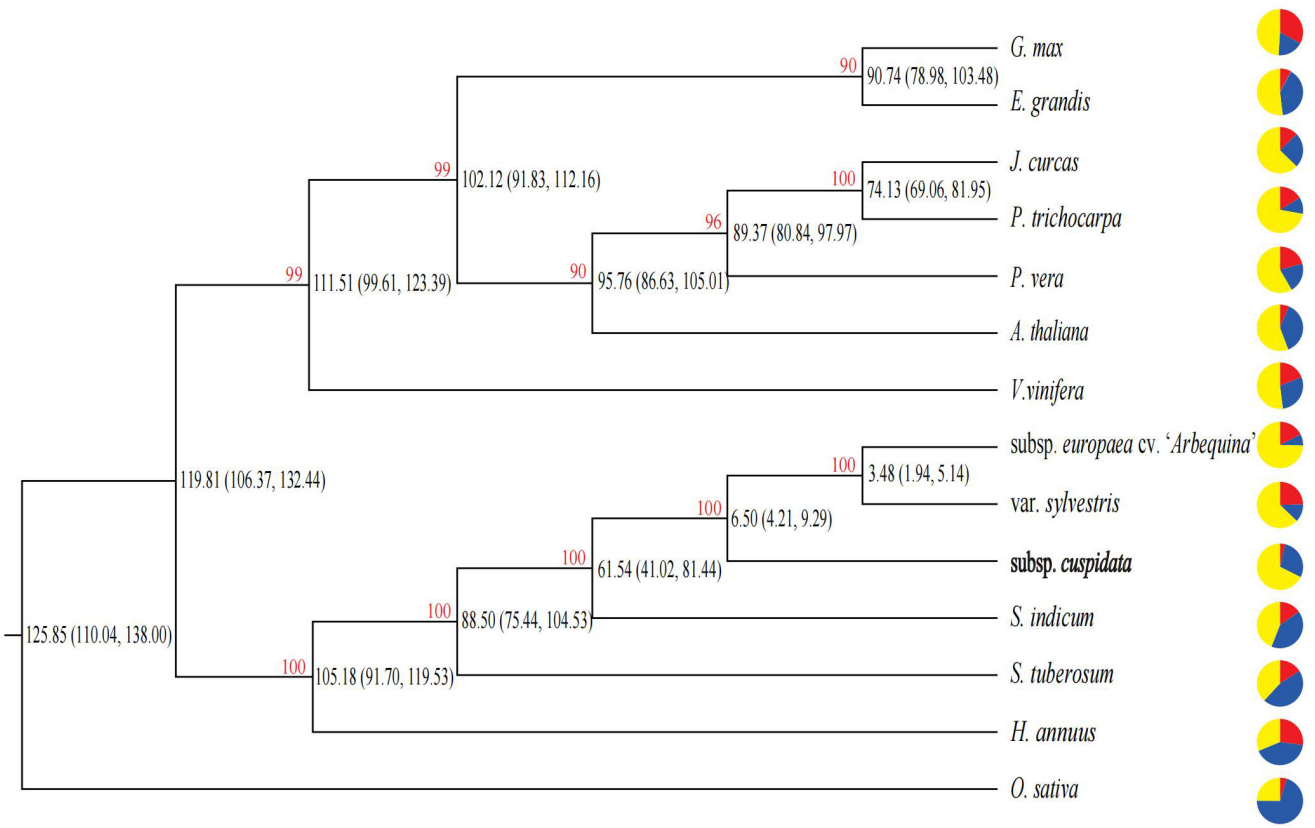


FIGURE 2 | Phylogenetic relationship and divergence time among species. Pie charts show the proportion of gene families that are expanded (red), contracted (blue), and conserved (yellow) across all gene families in the 14 species. The red number in each node represents the bootstrap value. The number in parentheses in each internal node indicates the estimated divergence time interval (in millions of years).

leaf spots in many plants (Groenewald et al., 2013). Since the symptom of leaf spot was also appeared in the infected leaves we collected, this result is largely reliable.

Differential Gene Analysis of Transcriptome

We performed differential gene analysis for the infected and healthy leaves of the two cultivars and obtained 248 and 475 DEGs. Among these, 117 and 172 genes were upregulated and 131 and 303 genes were downregulated in "*Arbequina*" and "*Arbosana*," respectively (**Supplementary Tables 7**, **8**). Moreover, 49 common genes were differentially expressed in two cultivars. To gain further insight into the function of the 49 genes in subsp. *cuspidata*, we performed GO enrichment analysis, categorizing the 49 DEGs into 19 functional groups, which included seven biological processes, seven cellular components, and five molecular function annotations (**Supplementary Figure 5**). Among them, we found only one gene (*evm.model.Chr16.1133*) with a term of response to stimulus in the biological process group; significantly, this gene also underwent positive selection. This gene sequence was then aligned to *A. thaliana* using Blastp with the e value of 1e-5, indicating *evm.model.Chr16.1133* is homologous to *AtMLO6* (AT1G61560), with a 61.78% identity. In our results, *evm.model.Chr16.1133* gene was moderately expressed in the healthy leaves of "*Arbequina*" and "*Arbosana*" (the mean FPKM value was 1.793 and 3.150 of three duplicates, respectively), but had a low expression in infected leaves of the two cultivars (the mean FPKM value was 0.558 and 0.818, respectively) (**Table 6**), and the log2(FC) value was separately –1.702 and –1.917 (**Supplementary Tables 7**, **8**), indicating the negative regulatory role of it against pathogens, which was in agreement with previous studies (Bai et al., 2008; Delventhal et al., 2011). Furthermore, this gene had also a low expression in subsp. *cuspidata* healthy leaves (mean FPKM: 0.583), implying that the low expression of this gene may be related to good resistance in subsp. *cuspidata*.

According to an underway olive-related study, the sequences of *evm.model.Chr16.1133* gene were separately obtained from 29 subsp. *cuspidata* and 25 olive cultivar individuals. We computed polymorphic sites, the values of Tajima's *D* and nucleotide polymorphism (θ_π) using DnaSP. We found no polymorphic sites of this gene in subsp. *cuspidata*, while the cultivars

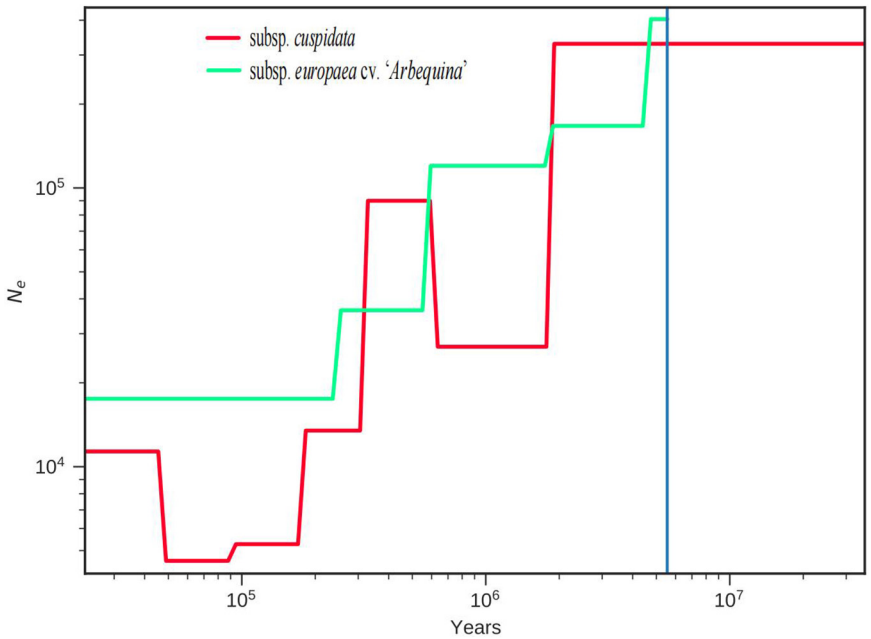


FIGURE 3 | Population history analysis of subsp. *cuspidata* and "*Arbequina*". SMC++ estimates the effective population size (*Ne*) changes for subsp. *cuspidata* and "*Arbequina*," and estimates the splice time between subsp. *cuspidata* and *Arbequina*".

showed a higher polymorphism with the Tajima's *D* value of 0.929 and θ_π value of 0.003 (**Table 7**), indicating that severe natural selection led to no polymorphism of this gene in subsp. *cuspidata*.

DISCUSSION

Phylogenetic Analysis

Olive is a world-renown tree species owing to its economic, ecological, cultural, and scientific values. The phylogenetic analysis showed that the ancestor of oleaster and "*Arbequina*" was a sister of subsp. *cuspidata*, and the divergence between them was approximately 6.5 (4.21, 9.29) Mya. SMC++ results showed the split time of 5.5 Mya between "*Arbequina*" and subsp. *cuspidata* and was similar to the phylogenetic analysis, which also showed a considerable decline in both population sizes, and the subsp. *cuspidata* population started approximately 13 Mya. These results were close to the formation of the high central plateau of the Qinghai Tibet Plateau (QTP) at 10–13 Mya. In the late Cretaceous period of approximately 60 Ma BP, however, most of the QTP was still in the ancient Mediterranean at that time. It had a hot tropical-subtropical climate and was a region where thermophilic plants developed and flourished at that time (Sun and Li, 2003), where *Canarium* was one of the common floras (Mai, 1989; Zheng, 1989). The retreat of the ancient Mediterranean and the uplift of the QTP changed the Asian climate system and promoted the formation of inland drought in Asia (Peng, 2013). Since subsp. *cuspidata* may be the remaining species of paleo-Mediterranean flora that originated from the ancient Mediterranean region, we thus speculated that the uplift of the QTP may have caused the differentiation between them, and potentially reduced subsp. *cuspidata* historic population sizes.

Determination of the Fungal Genus

The rainy summer climate in East Asia is conducive to the reproduction of pathogens, and the introduced olives are thus susceptible to pathogen invasion, leading to a decline in fruit production and even the trees' death. The fungi causing the leaf spot disease in the two olive cultivars were identified to be the *Cercospora* genus, which was known as one of the main groups of plant pathogenic fungi. *C. beticola* is a worldwide distributed fungal disease and severely destroys the leaves of *Beta vulgaris* L., causing leaf spots and further resulting in the reduction of production and sugar content (Shane, 1992). *C. zeina* is distributed in many countries, it causes gray leaf spot of maize and leads to the reduction of maize yield (Meisel et al., 2009). *C. kikuchii* occurs in all soybean producing regions around the world, it causes purple seed stain on seed pods and seeds, and leaf blight on leaves and petioles, which has seriously affected the quality of soybean (Takeshi and Tomohiro, 2021). These three fungi are associated with foliar diseases, which is consistent with the symptom of the infected leaves that we collected. We thus speculate that the *Cercospora* genus may be causing the leaf spot disease in the infected leaves of the two olive cultivars.

Identification of the Susceptibility Gene

Compared with olive cultivars, subsp. *cuspidata* has lower oil content but higher resistance to fungal diseases and abiotic stress (Hannachi et al., 2009; Trapero et al., 2015). Thus, it

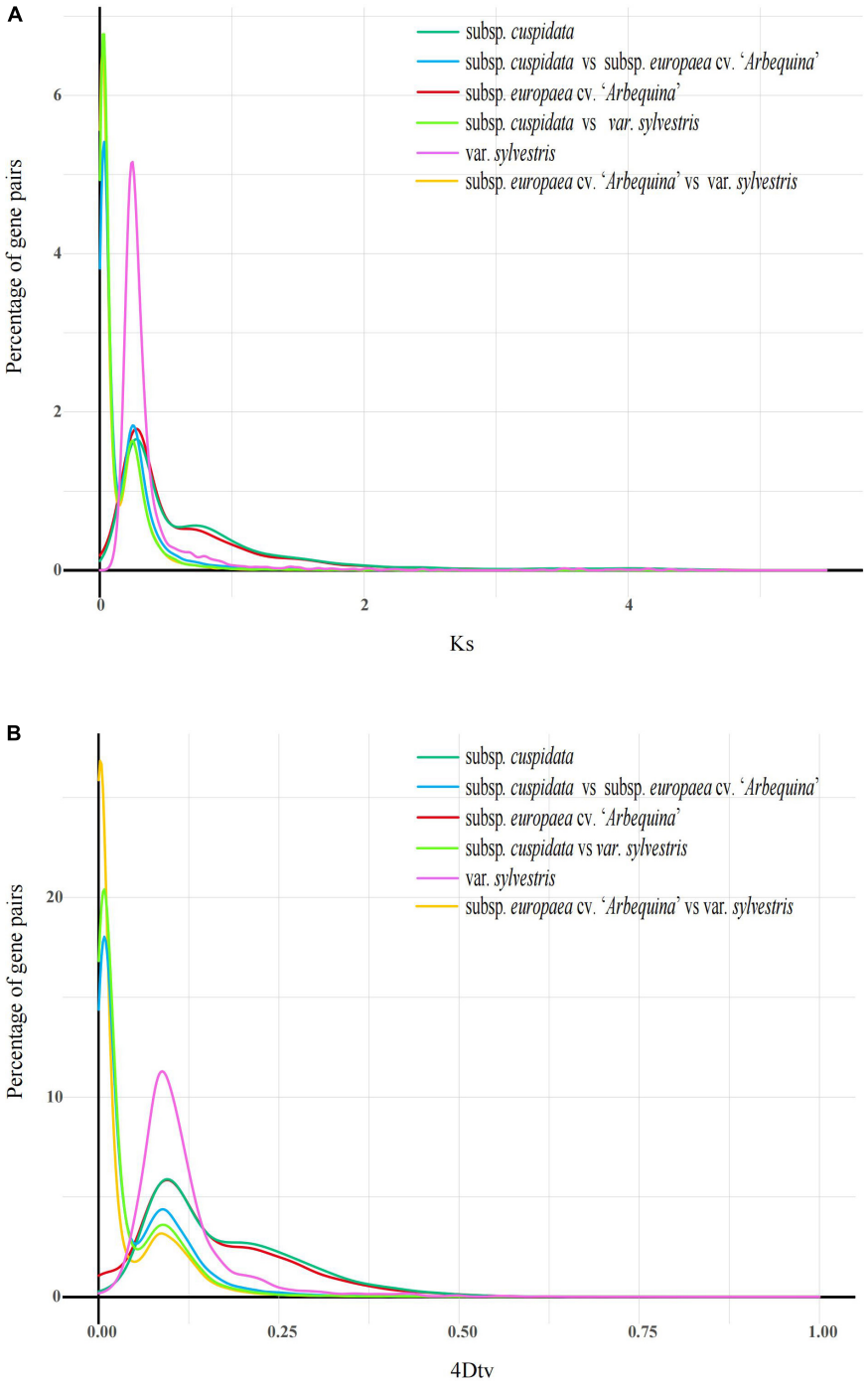


FIGURE 4 | Whole-genome duplication (WGD) analysis. **(A)** *K*s distributions analysis. Peaks of intraspecies *K*s distributions indicate whole genome polyploidization events, and peaks of interspecies *K*s distributions indicate speciation events. **(B)** The 4DTv distribution of gene pairs in subsp. *cuspidata* and other genomes. The *x*-coordinate is the 4DTv value, and the *y*-coordinate represents the proportion of genes corresponding to the 4DTv values.

is a valuable genetic resource to investigate the differences in oil content and resistance between subsp. *europaea* and *cuspidata*. Here, we performed a GO analysis for the positive selection of genes and found 66 genes belonging to the term of response to stimulus, indicating that the genes associated with environmental adaptation were under selection pressure in subsp. *cuspidata*. Interestingly, one of the positive selected genes, *evm.model.Chr16.1133*, belongs to the term of defense response (GO:0006952); it is also found in the results of transcriptome differential gene analysis. *AtMLO6* is the homolog of this gene in *A. thaliana*; it is a well-characterized susceptibility gene belonging to the mildew resistance locus O (MLO) gene family, which is a

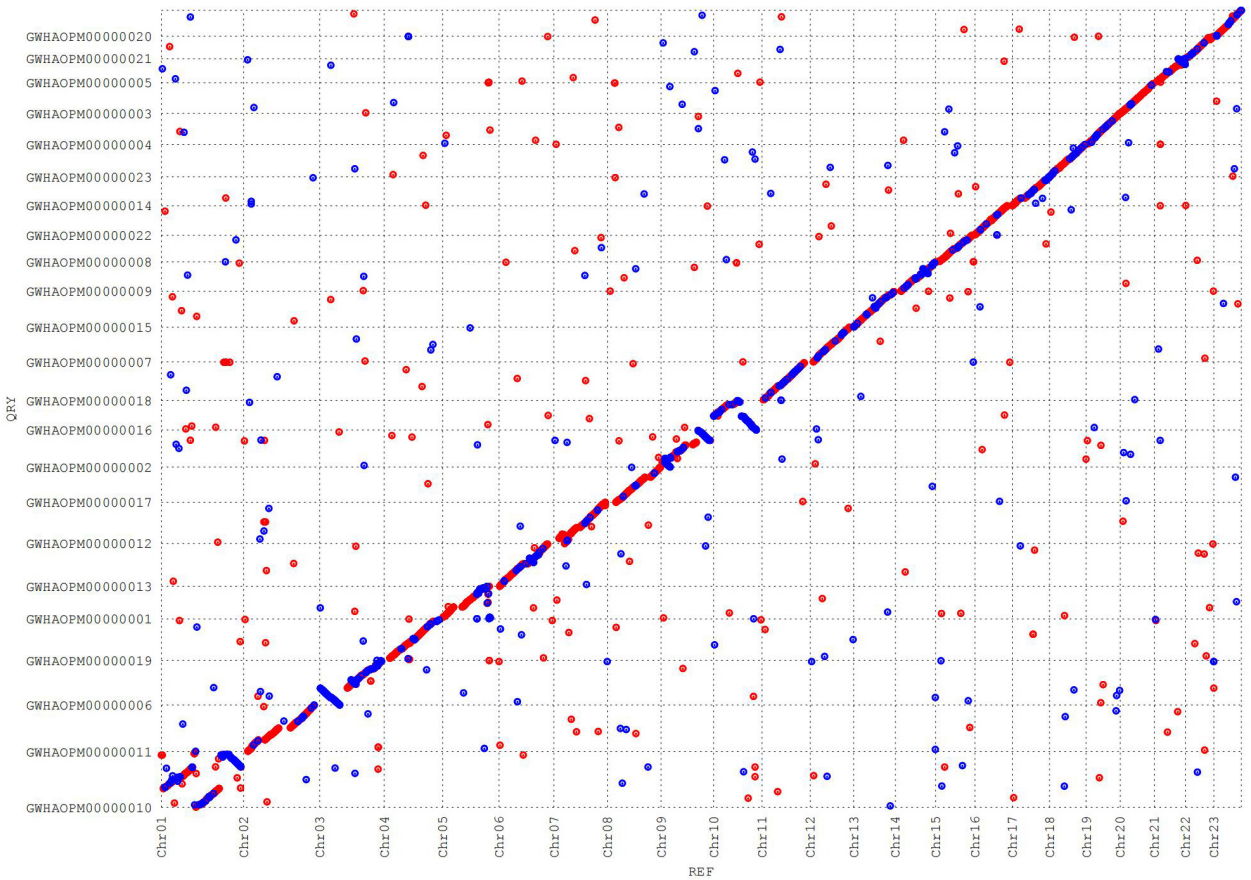


FIGURE 5 | Dot synteny diagram of chromosomes in subsp. *cuspidata* and "*Arbequina*".

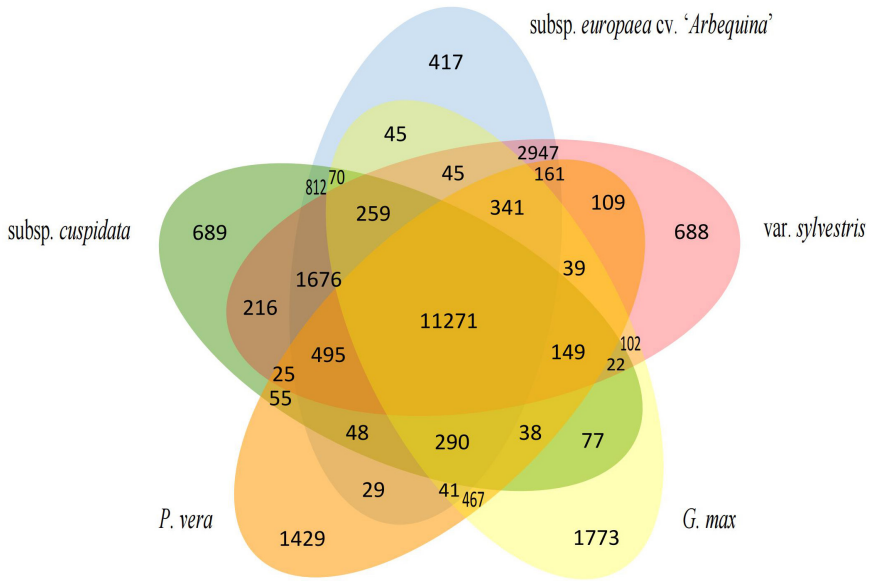


FIGURE 6 | Petal diagram of the gene families for six oil species. The middle number represents the gene families shared by all species and the number of gene families unique to each species is on the side.

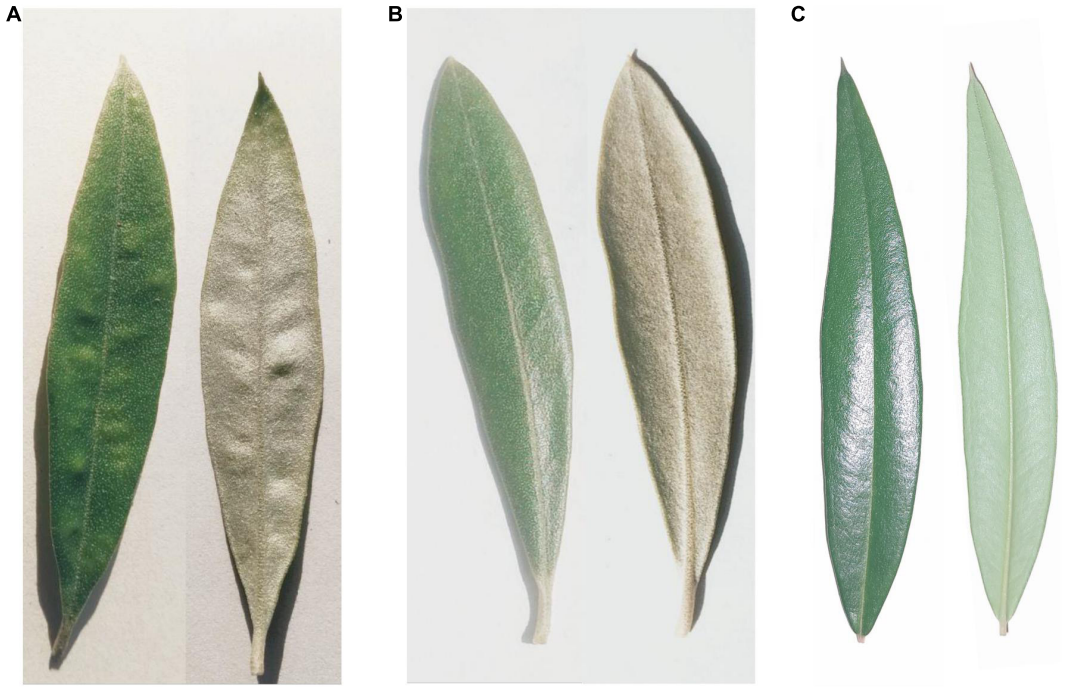


FIGURE 7 | The leaves of the three subspecies. The symptom of infected leaves is identical in "*Arbequina*" and "*Arbosana*". **(A,B)** Represents the front and back of the infected and healthy leaves, respectively. **(C)** Represents the front and back of the healthy leaves of subsp. *cuspidata*.

class of single-gene controlled recessive disease resistance genes that negatively regulates the disease resistance process and leaf cell death in plants (Buschges et al., 1997; Piffanelli et al., 2002). The *MLO* gene was originally found in barley and also found in some plants, such as *A. thaliana* (Vogel et al., 2006), *Rosa multiflora* (Xiang et al., 2018), *Pisum sativum* (Humphry et al., 2011), *Malus domestica* (Pessina et al., 2014), and *V. vinifera* (Feechan et al., 2008). The loss-of-function mutants, *mlo*, were obtained by using X-ray, which has a broad-spectrum resistance to powdery mildew (*Blumeria graminis* f.sp. *hordei*) (Freisleben and Lein, 1942). Moreover, the downregulation of the *MLO* gene also caused a higher resistance to powdery mildew in barley (Delventhal et al., 2011). In addition, silencing *SlMLO1* gene confers robust powdery mildew resistance in tomato (Bai et al., 2008). All these indicate the important role of the *MLO* gene in plant disease resistance. Consistent with previous studies, this gene's expression in infected leaves was lower than that in the healthy leaves of the two olive cultivars, suggesting this gene's negative regulatory role. It is worth mentioning that subsp. *cuspidata* has a lower expression of this gene than the two cultivars in healthy leaves. Besides, we computed polymorphic sites, Tajima's *D* and θ_π for this gene sequences of all 29 subsp. *cuspidata* and 25 olive cultivar individuals. No polymorphism site was found in subsp. *cuspidata*. All results indicate that this gene has undergone strict positive selection and provide a validated explanation for the higher resistance against pathogens in subsp. *cuspidata*.

TABLE 6 | Statistics of the FPKM values for *evm.model.Chr16.1133*.

Species	Healthy leaves	Mean	Infected leaves	Mean
	1.948		1.047	
'*Arbequina*'	0.999	1.793	0.240	0.558
	2.433		0.388	
	2.871		1.212	
'*Arbosana*'	2.239	3.150	0.677	0.818
	4.341		0.566	
	0.449		-	
subsp. *cuspidata*	0.870	0.583	-	-
	0.431		-	

TABLE 7 | Genetic diversity of *evm.model.Chr16.1133* in 29 *cuspidata* and 25 cultivar individuals.

Species	Tajima's *D*	θ_π	Polymorphic sites
subsp. *cuspidata*	–	–	–
Cultivars	0.929	0.003	21

Overall, we used high-accuracy PacBio HiFi sequencing and Hi-C technologies to assemble a chromosome-level genome of subsp. *cuspidata*, which significantly improved the assembly quality of olive. We performed transcriptome analysis and identified the fungi genus of infected leaves as well as a susceptible gene that was also found in our positive selection analysis. Given the characteristics of smaller fruit size and lower oil content but higher resistance of subsp. *cuspidata* compared with those of subsp. *europaea*, the genome assembly presented here will provide a valuable molecular resource to investigate the differences of oil content and resistance between them.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://db.cngb.org/cnsa/, CNP0002655.

AUTHOR CONTRIBUTIONS

ZWa and JZ planned and designed the research. LW, JZ, DP, YTi, DZ, WN, JLo, JLi, and YZ analyzed the data. LW and ZWa contributed to writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Financial support for this research was provided by the National Key R&D Program of China (No. 2019YFD1000602).

ACKNOWLEDGMENTS

The support of this work from the China National Gene Bank (CNGB) is gratefully acknowledged. We thank Bai Shenglong in Henan University, Kaifeng City, Henan Province and Gu Yincong, Shanghai OE Biotech. Co., Ltd for their valuable suggestions on data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.879822/full#supplementary-material

Supplementary Figure 1 | Graph of the k-mer distribution (K = 21) generated using GenomeScope2.0. The big peak at the coverage of ∼60 in the graph is the homozygous portion of the genome, which accounts for the strands of the DNA having identical 21-mers. The smaller shoulder to the left of the peak corresponds to the heterozygous portion of the genome, which accounts for the strands of the DNA having different 21-mers. If the genome is highly heterozygous, the height of the shoulder peak would be closer to that of the homozygous peak.

Supplementary Figure 2 | GO terms (level 2) distribution of "*Arbequina*" unique genes.

Supplementary Figure 3 | GO terms (level 2) distribution of subsp. *cuspidata* unique genes.

Supplementary Figure 4 | GO terms (level 2) distribution of subsp. *cuspidata* positive selection genes.

Supplementary Figure 5 | GO terms (level 2) distribution of the 49 shared DEGs between "*Arbequina*" and "*Arbosana*".

REFERENCES

Bai, Y., Pavan, S., Zheng, Z., Zappel, N. F., Reinstädler, A., Lotti, C., et al. (2008). Naturally occurring broad-spectrum powdery mildew resistance in a central american tomato accession is caused by loss of MLO function. *Mol. Plant Microbe Interact.* 21, 30–39. doi: 10.1094/MPMI-21-1-0030

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Besnard, G., Baradat, P., Chevalier, D., Tagmount, A., and Bervillé, A. (2001). Genetic differentiation in the olive complex (*Olea europaea*) revealed by RAPDs and RFLPs in the rRNA genes. *Genet. Resour. Crop Evol.* 48, 165–182.

Bolger, A. M., Marc, L., and Bjoern, U. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Buschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., et al. (1997). The barley MLO gene: a novel control element of plant pathogen resistance. *Cell* 88, 695–705. doi: 10.1016/s0092-8674(00)81912-1

Centeno, A., Hueso, A., and Gómez-Del-Campo, M. (2019). Long-term evaluation of growth and production of olive cultivars in super high-density orchard under cold-weather conditions. *Sci. Horticult.* 257, 108657–108657. doi: 10.1016/j.scienta.2019.108657

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., and Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* 364:eaav6202. doi: 10.1126/science.aav6202

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 1–6. doi: 10.1038/s41592-020-01056-5

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016). Genome sequence of the olive tree, *Olea europaea*. *Gigascience* 5:29.

Delventhal, R., Zellerhoff, N., and Schaffrath, U. (2011). Barley stripe mosaic virus-induced gene silencing (BSMV-IGS) as a tool for functional analysis of barley genes potentially involved in nonhost resistance. *Plant Signal. Behav.* 6, 867–869. doi: 10.4161/psb.6.6.15240

Diez, C. M., Trujillo, I., Martinez-Urdiroz, N., Barranco, D., Rallo, L., Marfil, P., et al. (2015). Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* 206, 436–447. doi: 10.1111/nph.13181

Donaire, L., Pedrola, L., Rosa, R., and Llave, C. (2011). High-throughput sequencing of RNA silencing-associated small RNAs in olive (*Olea europaea* L.). *PLoS One* 6:e27916. doi: 10.1371/journal.pone.0027916

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 3, 99–101. doi: 10.1016/j.cels.2015.07.012

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238. doi: 10.1186/s13059-019-1832-y

Feechan, A., Jermakow, A. M., Torregrosa, L., Panstruga, R., and Dry, I. B. (2008). Identification of grapevine MLO gene candidates involved in susceptibility to powdery mildew. *Funct. Plant Biol.* 35, 1255–1266. doi: 10.1071/FP08173

Freisleben, R., and Lein, A. (1942). Über die Auffindung einer mehltauresistenten Mutante nach R?ntgenbestrahlung einer anf?lligen reinen Linie von Sommergerste. *Naturwissenschaften* 30, 608–608. doi: 10.1007/bf01488231

Govindprasad, B., Martin, K., Rodrigo, L. A., Stéphane, T., Rechberger, G. N., Jean-Marc, N., et al. (2017). Sugar versus fat: elimination of glycogen storage improves lipid accumulation in Yarrowia lipolytica. *FEMS Yeast Res.* 17:3. doi: 10.1093/femsyr/fox020

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Green, P. S. (2002). A revision of Olea, *L. (Oleaceae)*. *Kew Bull.* 57, 91–140. doi: 10.2307/4110824

Groenewald, J. Z., Nakashima, C., Nishikawa, J., Shin, H.-D., Park, J.-H., Jama, A. N., et al. (2013). Species concepts in *Cercospora*: spotting the weeds among the roses. *Stud. Mycol.* 75, 115–170. doi: 10.3114/sim0012

Gros-Balthazard, M., Besnard, G., Sarah, G., Holtz, Y., Leclercq, J., Santoni, S., et al. (2019). Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *Plant J.* 100, 1–15. doi: 10.1111/tpj.14435

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, J., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Hackett, N. R., Butler, M. W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Flores, J. L., et al. (2012). RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics* 13:82. doi: 10.1186/1471-2164-13-82

Han, H., and He, F. (2007). Research progress of olive introduction in China. *South China Fruits* 36, 37–42. doi: 10.1016/j.vascn.2019.106600

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi: 10.1093/molbev/mst100

Hannachi, H., Sommerlatte, H., Breton, C., Msallem, M., Gazzah, M. E., Hadj, S., et al. (2009). Oleaster (var. sylvestris) and subsp. cuspidata are suitable genetic resources for improvement of the olive (*Olea europaea* subsp. europaea var. europaea). *Genet. Resour. Crop Evol.* 56, 393–403. doi: 10.1007/s10722-008-9374-2

Humphry, M., Reinstdler, A., Ivanov, S., Bisseling, T., and Panstruga, R. (2011). Durable broad-spectrum powdery mildew resistance in pea er1 plants is conferred by natural loss-of-function mutations in PsMLO1. *Mol. Plant Pathol.* 12, 866–878. doi: 10.1111/j.1364-3703.2011.00718.x

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'donnell, C. J., and de Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938–2939. doi: 10.1093/bioinformatics/btn564

Julca, I., Marcet-Houben, M., Cruz, F., Gómez-Garrido, J., Gaut, B. S., Díez, C. M., et al. (2020). Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biol.* 18:148. doi: 10.1186/s12915-020-00881-6

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Laver, T., Harrison, J., O'neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3, 1–8. doi: 10.1016/j.bdq.2015.02.001

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Y., Han, D., Hu, G., Sommerfeld, M., and Hu, Q. (2010). Inhibition of starch synthesis results in overproduction of lipids in *Chlamydomonas reinhardtii*. *Biotechnol. Bioeng.* 107, 258–268. doi: 10.1002/bit.22807

López-Escudero, F., and Mercado-Blanco, J. (2011). Verticillium wilt of olive: a case study to implement an integrated strategy to control a soil-borne pathogen. *Plant Soil* 344, 1–50. doi: 10.1007/s11104-010-0629-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Ma, T., Ning, D., Yang, W., Zzhang, Z., Li, Y., Xu, T., et al. (2014). *The Breeding of New Olive Cultivar "Jinyefoxilan". China Fruits*, Kunming, 3–4.

Mai, D. H. (1989). Development and regional differentiation of the European vegetation during the Tertiary. *Plant Syst. Evol.* 161, 79–91. doi: 10.1007/978-3-7091-3972-1_4

Majoros, W., Pertea, M., and Salzberg, S. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Marais, G., Delcher, A. L., Phillippy, A. M., Coston, R., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14:e1005944. doi: 10.1371/journal.pcbi.1005944

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mcgrath, C. L., and Lynch, M. (2012). *Evolutionary Significance of Whole-Genome Duplication*. Berlin: Springer, 1–20.

Meisel, B., Korsman, J., Kloppers, F. J., and Berger, D. K. (2009). Cercospora zeina is the causal agent of grey leaf spot disease of maize in southern Africa. *Eur. J. Plant Pathol.* 124, 577–583. doi: 10.1007/s10658-009-9443-1

Moral, J., and Trapero, A. (2009). Assessing the susceptibility of olive cultivars to anthracnose caused by colletotrichum acutatum. *Plant Disease* 93, 1028–1036. doi: 10.1094/PDIS-93-10-1028

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20:275.

Peng, X. L. (2013). *Evolution of Ephedra in the Qinghai-Tibetan Plateau and Adjacent Regions*. Beijing: University of Chinese Academy of Sciences.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Pessina, S., Pavan, S., Catalano, D., Gallotta, A., Visser, R. G. F., Bai, Y., et al. (2014). Characterization of the MLO gene family in Rosaceae and gene expression analysis in *Malus domestica*. *BMC Genomics* 15:618. doi: 10.1186/1471-2164-15-618

Piffanelli, P., Zhou, F., Casais, C., Orme, J., Jarosch, B., Schaffrath, U., et al. (2002). The Barley MLO modulator of defense and cell death is responsive to biotic and abiotic stress stimuli. *Plant Physiol.* 129, 1076–1085. doi: 10.1104/pp.010954

Priore, P., Siculella, L., and Gnoni, G. V. (2014). Extra virgin olive oil phenols down-regulate lipid synthesis in primary-cultured rat-hepatocytes. *J. Nutr. Biochem.* 25, 683–691. doi: 10.1016/j.jnutbio.2014.01.009

Rao, G., Zhang, J., Liu, X., Lin, C., Xin, H., Xue, L., et al. (2021). *De novo* assembly of a new Olea europaea genome accession using nanopore sequencing. *Horticult. Res.* 8:64. doi: 10.1038/s41438-021-00498-y

Roach, M. J., Schmidt, S. A., and Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460. doi: 10.1186/s12859-018-2485-7

Salas, J. J., Harwood, J. L., and Martínez-Force, E. (2013). "Lipid metabolism in olive: biosynthesis of triacylglycerols and aroma components," in *Handbook of Olive Oil*, eds R. Aparicio and J. Harwood (Boston, MA: Springer), 97–127. doi: 10.1007/978-1-4614-7777-8_4

Sanchez-Ortiz, A., Romero-Segura, C., Gazda, V. E., Graham, I. A., Sanz, C., and Perez, A. G. (2012). Factors Limiting the Synthesis of Virgin Olive Oil Volatile Esters. *J. Agric. Food Chem.* 60, 1300–1307. doi: 10.1021/jf203871v

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., et al. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16:259. doi: 10.1186/s13059-015-0831-x

Shane, W. W. (1992). Impact of Cercospora leaf spot on root weight, sugar yield, and purity of *Beta vulgaris*. *Plant Disease* 76:812. doi: 10.1094/pd-76-0812

Shi, Z., Luo, F., Li, Y., Yang, F., Xie, K., and Yang, W. (1991). Study on grafting *Olea europaea* L. with Olea ferruginea Royle as Rootstock. *Acta Bot. Yunnanica* 13, 65–75.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Sun, H., and Li, Z. M. (2003). Evolution and development of the ancient Mediterranean flora after uplift on the Qinghai-Tibetan Plateau. *Adv. Earth Sci.* 18, 852–862.

Takeshi, K., and Tomohiro, S. (2021). High-quality genome assembly of the soybean fungal pathogen *Cercospora kikuchii*. *G3 (Bethesda, Md.)* 11:jkab277. doi: 10.1093/g3journal/jkab277

Tempel, S. (2012). Using and Understanding RepeatMasker. *Methods Mol. Biol.* 859, 29–51. doi: 10.1007/978-1-61779-603-6_2

Terhorst, J., Kamm, J. A., and Song, Y. S. (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* 49, 303–309. doi: 10.1038/ng.3748

Trapero, C., Rallo, L., López-Escudero, F. J., Barranco, D., and Díez, C. M. (2015). Variability and selection of verticillium wilt resistant genotypes in cultivated olive and in the *Olea* genus. *Plant Pathol.* 64, 890–900. doi: 10.1111/ppa.12330

Unver, T., Wu, Z., Sterck, L., Turktas, M., and Peer, Y. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Vogel, J., Lipka, V., Kemmerling, B., Schulze-Lefert, P., Consonni, C., Humphry, M. E., et al. (2006). Conserved requirement for a plant host cell protein in powdery mildew pathogenesis. *Nat. Genet.* 38, 716–720. doi: 10.1038/ng1806

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinform.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, G., Yu, N., Deng, M., and Liu, X. (2000). The development situation of olive in China. *For. Sci. Technol.* 1, 18–19.

Xiang, G., Wang, K., Yan, H., Li, S., Zhou, N., Tang, K., et al. (2018). Bioinformatics analysis of MLO protein family in rosaceae plants. *Genomics Appl. Biol.* 37, 2043–2059.

Xie, Z., Wang, L., Wang, L., Wang, Z., Lu, Z., Tian, D., et al. (2016). Mutation rate analysis *via* parent-progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proc. R. Soc. Biol. Sci.* 283:20161016. doi: 10.1098/rspb.2016.1016

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Ye, C., and Ma, Z. S. (2016). Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *Peerj* 4:e2016. doi: 10.7717/peerj.2016

Ye, X., Yang, W., and Zhang, Z. (1981). Investigation on the effect of Olea europaea L. subsp. cuspidata grafting olive. *Pract. For. Technol.* 3, 12–15.

Zhang, K., Wang, G., Luo, M., Ji, J., Xu, Y., Chen, R., et al. (2010). Evolution of tectonic lithofacies paleogeography of cenozoic of Qinghai-Tibet Plateau and its response to uplift of the plateau. *Earth Sci.* 35:16.

Zheng, Z. (1989). Flora evolution of northwestern mediterranean area since the miocene and the appearance of mediterranean vegetation. *Guihaia* 9, 13–20.

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 19 May 2022
doi: 10.3389/fpls.2022.858526

Population and Landscape Genetics Provide Insights Into Species Conservation of Two Evergreen Oaks in Qinghai–Tibet Plateau and Adjacent Regions

*Keke Liu, Min Qi and Fang K. Du**

School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China

The combination of population and landscape genetics can facilitate the understanding of conservation strategy under the changing climate. Here, we focused on the two most diverse and ecologically important evergreen oaks: *Quercus aquifolioides* and *Quercus spinosa* in Qinghai–Tibetan Plateau (QTP), which is considered as world's biodiversity hotspot. We genotyped 1,657 individuals of 106 populations at 15 nuclear microsatellite loci throughout the species distribution range. Spatial patterns of genetic diversity were identified by mapping the allelic richness (AR) and locally common alleles (LCA) according to the circular neighborhood methodology. Migration routes from QTP were detected by historical gene flow estimation. The response pattern of genetic variation to environmental gradient was assessed by the genotype–environment association (GEA) analysis. The overall genetic structure showed a high level of intra-species genetic divergence of a strong west-east pattern. The West-to-East migration route indicated the complex demographic history of two oak species. We found evidence of isolation by the environment in *Q. aqu*-East and *Q. spi*-West lineage but not in *Q. aqu*-West and *Q. spi*-East lineage. Furthermore, priority for conservation should be given to populations that retain higher spatial genetic diversity or isolated at the edge of the distribution range. Our findings indicate that knowledge of spatial diversity and migration route can provide valuable information for the conservation of existing populations. This study provides an important guide for species conservation for two oak species by the integration of population and landscape genetic methods.

Keywords: species conservation, migration routes, genotype-environment association, *Quercus aquifolioides*, *Quercus spinosa*, Hengduan Mountains, Qinling Mountains

OPEN ACCESS

Edited by:
Alison G. Nazareno, Federal University of Minas Gerais, Brazil

Reviewed by:
Thaís C. S. Dal'Sasso, Universidade Federal de Viçosa, Brazil
Alejandra Moreno-Letelier, National Autonomous University of Mexico, Mexico

***Correspondence:**
Fang K. Du
dufang325@bjfu.edu.cn
orcid.org/0000-0002-7377-5259

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *20 January 2022*
Accepted: *11 April 2022*
Published: *19 May 2022*

Citation:
Liu K, Qi M and Du FK (2022) Population and Landscape Genetics Provide Insights Into Species Conservation of Two Evergreen Oaks in Qinghai–Tibet Plateau and Adjacent Regions. Front. Plant Sci. 13:858526. doi: 10.3389/fpls.2022.858526

INTRODUCTION

The Qinghai–Tibet Plateau (QTP) is the highest and largest plateau with its southern (Himalayas) and southeastern border (Hengduan Mountains, HDM) considered as world's biodiversity hotspots (Myers et al., 2000; Zhang et al., 2002; Mulch and Chamberlain, 2006; Wen et al., 2014). This plateau harbors abundant species richness with more than 12,000 species of vascular plants, many of which

are alpine endemics (Wu et al., 1995; Liu et al., 2000). However, due to anthropogenic habitat loss or fragmentation and climate change, the species diversity has decreased rapidly and led to a sharp decrease in the natural distribution of some species in this region (Xu et al., 2017; Song et al., 2018). Hence, facing the crisis of diversity decrease in the QTP, establishing biodiversity richness areas of conservation priorities is considered one of the most effective strategies for halting the loss of biodiversity (Myers et al., 2000; Geldmann et al., 2018).

Population genetics approach is a useful tool for biodiversity conservation by detecting population substructure, measuring genetic diversity, and identifying potential risks associated with demographic change and inbreeding (Frankham, 1995). One limitation of this approach is the inability to assess spatial patterns of genetic diversity of species across species distribution ranges (Petit et al., 1998). The development of molecular tools in combination with population genetics and geographic information system (GIS) provides opportunities to carry out spatial analyses of genetic diversity patterns (Degen and Scholz, 1998). For example, allelic richness (AR) and locally common alleles (LCA) between circular neighborhoods of sampled populations can be used to interpolate genetic parameters (Hanotte et al., 2002; Hoffmann et al., 2003; Van Zonneveld et al., 2012). Furthermore, recently appearing landscape genetics or genomics approaches integrating genetic variations and landscape characteristics provide novel insights into the molecular basis of local adaptation and conservation strategies (e.g., Manel et al., 2010; Sork et al., 2013; McKinney et al., 2017; Feng and Du, 2022). Therefore, a combination of population and landscape genetics or genomics is likely to provide the best understanding of the molecular imprint of local adaptation and further guide the conservation strategies.

In addition to local adaptation, migrating to new favorable locations is also a response pattern of plants to rapid climate changes, which is important for species conservation (Wulff, 1943; Ozenda, 1988; Donoghue et al., 2001; Donoghue and Smith, 2004). Studies have suggested that plants from the QTP might undergo specific migration patterns, that is, the out-of-QTP hypothesis (Wen et al., 2014, and references therein). Recent phylogenetic studies from various plants have provided evidence to support this hypothesis. For example, *Gentiana* L. diversified initially on the QTP, then dispersed to eastern China, Europe, and other areas (Favre et al., 2016). Similar patterns have been reported in Allium L. Li M. J. et al. (2021), Lagotis Gaertn. Li et al. (2014), Rhodiola L. Zhang et al. (2014), and Picea A. Dietrich Lockwood et al. (2013) (see summary in **Table 1** and reference in Qiu et al., 2011; Liu et al., 2012).

Quercus L. is one of the most diverse and ecologically important tree genera in the QTP and adjacent areas (Huang et al., 1999; Denk et al., 2018). Among these oaks in QTP, two evergreen oak species, *Quercus aquifolioides* Rehd. et Wils. and *Quercus spinosa* David ex Franchet, belonging to a species complex of the genus *Quercus* of section *Ilex*, are the most widely distributed oak species across QTP, HDM, and Qinling Mountains (QM) (Huang et al., 1999). Similar to other oak species, the two species are characterized by monoecious, outcrossing features, wind pollination, and seed dispersal by animals and gravity (Huang et al., 1999; Du et al., 2017; Meng et al., 2017). They display different geographically intraspecific lineages: *Q. spinosa* was diverged into West and East lineages (Feng et al., 2016; Ju et al., 2019), while *Q. aquifolioides* was divided into Tibet and Hengduan Mountains–Western Sichuan Plateau (HDM-WSP) (Du et al., 2017). A recent study further suggested that climatic shift triggered a split of two oak species between the cold highlands and warm lowlands (Meng et al., 2017). In addition, studies using ecological niche models (ENMs) suggested that the two species are relatively stable (Feng et al., 2016; Du et al., 2017; Meng et al., 2017; Ju et al., 2019), but might endure contraction because of spatial constraints, such as land use/cover and human influence (Liao et al., 2021). All the above studies have yielded a substantial understanding of the evolutionary history, phylogeographic patterns, and potential distribution of *Q. aquifolioides* and *Q. spinosa*. However, there are few studies focusing on oak species conservation in this region, despite now they were increasingly threatened by climatic change and habitat fragments. Here, we genotyped 1,657 oak individuals from 106 populations collected across four major regions: QTP, HDM, QM, and warm lowlands in East China based on a dense range-wide sampling of the two species. We aimed to identify the priority areas of *Q. aquifolioides* and *Q. spinosa* for conservation by a combination of population and landscape genetic approaches by answering the following questions: (1) What is the spatial pattern of genetic diversity of two species? (2) What is the species migration route from QTP? and (3) How do the species respond to the environmental gradients?

MATERIALS AND METHODS

Field Sampling, DNA Isolation, and Microsatellite Genotyping

We sampled leaf material from 996 individuals in 60 sites of *Quercus aquifolioides* and 661 individuals in 46 sites of *Quercus spinosa* throughout the species distribution range. The study sites were at least 30 km apart, and individuals were at least 100 m apart from each other to avoid sampling clone individuals. All leaf materials were rapidly dried in silica gel and stored for DNA isolation. The detailed information on sampling sites is depicted in **Figure 1** and **Supplementary Table 1**.

Total genomic DNA was extracted from leaf samples for each individual using an improved cetyltrimethylammonium bromide (CTAB) method (Richards et al., 1994). We randomly selected one individual from each of six distant sites for pre-amplification experiments with 25 nuclear microsatellite (nSSR) loci developed for other oak species (Dow et al., 1995; Steinkellner et al., 1997; Kampfer et al., 1998; Ueno et al., 2008; Durand et al., 2010; **Supplementary Table 2**). We excluded loci harboring null alleles using MICRO-CHECKER 2.2.3 (Van Oosterhout et al., 2004). Departure from Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) was evaluated using GenALEx 6 (Peakall and Smouse, 2006) and FDIST2 (Beaumont and Nichols, 1996). Finally, fifteen successfully amplified SSR loci were retained for subsequent analyses. The reaction procedures are modified from Du et al. (2017). The allele sizes were subsequently

TABLE 1 | Summary of plant studies on the out-of-QTP hypothesis.

Genus/Species	Family	Sample range	Methods	Migration route from QTP	References
Allium spp.	Amaryllidaceae	Europe, Caucasus and southwest Asia	cpDNA, ITS	To Caucasus and Europe.	Li M. J. et al., 2021
Gentiana spp.	Gentianaceae	Global	cpDNA, ITS	To eastern China, Taiwan, Europe, North and South America, Australia and New Guinea.	Favre et al., 2016
Lagotis spp.	Plantaginaceae	Southwest China, northeastern Russia, Kazakhstan and India	cpDNA, ITS	To the central Asian highlands, followed by the northward migration into the arctic.	Li et al., 2014
Rhodiola spp.	Crassulaceae	QTP, north-east Asia, Europe and North America	cpDNA, ITS	To eastern Asia, central Asia, Europe and North America.	Zhang et al., 2014
Picea spp.	Pinaceae	Eastern North America, western North America and QTP	ITS	To western North America and another dispersal into Taiwan.	Lockwood et al., 2013
Anaphalis spp.	Asteraceae	Asia and North America	ITS, ETS	To the eastern Himalayas, eastern Asia, western Himalayas, North America, and southeast Asia.	Nie et al., 2013
Leontopodium spp.	Asteraceae	Europe, central and eastern Asia	AFLP	To Mongolian and central China.	Safer et al., 2011
Leontopodium spp.	Asteraceae	Europe, north and east Asia	ITS, ETS	To middle Asia and eastern Europe.	Blöch et al., 2010
Kelloggia spp.	Rubiaceae	Eastern Asia and western north America	cpDNA	To western North America.	Nie et al., 2005
Sophora davidii	Fabaceae	QTP, Southeast and northeast China	cpDNA, ITS	To the southeast and northeast China.	Fan et al., 2013
Hippophae rhamnoides	Elaeagnaceae	Eastern Asia and Europe	cpDNA, ITS	To central Asia, Asia Minor/Europe, northern China and the Mongolian plateau.	Jia et al., 2012
Lepisorus clathratus	Polypodiaceae	QTP and north-central China	cpDNA	To the north-central China northward into the Altai.	Wang et al., 2011

cpDNA: chloroplast DNA, ITS: internal transcribed spacers, ETS: external transcribed spacers, AFLP: Amplified Fragments Length Polymorphism.

scored using GeneMarker v. 2.2 (Softgenetics, United States), and the genotypes were checked visually two times. A subset of the data, 959 individuals from 58 study sites of *Q. aquifolioides* at 15 nSSRs, were from Du et al. (2017) and Li Y. et al. (2021), and the additional data were first reported in this study.

Genetic Diversity and Differentiation

We estimated genetic diversity indices including mean observed heterozygosity (H_O), mean expected heterozygosity (H_E), mean unbiased expected heterozygosity (uH_E), mean effective population size (N_E), and mean Shannon index (I) by GenAlEx 6 (Peakall and Smouse, 2006). The significance of genetic diversity was evaluated by t-test in SPSS 22 (SPSS Inc., Chicago, IL, United States) with a significance level of 0.05. In order to formulate optimal conservation strategies by revealing priority areas for conservation, we applied spatial analysis to improve the understanding of the geographic distribution of genetic diversity across the oak distribution range. We calculated and mapped the AR and LCA according to the circular neighborhood methodology described by Van Zonneveld et al. (2012). AR, also referred to mean number of alleles per locus, is a straightforward measure of genetic diversity based on molecular markers that aim at selecting populations for conservation (Frankel et al., 1995; Petit et al., 1998). LCA are alleles that occur in 25% or less of all grid cells and with a frequency of at least 5% in a grid cell per locus. Population with high LCA indicate the presence of genotypes adapted to specific environments; therefore, priority for conservation should be given to those populations (Frankel et al., 1995). After applying circular neighborhood to all samples, we calculated the AR and LCA for all 10-minute grid cells by GenAlEx 6 (Peakall and Smouse, 2006). AR was corrected by rarefaction to a minimum sample size of 10 re-sampled trees per cell with the HP-RARE software (Kalinowski, 2005).

We examined the genetic differentiation using hierarchical analysis of molecular variance (AMOVA, Excoffier et al., 1992) in Arlequin 3.5 (Excoffier and Lischer, 2010). The significance of fixation indices was tested using 10,000 permutations in Arlequin 3.5. We used a model-based clustering program implemented in STRUCTURE 2.3 (Pritchard et al., 2000) to infer the genetic clustering without consideration of sampling information. The program was run with the number of clusters (K) varied from 1 to 10 with 20 independent replicates conducted for each K-value, and the length of the burn-in period was set to 100,000 steps followed by the number of Markov chain Monte Carlo (MCMC) after burn-in of 100,000. We selected the optimal K-value by ΔK statistics performed in the web-based program STRUCTURE HARVESTER (Earl and Vonholdt, 2012). Graphic visualization of the STRUCTURE results was produced using DISTRUCT 1.1

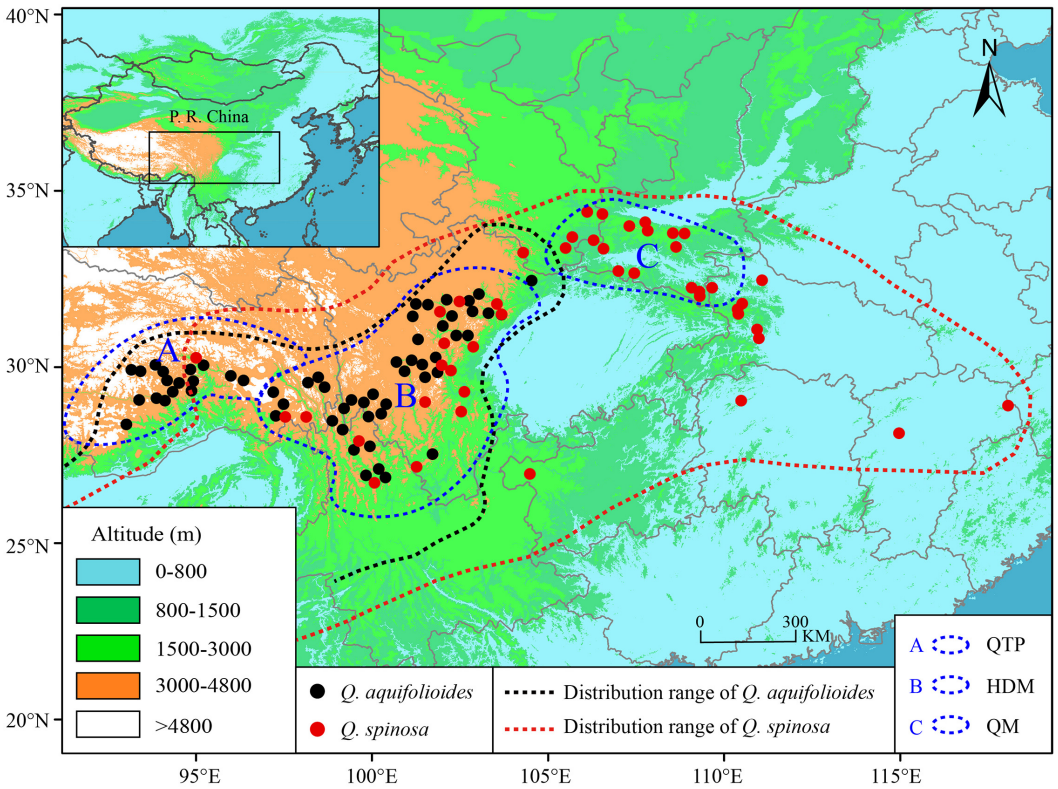


FIGURE 1 | Geographic distribution and sampling sites of *Q. aquifolioides* and *Q. spinosa*. Black and red dashed lines indicate the geographic distribution of *Q. aquifolioides* and *Q. spinosa*, respectively. Three blue dashed lines represent defined research areas. The black rectangle on left top map represents the whole research area. QTP: Qinghai–Tibet Plateau, HDM: Hengduan Mountains, QM: Qinling Mountains.

(Rosenberg, 2004). We also conducted a principal component analysis (PCA) to visualize the genetic relatedness among individuals by calculating principal components (Novembre and Stephens, 2008) using "adegenet" R package (Jombart and Ahmed, 2011). The first two eigenvectors were plotted, and the discrete points reflect the real structure of populations. In addition, we conducted a principal coordinate analysis (PCoA, Gower, 1966) based on genetic covariance among populations in GenAlEx 6 (Peakall and Smouse, 2006) and plotted the first two eigenvectors to visualize genetic relatedness.

Historical Gene Flow Among Lineages

The historical gene flow of two oak species was assessed by Migrate-n 3.6 (Beerli and Felsenstein, 2001; Beerli, 2006) based on the Bayes factor value. First, we generated initial θ ($4N_e\mu$, four times effective population size multiplied by mutation rate per site per generation) and *M* (immigration rate divided by the mutation rate) to estimate the amount and direction of gene flow. A continuous Brownian motion model and the default genetic differentiation were used to generate initial theta and migration values. Then, we started three independent MCMC chains with 500,000 iterations, respectively. We sampled every 100 steps under a constant mutation model and discarded the first 10,000 records as burn-in. After checking the model convergence, we calculated the mode value and 95% posterior probability.

Genotype–Environment Associations Climatic Variables

We obtained climatic variables of the current conditions (~1970–2000) from WorldClim[1], a database of high spatial resolution global weather and climate data (Fick and Hijmans, 2017). A total of 31 climatic variables, including the full suite of 19 mean annual bioclimatic variables and 12 average monthly climate data for precipitation, were downloaded. We excluded climatic variables that were highly correlated with the threshold values of 0.7 using a variance inflation factor (VIF) test in "usdm" R package (Naimi et al., 2014). After avoiding the high multicollinearity bias, four climatic variables, namely precipitation seasonality (bio15, coefficient of variation), mean temperature of the driest quarter (bio09), temperature annual range (bio07, between the minimum temperature of the coldest month and the maximum temperature of the warmest month), and precipitation during June (prec06), were finally remained for downstream analyses (**Supplementary Table 1**).

Linear Relationships

The linear relationships analysis can integrate environmental variables and spatial genetic structure into the analytical framework to assess the contributions of geography

[1]http://www.worldclim.org/version2

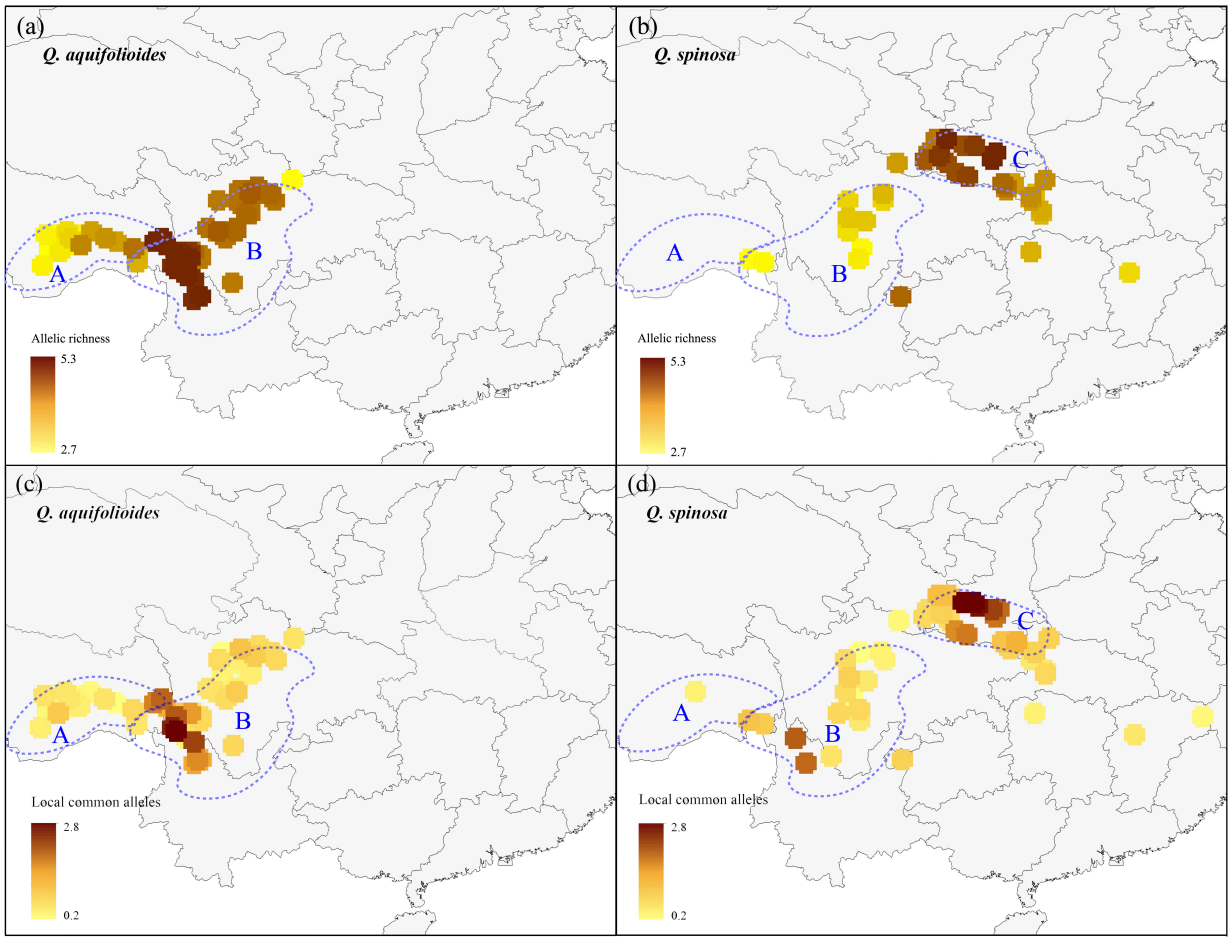


FIGURE 2 | The allelic richness and locally common alleles map of *Q. aquifolioides* and *Q. spinosa*. The light blue dotted lines represent defined three research areas: **(a)** Qinghai–Tibet Plateau (QTP); **(b)** Hengduan Mountains (HDM); **(c)** Qinling Mountains (QM).

and environment in driving genetic differentiation (Feng and Du, 2022). The loading results of this analysis can be interpreted as the response proportion of environmental factors to genetic variation. In this study, we performed isolation-by-resistance (IBR) to illustrate the effects of the heterogeneous landscapes on the population genetic connectivity of two oak species in "ade4" R package (Dray and Dufour, 2007). We first predicted the potential distribution of two oak species based on the current ecological niche model (ENM) in MAXENT (Phillips and Dudik, 2008) and then transformed the environmental rasters into resistance surfaces. We generated the resistance distance based on circuit theory in CIRCUITSCAPE 4.0.5 (McRae, 2006; McRae et al., 2008) and "ResistanceGA" R package (Peterman, 2018). We performed Mantel tests of isolation by distance (IBD; Van Strien et al., 2015) and isolation by environment (IBE; Manthey and Moyle, 2015) to test the linear relationships between geographic or environmental distance and genetic distance using "ecodist" R package (Goslee and Urban, 2007). To distinguish the impact of IBD and IBE, a partial Mantel test was used to evaluate IBE/IBD by controlling the linear influence of geographic/environmental distance (Smouse et al., 1986). In addition, we performed multiple regression on distance matrices (MRM, Lichstein, 2007) to test the multivariate correlation between genetic distance matrix and climate distance using "ecodist" R package (Goslee and Urban, 2007). The significance for Mantel tests and MRM was evaluated by 10,000 permutation tests with the significance level set to 0.05.

We performed redundancy analyses (RDAs) to detect the multivariate relationship between genetic variation and climate variation (Van den Wollenberg, 1977; Legendre and Legendre, 1998) using "vegan" R package (Oksanen et al., 2017). A partial redundancy analysis (*p*RDAs, Legendre and Legendre, 1998, 2012) was performed to avoid the linear influence of geographic/climate variables when analyzing the climate/geographic variables. Statistical significance was evaluated from 999 permutations.

Non-linear Relationships

A limitation of the linear relationships analysis is the inability to fit the variation in the rate of compositional turnover along environmental gradients and the curvilinear relationship between genetic distance and environmental and geographic

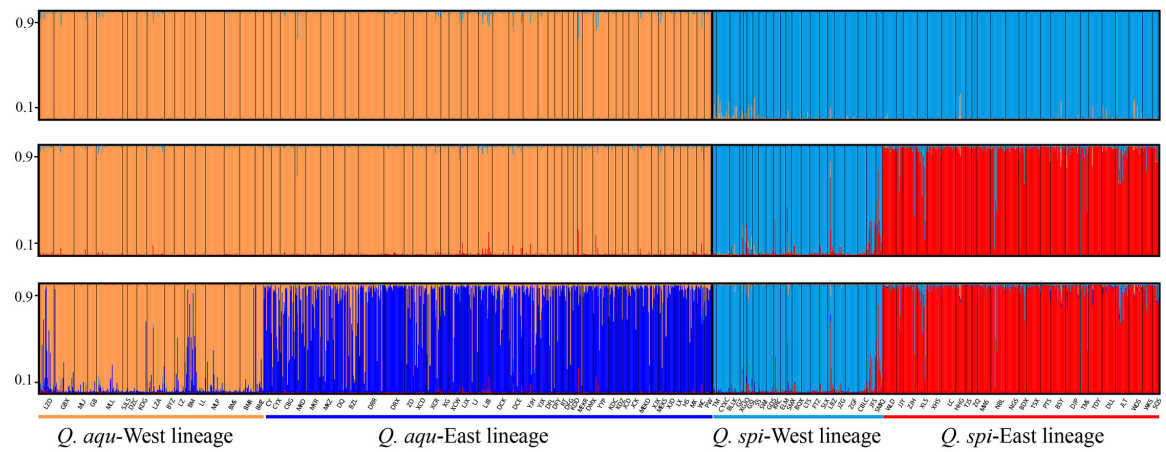


FIGURE 3 | Individual assignment to two **(top)**, three **(middle)**, and four **(below)** genetic clusters by STRUCTURE of *Q. aquifolioides* and *Q. spinosa*. Each bar represents a single individual, with portions of the bar colored depending on the ancestry proportions estimated. The *y*-axis quantifies subgroup membership, and the *x*-axis shows the sample ID for each individual.

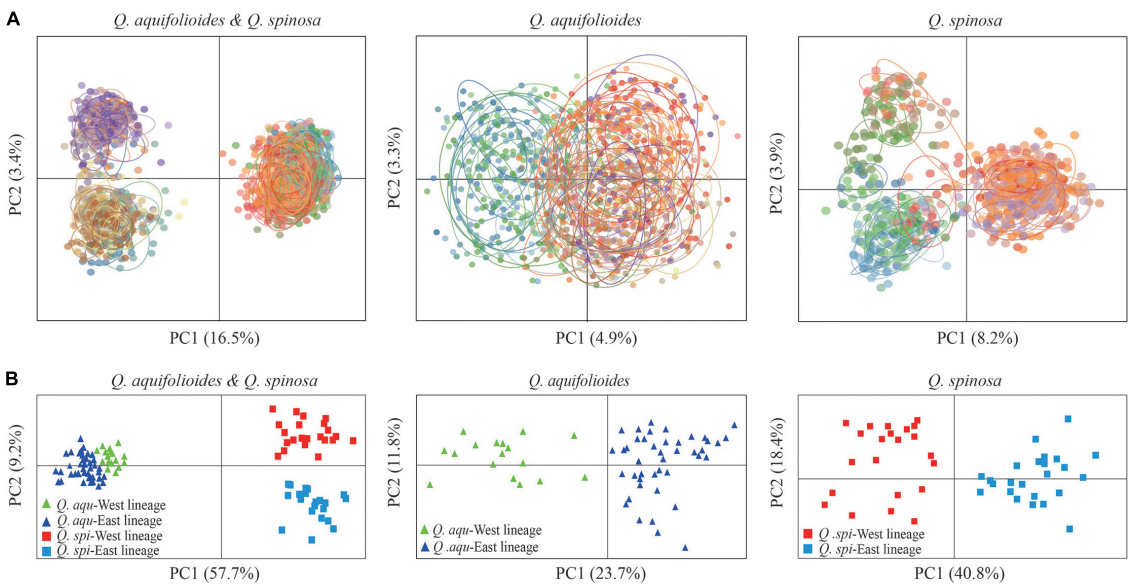


FIGURE 4 | Genetic covariance of *Q. aquifolioides* and *Q. spinosa*. **(A)** Principal component analysis (PCA) plots based on genetic covariance among individuals. The first two principal components (PCs) are shown; **(B)** principal coordinate analysis (PCoA) plots of the first two components based on genetic covariance among populations.

distance. Therefore, a non-linear relationship is essential for applying the associated turnover function to each mapped environmental variable (Fitzpatrick and Keller, 2015). In this study, we performed generalized dissimilarity modeling (GDM) to identify non-linear relationships between genetic distance matrix (response variable) and geographic/environmental distances (predictors) using "gdm" package (Ferrier, 2002; Ferrier et al., 2007). We also evaluated the variation in the rate of allelic compositional change along environmental gradients by fitting splines (Fitzpatrick and Keller, 2015). Genetic distances among individuals were calculated based on allele frequency, and geographic distance was based on Euclidean distance among coordinates. We assessed the variable significance by randomization tests and assessed uncertainty due to sampling error by simulating 1,000 bootstrap iterations (Ferrier et al., 2007; Fitzpatrick et al., 2013).

RESULTS

Genetic Diversity

We found that the genetic diversity was higher in *Q. aquifolioides* than in *Q. spinosa* (H_O: 0.59 vs. 0.41; H_E: 0.58 vs. 0.49; uH_E: 0.61 vs. 0.52; $P < 0.01$) (**Supplementary Table 3**). We also identified

significantly higher genetic diversity in *Q. aqu*-East than *Q. aqu*-West lineage (H_O: 0.60 vs. 0.53; H_E: 0.61 vs. 0.52; uH_E: 0.64 vs. 0.54; $P < 0.01$) and a slightly higher genetic diversity in *Q. spi*-East lineage than *Q. spi*-West lineage (H_O: 0.42 vs. 0.41; H_E: 0.51 vs. 0.47; uH_E: 0.53 vs. 0.50; $P < 0.01$) (**Supplementary Table 3**).

We applied a circular neighborhood re-sampling technique to ensure sufficiently and more evenly distributed data points for spatial diversity analysis. A total dataset of 31,872 trees for *Q. aquifolioides* and 21,152 trees for *Q. spinosa* was used for further AR and LCA analyses (**Supplementary Figure 1**). Our results showed that the enriched regions of AR and

TABLE 2 | Hierarchical analyses of molecular variance (AMOVA) of *Q. aquifolioides* and *Q. spinosa* populations.

	d.f.[1]	SS[2]	VC[3]	Percentage of variation (%)	Fixation indices
All samples					
Between species	1	2738.131	1.70784	26.1	F_{CT} = 0.26
Among populations within species	104	2253.579	0.55798	8.5	F_{SC} = 0.12
Within populations	3208	13710.155	4.27383	65.4	F_{ST} = 0.35
Q. aquifolioides					
Between lineage	1	205.911	0.21585	4.4	F_{CT} = 0.04
Among populations within lineages	58	728.025	0.24428	4.9	F_{SC} = 0.05
Within populations	1932	8704.617	4.5055	90.7	F_{ST} = 0.09
***Q. aqu*-West lineage**					
Among populations	16	179.949	0.18603	4.4	F_{ST} = 0.04
Within populations	647	2621.183	4.05129	95.6	
***Q. aqu*-East lineage**					
Among populations	42	553.323	0.27335	5.4	F_{ST} = 0.05
Within populations	1285	6126.11	4.7674	94.6	
Q. spinosa					
Between lineages	1	324.207	0.48268	9.6	F_{CT} = 0.09
Among populations within lineages	44	983.166	0.64466	12.9	F_{SC} = 0.14
Within populations	1276	4963.162	3.88963	77.5	F_{ST} = 0.22
***Q. spi*-West lineage**					
Among populations	21	462.271	0.8055	18.1	F_{ST} = 0.18
Within populations	482	1767.044	3.66607	81.9	
***Q. spi*-East lineage**					
Among populations	23	520.894	0.54821	11.99	F_{ST} = 0.12
Within populations	794	3196.118	4.02534	88.01	

Significance tests (1,000 permutations) showed all fixation indices were significant ($P < 0.001$). [1]d.f., degrees of freedom; [2]SS, sum of squares; [3]VC, variance component.

TABLE 3 | Historical gene flow as estimated by Migrate-n among *Q. aquifolioides* and *Q. spinosa* based on SSR datasets.

		N_em					
	θ	***Q. aquifolioides* →**	***Q. aqu*-West →**	***Q. aqu*-East →**	***Q. spinosa* →**	***Q. spi*-West →**	***Q. spi*-East →**
Q. aquifolioides	2.0 [1.5-2.3]				47.5 [34.3-58.1]		
Q. aqu-West	4.0 [3.3-4.6]			34.5 [31.7-36.6]		30.6 [24.7-35.9]	10.2 [5.5-15.1]
Q. aqu-East	7.7 [7.2-8.1]		46.6 [43.7-49.1]			36.6 [31.7-41.5]	17.9 [10.1-24.1]
Q. spinosa	3.5 [2.4-4.5]	56.1 [33.6-49.1]					
Q. spi-West	8.4 [7.2-8.8]		34.7 [24.5-44.1]	56.8 [43.7-69.1]			43 [39.3-46.3]
Q. spi-East	2.0 [1.5-2.3]		27.3 [19.1-35.4]	26.6 [13.7-39.1]		65.6 [52.9-69.6]	

The values in square brackets give the 95% credibility interval; θ, 4Neμ; →, source populations; Ne, effective population size; M, mutation-scaled immigration rate; m, immigration rate; μ, mutation rate.

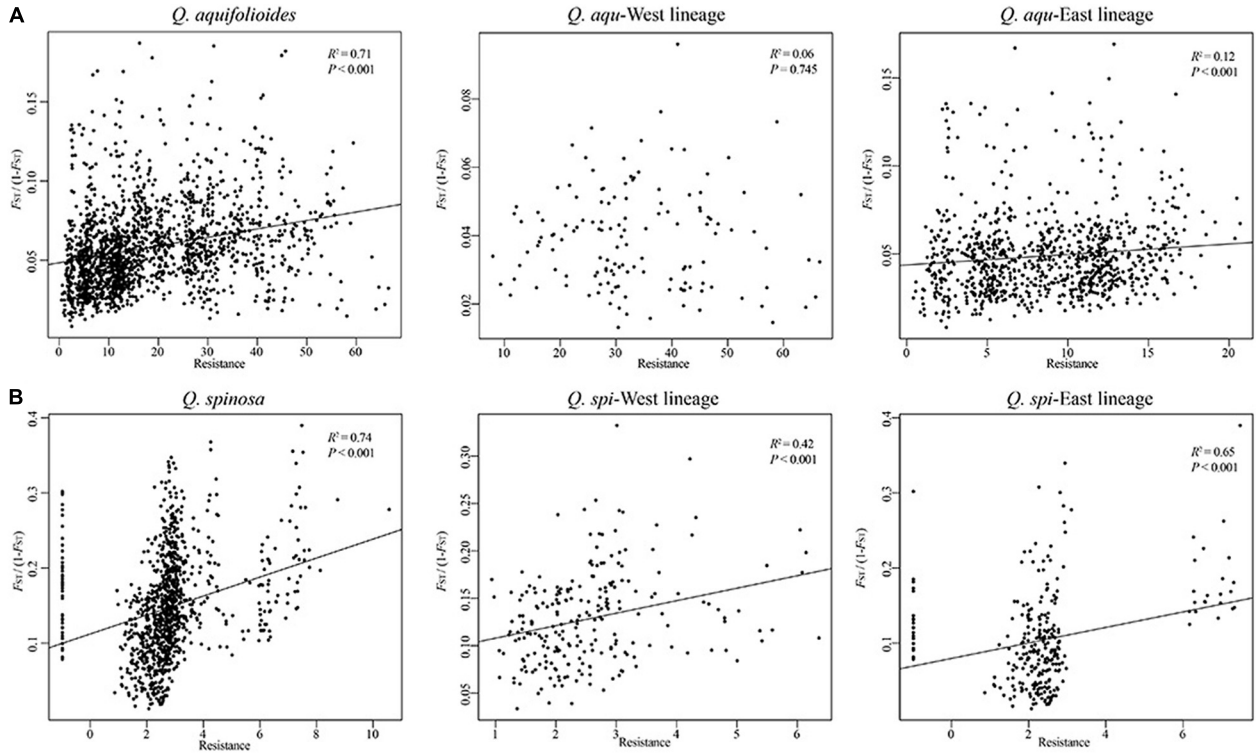


FIGURE 5 | Relationship of genetic distance ($F_{ST}/(1\text{-}F_{ST})$) and resistance distance based on climatic niche suitability of **(A)** *Q. aquifolioides* and **(B)** *Q. spinosa*.

LCA of two oak species were different (**Figure 2**). For *Q. aquifolioides*, the populations located at HDM (*Q. aqu*-East lineage) in southwest Sichuan province and northwest Yunnan province contained higher AR and LCA than QTP (*Q. aqu*-West lineage) (**Figures 2a,c**). The marginal population PW of *Q. aquifolioides* located at the easternmost end of the HDM with a lower AR than in other areas of the HDM (**Figure 2a**). For *Q. spinosa*, populations from QM (*Q. spi*-East lineage) revealed higher AR and LCA than HDM (**Figures 2b,d**).

Genetic Differentiation

Bayesian clustering identified $K = 2$ as the optimal number of evolutionary clusters (**Supplementary Figure 2**), subdivided all individuals into two clusters, one corresponded to *Q. aquifolioides* and the other to *Q. spinosa*. When $K = 3$, *Q. aquifolioides* was maintained unchanged while *Q. spinosa* is further subdivided into two geographically related lineages: *Q. spi*-West (22 sites from HDM) and *Q. spi*-East lineage (24 sites from QM and lowlands in East China). When $K = 4$, *Q. aquifolioides* is divided into *Q. aqu*-West (17 sites from QTP) and *Q. aqu*-East lineage (43 sites from HDM) (**Figure 3**). The results of PCA and PCoA were largely consistent with the STRUCTURE analysis with clear separation in interspecific and intraspecific levels (**Figure 4**). AMOVA showed a high level of genetic differentiation between *Q. aquifolioides* and *Q. spinosa*, and most of the variation occurred within populations (F_{ST} = 0.35, 65.4%) (**Table 2**).

Historical Gene Flow Among Lineages

The Migrate-n analysis generated θ and M values greater than zero, which revealed an asymmetric historical gene flow between two species, mainly from *Q. aquifolioides* to *Q. spinosa* (56.1 vs. 47.5) (**Table 3**). Moreover, we found gene movements occurred predominantly from *Q. aqu*-West into *Q. aqu*-East lineage (46.6 vs. 34.5) and from *Q. spi*-West into *Q. spi*-East lineage (65.6 vs. 43.0) (**Table 3**).

Linear Relationships

Our analyses revealed highly significant correlations between pairwise genetic distances and resistance distance in *Q. aquifolioides* and *Q. spinosa*, but not in *Q. aqu*-West lineage (P = 0.745). The pattern of IBR in *Q. spi*-East was stronger than in *Q. spi*-West lineage (R^2 = 0.65 vs. R^2 = 0.42; **Figure 5**). In addition, we found significant patterns of IBD and IBE by Mantel and partial Mantel tests in *Q. aquifolioides* and *Q. spinosa* (**Table 4** and **Supplementary Figure 3**). However, significant IBD was only detected in *Q. aqu*-West and *Q. spi*-East lineage; significant IBE was detected in *Q. aqu*-East and *Q. spi*-West lineage (**Table 4** and **Supplementary Figure 3**). More specifically, genetic distance was significantly associated with annual range temperature (bio07) and seasonal precipitation (bio15) in *Q. aquifolioides* and *Q. aqu*-East lineage; bio15 and

TABLE 4 | Mantel tests and partial Mantel tests (conditioned with geographic or environmental distances) between pairwise genetic distance ($F_{ST}/(1 - F_{ST})$) and geographic or environmental distances in different lineages and all populations of *Q. aquifolioides* and *Q. spinosa*.

	Q. aquifolioides		***Q. aqu*-West lineage**		***Q. aqu*-East lineage**		***Q. spinosa***		***Q. spi*-West lineage**		***Q. spi*-East lineage**	
	Mantel's *r*	***P***	**Mantel's *r***	***P***	**Mantel's *r***	***P***	**Mantel's *r***	***P***	**Mantel's *r***	***P***	**Mantel's *r***	***P***
Mantel test												
Isolation by Distance (IBD)	0.52	**<0.001**	0.15	**0.006**	0.39	**<0.001**	0.55	**<0.001**	0.38	**<0.001**	0.49	**0.005**
Isolation by Environment (IBE)	0.18	**0.025**	−0.07	0.675	0.3	**0.014**	0.28	**<0.001**	0.25	**0.006**	0.44	0.005
partial Mantel test												
IBD conditioned with environmental distances	0.47	**<0.001**	0.24	**0.007**	0.31	0.059	0.5	**<0.001**	0.13	0.052	0.25	**0.044**
IBE conditioned with geographical distances	0.21	**0.031**	−0.2	0.875	0.19	**<0.001**	0.01	**0.048**	0.32	**0.002**	−0.02	0.566

The bolded text indicates that data are significant.

precipitation during June (prec06) in *Q. spinosa* and *Q. spi*-West lineage (**Supplementary Table 4**). These findings were consistent with optimal MRM models that yielded similar results (**Supplementary Table 5**).

The percentages of variance explained by RDA and *p*RDA were similar, and we thus report results for *p*RDA. Geography (4.0 and 5.2%) explained more genetic variance than climate variables (1.5 and 3.0%) in *Q. aqu*-West and *Q. spi*-East lineage, whereas climate variables (2.7 and 5.6%) explained more genetic variance than geography (1.4 and 5.1%) in *Q. aqu*-East and *Q. spi*-West lineage (**Table 5**). Partitioning of the total genetic variance revealed that bio07 and prec06 explained most genetic variance in *Q. aqu*-East lineage (44.5 and 23.5%), while in *Q. spi*-West lineage, prec06 and bio15 were the two most explanatory environmental variables (46.0% and 20.3%) (**Table 5** and **Supplementary Figure 4**).

Non-linear Relationships

Generalized dissimilarity modeling analyses suggested that geography was the most important predictor among all variables considered in *Q. aquifolioides*, *Q. aqu*-West, *Q. spinosa,* and *Q. spi*-East lineage (59.3, 62.9, 53.4, and 8.8%) (**Supplementary Table 6**). However, there was almost no contribution of geography in *Q. aqu*-East and *Q. spi*-West lineage (2.8 and 7.2%), while bio15 was the most important environmental factor (36.2 and 30.8%) (**Figure 6** and **Supplementary Table 6**). These results were consistent with I-spline analysis (**Supplementary Figure 5**).

DISCUSSION

Population and landscape genetic methods were used to identify priority areas for conservation throughout the species range of *Q. aquifolioides* and *Q. spinosa*. We found that the two evergreen oak species might originate from QTP and then dispersal into HDM and QM. In addition, the intraspecific genetic variation of different lineages of the two species showed different response patterns to environmental factors. Therefore, priority conservation areas were different for the two species: for *Q. aquifolioides*, a priority area for conservation should be at HDM, whereas for *Q. spinosa*, populations from QM should be considered in conservation.

The West-To-East Migration Route From Qinghai–Tibetan Plateau

Quercus aquifolioides and *Quercus spinosa* were assigned to two distinct genetic clusters (**Figures 3**, **4**), and the direction of the interspecific and intraspecific gene flow was from west to east (**Table 3**). These results showed a strong West-to-East migration pattern and likely reflect long-term geographic isolation due to the orogenic history of QTP and adjacent regions (Qiu et al., 2011; Wen et al., 2014). In addition, the amount of historical gene flow was asymmetric, mainly from *Q. aquifolioides* to *Q. spinosa* at the interspecific level (56.1 vs. 47.5) and West lineages to East lineages at the intraspecific level (46.6 vs. 34.5 and 65.6 vs. 43.0) indicating the asymmetric introgression of the species (Currat et al., 2008; Du et al., 2011).

TABLE 5 | Summary of the genetic variations associated with climate and geographic variables based on RDA and *p*RDA in *Q. aquifolioides* and *Q. spinosa*.

	RDA			*p*RDA				RDA			*p*RDA		
	PVE	**Eigenvalue**	***P***	**PVE**	**Eigenvalue**	***P***		**PVE**	**Eigenvalue**	***P***	**PVE**	**Eigenvalue**	***P***
Q. aquifolioides							***Q. spinosa***						
climate	3.3	8.46	0.001	1.44	4.9	0.001	climate	10.34	18.91	0.001	2.67	5.1	0.001
geography				1.89	7.47	0.001	geography				4.13	15.79	0.001
bio15	33.51	11.33	0.001	15.19	2.97	0.002	bio15	60.79	45.99	0.001	27.11	5.53	0.001
bio09	37.06	12.54	0.001	50.62	3.92	0.001	bio09	17.14	12.96	0.001	25.11	5.12	0.001
bio07	14.28	4.83	0.001	11.69	2.29	0.004	bio07	14.14	10.69	0.001	23.72	4.83	0.001
prec06	15.15	5.12	0.001	22.5	4.41	0.001	prec06	7.93	5.99	0.001	24.06	4.91	0.001
Whole model			0.001			0.001	Whole model			0.001			0.001
***Q. aqu*-West lineage**							***Q. spi*-West lineage**						
climate	3.16	2.67	0.001	1.45	3.43	0.001	climate	7.58	5.07	0.001	5.62	3.94	0.001
geography				4.03	2.47	0.001	geography				5.12	7.18	0.001
bio15	18.08	1.93	0.029	23.78	3.26	0.002	bio15	25.67	3.2	0.002	20.31	3.2	0.001
bio09	30.83	3.29	0.001	30.76	4.22	0.001	bio09	23.41	4.74	0.001	16.87	2.65	0.003
bio07	40.88	4.37	0.001	29.41	4.03	0.001	bio07	15.76	5.2	0.001	10.78	2.66	0.004
prec06	10.21	1.09	0.354	16.05	2.2	0.001	prec06	35.16	7.12	0.001	46.04	7.26	0.001
Whole model			0.001			0.001	Whole model			0.001			0.001
***Q. aqu*-East lineage**							***Q. spi*-East lineage**						
climate	3.66	6.25	0.001	2.7	4.66	0.001	climate	6.04	6.48	0.001	2.98	5.74	0.001
geography				1.38	4.75	0.001	geography				5.19	6.57	0.001
bio15	27.17	6.79	0.001	17.27	2.73	0.001	bio15	32.37	8.4	0.001	19.2	4.4	0.001
bio09	14.53	9.3	0.001	14.64	8.3	0.001	bio09	30.59	7.94	0.001	26	5.97	0.001
bio07	37.19	3.63	0.001	44.57	3.22	0.001	bio07	11.28	2.93	0.001	16.4	3.76	0.001
prec06	21.11	5.28	0.001	23.52	4.38	0.001	prec06	25.76	6.69	0.001	38.4	8.81	0.001
Whole model			0.001			0.001	Whole model			0.001			0.001

PVE, percentage of explained variance.

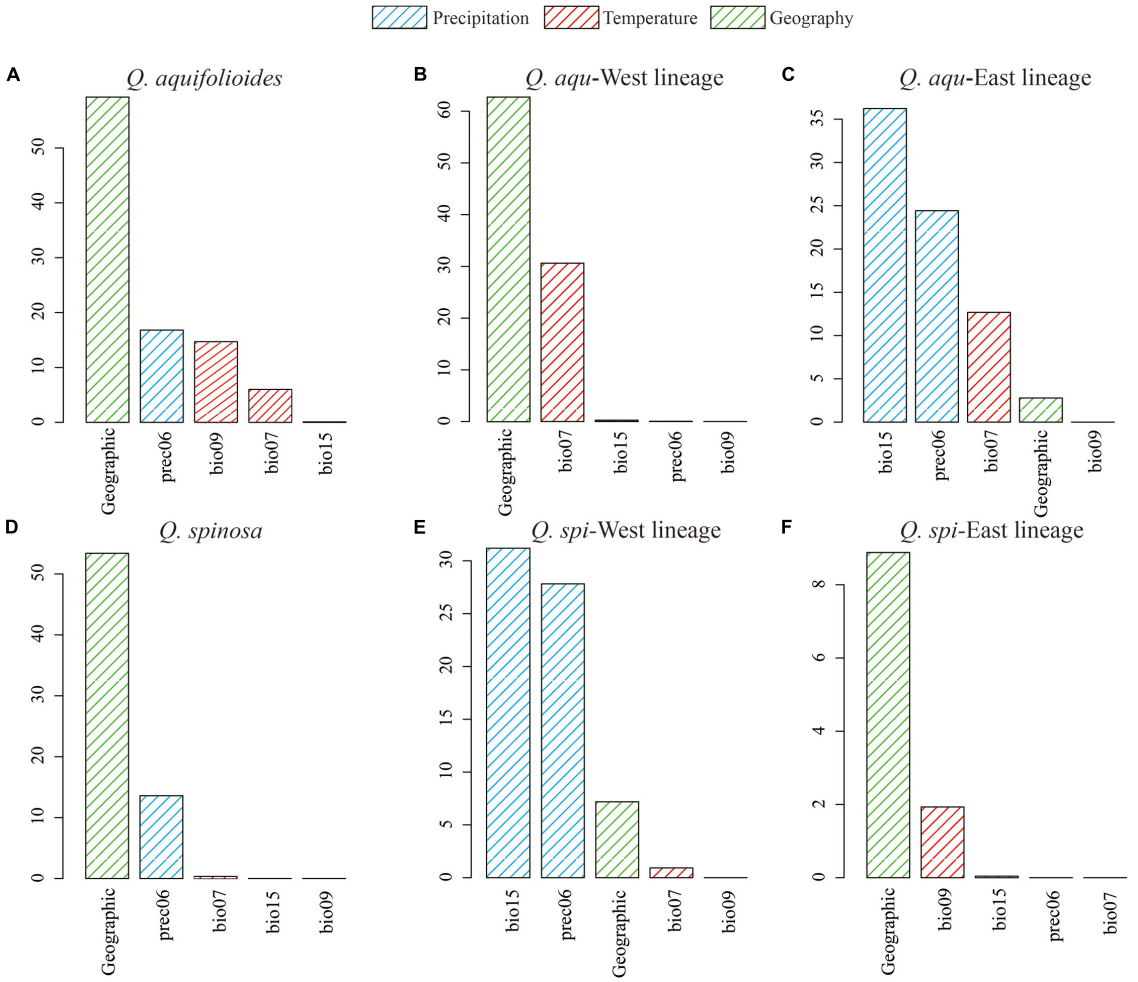


FIGURE 6 | Variable importance of environmental variables based on analysis of generalized dissimilarity model (GDM) for **(A)** all populations of *Q. aquifolioides*, **(B)** *Q. aqu*-West lineage, **(C)** *Q. aqu*-East lineage, **(D)** all populations of *Q. spinosa*, **(E)** *Q. spi*-West lineage, and **(F)** *Q. spi*-East lineage.

There are already several migration routes for the out-of-QTP hypothesis (**Table 1**), and all of the studies suggested that migration, orographic, and climate oscillations catalyzed intraspecific differentiation, diversification, and adaptation of species in this region (see the summary of Wen et al., 2014). It is suggested that orographic and climatic oscillations might result in lots of small fragmented habitats with different microclimates, which could influence the direction of natural selection, and might promote intraspecific high differentiation of species (Sobel et al., 2010). Our results represent a very typical case of a West-to-East migration pattern, which might be triggered by extensive uplifts of the QTP (see the summary of Favre et al., 2015). The QTP uplift events provided opportunities for the ancestral population in this region continually expanded to its eastward ranges and gradually triggered and facilitated speciation and diversifications of oak species (Zhou, 1992). Meanwhile, the West-to-East migration pattern indicated that migrating to new favorable locations might be a survival strategy of species to rapid climate changes as in *Sophora davidii* (Fan et al., 2013) and *Gentiana* (Favre et al., 2016).

Response Pattern of Genetic Variation Under Genotype–Environment Association

Genotype–environment association (GEA) analysis, including mantel tests, redundancy analyses, and generalized dissimilarity modeling, integrates environmental variables and spatial genetic structure into the analytical framework to detect the adaptive variation (Feng and Du, 2022). GEA analysis is essential for understanding the mechanisms underlying the evolutionary responses to environments and was used to quantify patterns of interaction between genetic variation and climate conditions (Hansen et al., 2012). In addition, resistance analysis is important to understand how the species respond to different resistance distances (McRae, 2006). Based on this theory, we identified the landscape resistance matrix that was most highly correlated with pairwise genetic distances in *Q. aquifolioides* and *Q. spinosa*, especially in *Q. aqu*-East and *Q. spi*-East lineage (**Figure 5**), and this result may be related to increased habitat isolation in this area resulted from the disjunct distribution of the two oak species.

However, the pairwise genetic distance was not correlated with resistance distance in *Q. aqu*-West lineage (**Figure 5**), indicating that populations from this lineage are low resistant to dispersal and might endure high genetic connectivity among populations. This result is confirmed with the lowest genetic differentiation in the *Q. aqu*-West lineage (**Table 2**).

The IBD and IBE results indicated the intraspecific lineages of oaks with different response patterns of genetic variation. We detected significant IBD patterns in *Q. aqu*-West and *Q. spi*-East lineage (**Table 4** and **Supplementary Figure 3**, **Supplementary Table 4**). These results were consistent with GDM and RDA (**Figure 6**, **Table 5** and **Supplementary Figure 4**, **Supplementary Table 6**), indicating that the genetic variation of *Q. aqu*-West and *Q. spi*-East lineage was mainly driven by selectively neutral evolutionary processes, not by strong selection pressure from the environment. The complex geological structure of mountains might form a natural geographic barrier for seeds or pollen dispersal, which can provide potential conditions for the formation and independent evolution of plants' intraspecific lineages (Liu et al., 2013; Li et al., 2014).

By contrast, we detected significant IBE patterns in *Q. aqu*-East and *Q. spi*-West lineage (**Table 4** and **Supplementary Figure 5**, **Supplementary Tables 5**, **6**), where the extreme environmental conditions on the plateau might be regarded as a significant climatic barrier, rather than a geographic barrier. It also suggests that geographic isolation may cause interspecific and intraspecific differentiation; adaptation to local climate and environmental factors reinforces this differentiation and gradually forms this significant IBE pattern (Gao et al., 2021). Accordingly, GDM and RDA both suggesting temperature annual range (bio07) and precipitation during June (prec06) were the most significant environmental factor driving genetic variation in *Q. aqu*-East and *Q. spi*-West lineage (**Figure 6**, **Table 5** and **Supplementary Figure 4**, **Supplementary Table 6**), respectively. The temperature may be the main driver of genetic variation for *Q. aqu*-East lineage, and it may influence the growth of plants in microhabitats by affecting the metabolic processes (Wahid et al., 2007). Precipitation might have a great impact on phenological and growth of oak from Q. spi-West lineage and then affect the ability to adapt to climate change.

Priority Areas for Conservation

A better understanding of the spatial distribution of genetic diversity is necessary for the formulation of effective and efficient conservation strategies (Petit et al., 1998). Priority for conservation should be given to populations that retain the highest AR and LCA because the likelihood to find interesting breeding materials is higher in the highest genetic diversity populations, which can indicate the presence of genotypes adapted to specific environments (Frankel et al., 1995; Tanksley and McCouch, 1997). We found that the priority conservation areas were different for the two species based on a large number of samples (1,657 individuals) across their distribution range (**Figure 2**). For *Q. aquifolioides*, a priority area for conservation should be the populations located at HDM (*Q. aqu*-East lineage), which contained the highest AR and LCA. In addition, the marginal population PW of *Q. aquifolioides* located at the easternmost end of the HDM with a lower AR and LCA than other areas. Risk of non-adaptedness (RONA) revealed that this marginal population might be at higher risk of extinction under future climate (Du et al., 2020). Therefore, populations isolated at the edge of the distribution range also should be considered in conservation activities to prevent the extinction of species in this area. For *Q. spinosa,* a priority area for conservation should be the populations located at QM (*Q. spi*-East lineage) with the highest genetic diversity. The second area of a higher diversity of *Q. spinosa* is located on the border zone between Sichuan and Yunnan province, probably related to the high species richness of this region. The higher AR and LCA in the *Q. aqu*-East and *Q. spi*-East lineage than in *Q. aqu*-West and *Q. spi*-West lineage suggested those populations received more immigrants and played an important role in evolution and diversification (López-Pujol et al., 2011; Favre et al., 2015), while the lower genetic diversity in *Q. aqu*-West and *Q. spi*-West lineage likely reflects population contraction or extinction–recolonization dynamics in this area (Qiu et al., 2011; Du et al., 2017; Meng et al., 2017).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FD designed the research. KL performed the experiments and the analyses. KL, MQ, and FD wrote the manuscript. All authors contributed to its revision.

FUNDING

This research was supported by a grant from the National Science Foundation of China (No. 42071060) and 111 Project (No. B20050) to FD.

ACKNOWLEDGMENTS

We thank Yang Xu, Yuyao Wang, and Tianrui Wang for leaf sampling and genotyping. We thank Han Xie for his help in spatial genetic diversity analysis. We thank Li Feng from Xi'an Jiaotong University and Wenting Wang from Northwest Minzu University for the guidance in resistance analysis. We thank Wei Qin for his comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.858526/full#supplementary-material

REFERENCES

Beaumont, M. A., and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. B Biol. Sci.* 263, 1619–1626. doi: 10.1098/rspb.1996.0237

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22, 341–345. doi: 10.1093/bioinformatics/bti803

Beerli, P., and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4563–4568. doi: 10.1073/pnas.081068098

Blöch, C., Dickoré, W. B., Samuel, R., and Stuessy, T. F. (2010). Molecular phylogeny of the edelweiss (*Leontopodium*, Asteraceae-Gnaphalieae). *Edinb. J. Bot.* 67, 235–264. doi: 10.1017/S0960428610000065

Currat, M., Ruedi, M., Petit, R. J., and Excoffier, L. (2008). The hidden side of invasion: massive introgression by local genes. *Evolution* 62, 1908–1920. doi: 10.1111/j.1558-5646.2008.00413.x

Degen, B., and Scholz, F. (1998). Spatial genetic differentiation among populations of European beech (*Fagus sylvatica* L.) in western Germany as identified by geostatistical analysis. *For. Genet.* 5, 191–199. doi: 10.1101/gad.9.21.2598

Denk, T., Grimm, G. W., Manos, P. S., Deng, M., and Hipp, A. L. (2018). "An updated infrageneric classification of the oaks: review of previous taxonomic schemes and synthesis of evolutionary patterns," in *Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus* L, eds E. Gil-Pelegrín, J. Peguero-Pina, and D. Sancho-Knapik (Cham: Springer), 13–38.

Donoghue, M. J., Bell, C. D., and Li, J. H. (2001). Phylogenetic patterns in Northern hemisphere plant geography. *Int. J. Plant Sci.* 162, 41–52. doi: 10.1086/323278

Donoghue, M. J., and Smith, S. A. (2004). Patterns in the assembly of temperate forests around the Northern hemisphere. *Philos. Trans. R. Soc. B Biol. Sci.* 359, 1633–1644. doi: 10.1098/rstb.2004.1538

Dow, B. D., Ashley, M. V., and Howe, H. F. (1995). Characterization of highly variable (GA/CT) $_n$ microsatellites in the bur oak, *Quercus macrocarpa*. *Theor. Appl. Genet.* 91, 137–141. doi: 10.1007/BF00220870

Dray, S., and Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22, 1–20. doi: 10.18637/jss.v022.i04

Du, F. K., Hou, M., Wang, W., Mao, K. S., and Hampe, A. (2017). Phylogeography of *Quercus aquifolioides* provides novel insights into the Neogene history of a major global hotspot of plant diversity in south-west China. *J. Biogeogr.* 44, 294–307. doi: 10.1111/jbi.12836

Du, F. K., Peng, X. L., Liu, J. Q., Lascoux, M., Hu, F. S., and Petit, R. J. (2011). Direction and extent of organelle DNA introgression between two spruce species in the Qinghai-Tibetan Plateau. *New Phytol.* 192, 1024–1033. doi: 10.1111/j.1469-8137.2011.03853.x

Du, F. K., Wang, T. R., Wang, Y. Y., Ueno, S., and De Lafontaine, G. (2020). Contrasted patterns of local adaptation to climate change across the range of an evergreen oak, *Quercus aquifolioides*. *Evol. Appl.* 13, 2377–2391. doi: 10.1111/eva.13030

Durand, J., Bodénès, C., Chancerel, E., Frigerio, J.-M., Vendramin, G., Sebastiani, F., et al. (2010). A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 11:570. doi: 10.1186/1471-2164-11-570

Earl, D. A., and Vonholdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491. doi: 10.1093/genetics/131.2.479

Fan, D. M., Yue, J. P., Nie, Z. L., Li, Z. M., Comes, H. P., and Sun, H. (2013). Phylogeography of *Sophora davidii* (Leguminosae) across the 'Tanaka-Kaiyong Line', an important phytogeographic boundary in Southwest China. *Mol. Ecol.* 22, 4270–4288. doi: 10.1111/mec.12388

Favre, A., Michalak, I., Chen, C. H., Wang, J. C., Pringle, J. S., Matuszak, S., et al. (2016). Out-of-Tibet: the spatio-temporal evolution of *Gentiana* (Gentianaceae). *J. Biogeogr.* 43, 1967–1978. doi: 10.1111/jbi.12840

Favre, A., Paeckert, M., Pauls, S. U., Jaehnig, S. C., Uhl, D., Michalak, I., et al. (2015). The role of the uplift of the Qinghai-Tibetan plateau for the evolution of Tibetan biotas. *Biol. Rev.* 90, 236–253. doi: 10.1111/brv.12107

Feng, L., and Du, F. K. (2022). Landscape genomics in tree conservation under a changing environment. *Front. Plant Sci.* 13:822217. doi: 10.3389/fpls.2022.822217

Feng, L., Qian, Z. Q., Yang, J., Zhang, Y. P., Li, Z. H., and Zhao, G. F. (2016). Genetic structure and evolutionary history of three alpine sclerophyllous oaks in east Himalaya-Hengduan Mountains and adjacent regions. *Front. Plant Sci.* 7:1688. doi: 10.3389/fpls.2016.01688

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Syst. Biol.* 51, 331–363. doi: 10.1080/10635150252899806

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* 13, 252–264. doi: 10.1111/j.1472-4642.2007.00341.x

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

Fitzpatrick, M. C., and Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecol. Lett.* 18, 1–16. doi: 10.1111/ele.12376

Fitzpatrick, M. C., Sanders, N. J., Normand, S., Svenning, J. C., Ferrier, S., Gove, A. D., et al. (2013). Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. *Proc. R. Soc. B Biol. Sci.* 280:20131201. doi: 10.1098/rspb.2013.1201

Frankel, O. H., Brown, A. H. D., and Burdon, J. (1995). "The genetic diversity of wild plants," in *The Conservation of Plant Biodiversity*, 1st Edn, eds O. H. Frankel, A. H. D. Brown, and J. Burdon (Cambridge: Cambridge University Press), 10–38.

Frankham, R. (1995). Conservation genetics. *Annu. Rev. Genet.* 29, 305–327. doi: 10.1146/annurev.ge.29.120195.001513

Gao, Y., Yin, S., Chu, H., Zhang, Y., Wang, H., Chen, H., et al. (2021). Genome-Wide SNPs provide insights on the cryptic genetic structure and signatures of climate adaption in *Amorphophallus albus* Germplasms. *Front. Plant Sci.* 12:683422. doi: 10.3389/fpls.2021.683422

Geldmann, J., Coad, L., Barnes, M. D., Craigie, I. D., Woodley, S., Balmford, A., et al. (2018). A global analysis of management capacity and ecological outcomes in terrestrial protected areas. *Conserv. Lett.* 11:e12434. doi: 10.1111/conl.12434

Goslee, S. C., and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* 22, 1–19. doi: 10.18637/jss.v022.i07

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325–338. doi: 10.1093/biomet/53.3-4.325

Hanotte, O., Bradley, D. G., Ochieng, J. W., Verjee, Y., Hill, E. W., and Gege, J. E. (2002). African pastoralism: genetic imprints of origins and migrations. *Science* 296, 336–339. doi: 10.1126/science.1069878

Hansen, M. M., Olivieri, I., Waller, D. M., and Nielsen, E. E. (2012). Monitoring adaptive genetic responses to environmental change. *Mol. Ecol.* 21, 1311–1329. doi: 10.1111/j.1365-294X.2011.05463.x

Hoffmann, M. H., Glaß, A. S., Tomiuk, J., Schmuths, H., Fritsch, R. M., and Bachmann, K. (2003). Analysis of molecular data of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS). *Mol. Ecol.* 12, 1007–1019. doi: 10.1046/j.1365-294X.2003.01791.x

Huang, C. J., Zhang, Y. L., and Bartholomew, R. (1999). *Fagaceae. Flora of China*. Beijing: Science Press.

Jia, D. R., Abbott, R. J., Liu, T. L., Mao, K. S., Bartish, L. V., and Liu, J. Q. (2012). Out of the Qinghai-Tibet Plateau: evidence for the origin and dispersal of Eurasian temperate plants from a phylogeographic study of *Hippophaë rhamnoides* (Elaeagnaceae). *New Phytol.* 194, 1123–1133. doi: 10.1111/j.1469-8137.2012.04115.x

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Ju, M. M., Feng, L., Yang, J., Yang, Y. C., Chen, X. D., and Zhao, G. F. (2019). Evaluating population genetic structure and demographic history of *Quercus spinosa* (Fagaceae) based on specific length amplified fragment sequencing. *Front. Genet*. 10:965. doi: 10.3389/fgene.2019.00965

Kalinowski, S. T. (2005). HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Mol. Ecol. Notes* 5, 187–189. doi: 10.1111/j.1471-8286.2004.00845.x

Kampfer, S., Lexer, C., Glossl, J., and Steinkellner, H. (1998). Brief report characterization of (GA) $_n$ microsatellite loci from *Quercus robur*. *Hereditas* 129, 183–186. doi: 10.1111/j.1601-5223.1998.00183.x

Legendre, P., and Legendre, L. (1998). *Numerical Ecology*. Amsterdam: Elsevier Science.

Legendre, P., and Legendre, L. (2012). *Complex ecological data sets*. Amsterdam: Elsevier.

Li, G., Kim, C., Zha, H., Zhou, Z., Nie, Z., and Sun, H. (2014). Molecular phylogeny and biogeography of the arctic-alpine genus *Lagotis* (Plantaginaceae). *Taxon* 63, 103–115. doi: 10.12705/631.47

Li, M. J., Yu, H. X., Guo, X. L., and He, X. J. (2021). Out of the Qinghai-Tibetan Plateau and rapid radiation across Eurasia for *Allium* section *Daghestanica* (Amaryllidaceae). *AoB Plants* 13:plab017. doi: 10.1093/aobpla/plab017

Li, Y., Zhang, Y., Liao, P. C., Wang, T. R., Wang, X. Y., Ueno, S., et al. (2021). Genetic, geographic, and climatic factors jointly shape leaf morphology of an alpine oak, *Quercus aquifolioides* Rehder & EH Wilson. *Ann. For. Sci.* 78:64. doi: 10.1007/s13595-021-01077-w

Liao, Z., Nobis, M. P., Xiong, Q., Tian, X. L., Wu, X. G., Pan, K., et al. (2021). Potential distributions of seven sympatric sclerophyllous oak species in Southwest China depend on climatic, non-climatic, and independent spatial drivers. *Ann. For. Sci*. 78:5. doi: 10.1007/s13595-020-01012-5

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecol*. 188, 117–131. doi: 10.1007/s11258-006-9126-3

Liu, J., Möller, M., Provan, J., Gao, L. M., Poudel, R. C., and Li, D. Z. (2013). Geological and ecological factors drive cryptic speciation of yews in a biodiversity hotspot. *New Phytol*. 199, 1093–1108. doi: 10.1111/nph.12336

Liu, J. Q., Chen, Z. D., and Lu, A. M. (2000). The phylogenetic relationships of an endemic genus *Sinadoxa* in the Qinghai-Xizang Plateau: evidence from ITS sequence analysis. *J. Integr. Plant Biol.* 42, 656–658.

Liu, J. Q., Sun, Y. S., Ge, X. J., Gao, L. M., and Qiu, Y. X. (2012). Phylogeographic studies of plants in China: advances in the past and directions in the future. *J. Syst. Evol*. 50, 267–275. doi: 10.1111/j.1759-6831.2012.00214.x

Lockwood, J. D., Aleksic, J. M., Zou, J., Wang, J., Liu, J. Q., and Renner, S. M. (2013). A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Mol. Phylogenet. Evol*. 69, 717–727. doi: 10.1016/j.ympev.2013.07.004

López-Pujol, J., Zhang, F. M., Sun, H. Q., Ying, T. S., and Ge, S. (2011). Centres of plant endemism in China: places for survival or for speciation? *J. Biogeogr*. 38, 1267–1280. doi: 10.1111/j.1365-2699.2011.02504.x

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S., et al. (2010). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol. Ecol*. 19, 3760–3772. doi: 10.1111/j.1365-294X.2010.04717.x

Manthey, J. D., and Moyle, R. G. (2015). Isolation by environment in white-breasted nuthatches (*Sitta carolinensis*) of the Madrean Archipelago sky islands: a landscape genomics approach. *Mol. Ecol*. 24, 3628–3638. doi: 10.1111/mec.13258

McKinney, G. J., Larson, W. A., Seeb, L. W., and Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). *Mol. Ecol. Resour*. 17, 356–361. doi: 10.1111/1755-0998.12649

McRae, B. H. (2006). Isolation by resistance. *Evolution* 60, 1551–1561. doi: 10.1111/j.0014-3820.2006.tb00500.x

McRae, B. H., Dickson, B. G., Keitt, T. H., and Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* 89, 2712–2724. doi: 10.1890/07-1861.1

Meng, H. H., Su, T., Gao, X. Y., Li, J., Jiang, X. L., Sun, H., et al. (2017). Warm-cold colonization: response of oaks to uplift of the Himalaya-Hengduan Mountains. *Mol. Ecol*. 26, 3276–3294. doi: 10.1111/mec.14092

Mulch, A., and Chamberlain, C. (2006). The rise and growth of Tibet. *Nature* 439, 670–671. doi: 10.1038/439670a

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi: 10.1038/35002501

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., and Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37, 191–203. doi: 10.1111/j.1600-0587.2013.00205.x

Nie, Z. L., Funk, V., Sun, H., Deng, T., Meng, Y., and Wen, J. (2013). Molecular phylogeny of *Anaphalis* (Asteraceae, Gnaphalieae) with biogeographic implications in the Northern Hemisphere. *J. Plant Res*. 126, 17–32. doi: 10.1007/s10265-012-0506-6

Nie, Z. L., Wen, J., Sun, H., and Bartholomew, B. (2005). Monophyly of *Kelloggia* Torrey ex Benth. (Rubiaceae) and evolution of its intercontinental disjunction between western North America and eastern Asia. *Am. J. Bot*. 92, 642–652. doi: 10.3732/ajb.92.4.642

Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet*. 40, 646–649. doi: 10.1038/ng.139

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2017). *Package 'Vegan'. Community Ecology Package: Ordination Methods, Diversity Analysis and Other Functions for Community and Vegetation Ecologists. Version, 2*. Nairobi: World Agroforestry.

Ozenda, P. (1988). *Die Vegetation der Alpen im Europaïschen Gebirgsraum*. Stuttgart: Fischer.

Peakall, R., and Smouse, P. E. (2006). GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Peterman, W. E. (2018). ResistanceGA: an R package for the optimization of resistance surfaces using genetic algorithms. *Methods Ecol. Evol*. 9, 1638–1647. doi: 10.1111/2041-210X.12984

Petit, R. J., El Mousadik, A., and Pons, O. (1998). Identifying populations for conservation on the basis of genetic markers. *Conserv. Biol.* 12, 844–855. doi: 10.1111/j.1523-1739.1998.96489.x

Phillips, S. J., and Dudik, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31, 161–175. doi: 10.1111/j.0906-7590.2008.5203.x

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1534/genetics.116.195164

Qiu, Y. X., Fu, C. X., and 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. doi: 10.1016/j.ympev.2011.01.012

Richards, E., Reichardt, M., and Rogers, S. (1994). Preparation of genomic DNA from plant tissue. *Curr. Protoc. Mol. Biol.* 27, 2.3.1–2.3.7. doi: 10.1002/0471142727.mb0203s27

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Safer, S., Tremetsberger, K., Guo, Y. P., Kohl, G., Samuel, M. R., Stuessy, T. F., et al. (2011). Phylogenetic relationships in the genus *Leontopodium* (Asteraceae: Gnaphalieae) based on AFLP data. *Bot. J. Linn. Soc*. 165, 364–377. doi: 10.1111/j.1095-8339.2011.01117.x

Smouse, P. E., Long, J. C., and Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool*. 35, 627–632. doi: 10.2307/2413122

Sobel, J. M., Chen, G. F., Watt, L. R., and Schemske, D. W. (2010). The biology of speciation. *Evolution* 64, 295–315. doi: 10.1111/j.1558-5646.2009.00877.x

Song, Y., Jin, L., and Wang, H. (2018). Vegetation changes along the Qinghai-Tibet Plateau engineering corridor since 2 000 induced by climate change and human activities. *Remote Sens* 10:95. doi: 10.3390/rs10010095

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., and Neale, D. B. (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. doi: 10.1007/s11295-013-0596-x

Steinkellner, H., Fluch, S., Turetschek, E., Lexer, C., Streiff, R., Kremer, A., et al. (1997). Identification and characterization of (GA/CT) $_n$-microsatellite

loci from *Quercus petraea*. *Plant Mol. Biol.* 33, 1093–1096. doi: 10.1023/a:1005736722794

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 227, 1063–1066. doi: 10.1126/science.277.5329.1063

Ueno, S., Taguchi, Y., and Tsumura, Y. (2008). Microsatellite markers derived from *Quercus mongolica* var. *crispula* (Fagaceae) inner bark expressed sequence tags. *Genes Genet. Syst*. 83, 179–187. doi: 10.1266/ggs.83.179

Van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrik* 42, 207–219. doi: 10.1007/BF02294050

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi: 10.1111/j.1471-8286.2004.00684.x

Van Strien, M. J., Holderegger, R., and Van Heck, H. J. (2015). Isolation-by-distance in landscapes: considerations for landscape genetics. *Heredity* 114, 27–37. doi: 10.1038/hdy.2014.62

Van Zonneveld, M., Scheldeman, X., Escribano, P., Viruel, M. A., Van Damme, P., Garcia, W., et al. (2012). Mapping genetic diversity of cherimoya (*Annona cherimola* Mill.): application of spatial analysis for conservation and use of plant genetic resources. *PLoS One* 7:e29845. doi: 10.1371/journal.pone.0029845

Wahid, A., Gelani, S., Ashraf, M., and Foolad, M. R. (2007). Heat tolerance in plants: an overview. *Environ. Exp. Bot*. 61, 199–223. doi: 10.1016/j.envexpbot.2007.05.011

Wang, L., Wu, Z. Q., Bystriakova, N., Ansell, S. W., Xiang, Q. P., Heinrichs, J., et al. (2011). Phylogeography of the Sino-Himalayan fern *Lepisorus clathratus* on "the roof of the world". *PLoS One* 6:e25896. doi: 10.1371/journal.pone.0025896

Wen, J., Zhang, J. Q., Nie, Z. L., Zhong, Y., and Sun, H. (2014). Evolutionary diversifications of plants on the Qinghai-Tibetan Plateau. *Front. Genet*. 5:4. doi: 10.3389/fgene.2014.00004

Wu, S. G., Yang, Y. P., and Fei, Y. (1995). On the flora of the alpine region in the Qinghai-Xizang (Tibet) plateau. *Acta Bot. Yunnanica* 17, 233–250.

Wulff, E. V. (1943). *An Introduction to Historical Plant Geography*. Waltham, MA: Chronica Botanica Company.

Xu, T. T., Wang, Q., Olson, M. S., Li, Z. H., Miao, N., and Mao, K. S. (2017). Allopatric divergence, demographic history, and conservation implications of an endangered conifer *Cupressus chengiana* in the eastern Qinghai-Tibet Plateau. *Tree Genet. Genomes* 13:100. doi: 10.1007/s11295-017-1183-3

Zhang, J. Q., Meng, S. Y., Allen, G. A., Wen, J., and Rao, G. Y. (2014). Rapid radiation and dispersal out of the Qinghai-Tibetan Plateau of an alpine plant lineage *Rhodiola* (Crassulaceae). *Mol. Phylogenet. Evol.* 77, 147–158. doi: 10.1016/j.ympev.2014.04.013

Zhang, Y. L., Li, B. Y., and Zheng, D. (2002). A discussion on the boundary and area of the Tibetan Plateau in China. *Geogr. Res.* 21, 1–8. doi: 10.3321/j.issn:1000-0585.2002.01.001

Zhou, Z. K. (1992). Origin, phylogeny and dispersal of *Quercus* from China. *Acta Bot. Yunnanica* 14, 227–236.

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fpls.2022.920054

Stepped Geomorphology Shaped the Phylogeographic Structure of a Widespread Tree Species (*Toxicodendron vernicifluum*, Anacardiaceae) in East Asia

Lu Wang[1†], Yao Li[1†], Shuichi Noshiro[2], Mitsuo Suzuki[3], Takahisa Arai[3], Kazutaka Kobayashi[3], Lei Xie[1], Mingyue Zhang[1], Na He[4], Yanming Fang[1†] and Feilong Zhang[4]*

[1]Co-Innovation Center for Sustainable Forestry in Southern China, Key Laboratory of State Forestry and Grassland Administration on Subtropical Forest Biodiversity Conservation, College of Biology and the Environment, Nanjing Forestry University, Nanjing, China, [2]Center for Obsidian and Lithic Studies, Meiji University, Tokyo, Japan, [3]Botanical Gardens, Tohoku University, Sendai, Japan, [4]Xi'an Research Institute of Chinese Lacquer, All China Federation of Supply and Marketing Cooperatives, Xi'an, China

OPEN ACCESS

Edited by:
Fang Du,
Beijing Forestry University, China

Reviewed by:
Baosheng Wang,
South China Botanical Garden (CAS), China
Eduardo Cires,
University of Oviedo, Spain

***Correspondence:**
Yanming Fang
jwu4@njfu.edu.cn

†ORCID:
Lu Wang
orcid.org/0000-0002-8684-4305
Yao Li
orcid.org/0000-0001-8081-3703
Yanming Fang
orcid.org/0000-0003-2320-9539

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *14 April 2022*
Accepted: *13 May 2022*
Published: *02 June 2022*

Citation:
Wang L, Li Y, Noshiro S, Suzuki M, Arai T, Kobayashi K, Xie L, Zhang M, He N, Fang Y and Zhang F (2022) Stepped Geomorphology Shaped the Phylogeographic Structure of a Widespread Tree Species (Toxicodendron vernicifluum, Anacardiaceae) in East Asia. Front. Plant Sci. 13:920054. doi: 10.3389/fpls.2022.920054

Species' phylogeographic patterns reflect the interplay between landscape features, climatic forces, and evolutionary processes. Here, we used two chloroplast DNA (cpDNA) markers (*trnL* and *trnL-F*) to explore the role of stepped geomorphology in shaping the phylogeographic structure of *Toxicodendron vernicifluum*, an economically important tree species widely distributed in East Asia. The range-wide pattern of sequence variation was analyzed based on a dataset including 357 individuals from China, together with published sequences of 92 individuals mainly from Japan and South Korea. We identified five chloroplast haplotypes based on seven substitutions across the 717-bp alignment. A clear east-west phylogeographic break was recovered according to the stepped landforms of mainland China. The wild trees of the western clade were found to be geographically restricted to the "middle step", which is characterized by high mountains and plateaus, while those of the eastern clade were confined to the "low step", which is mainly made up of hills and plains. The two major clades were estimated to have diverged during the Early Pleistocene, suggesting that the cool glacial climate may have caused the ancestral population to retreat to at least two glacial refugia, leading to allopatric divergence in response to long-term geographic isolation. Migration vector analyses based on the outputs of ecological niche models (ENMs) supported a gradual range expansion since the Last Interglacial. Mountain ranges in western China and the East China Sea land bridge were inferred to be dispersal corridors in the western and eastern distributions of *T. vernicifluum*, respectively. Overall, our study provides solid evidence for the role of stepped geomorphology in shaping the phylogeographic patterns of *T. vernicifluum*. The resulting east-west genetic discontinuities could persist for a long time, and could occur at a much larger scale than previously reported, extending from subtropical (e.g., the Xuefeng Mountain) to warm-temperate China (e.g., the Taihang Mountain).

Keywords: chloroplast haplotype, dispersal corridor, East Asia, geological isolation, phylogeographic break, refugia, stepped geomorphology, *Toxicodendron vernicifluum*

INTRODUCTION

Species' phylogeographic patterns contain valuable information regarding the impacts of past climatic and geological events on neutral evolutionary processes such as gene flow and genetic drift (Taberlet et al., 1998; Avise, 2000; Hewitt, 2004; Hickerson et al., 2010; Qiu et al., 2011; Feliner, 2014). One of the most remarkable genetic legacies within such patterns is the occurrence of phylogeographic breaks, where intraspecific gene flow is highly restricted and distinct lineages are geographically separated (Avise, 1992; Soltis et al., 1997, 2006; Schaal et al., 1998; Ye et al., 2017a). These breaks may arise when allopatric populations have experienced long-term isolation across major physiographic barriers to dispersals, such as mountains, rivers, glaciers, and oceans (Soltis et al., 2006; Jaramillo-Correa et al., 2009). However, sometimes obvious physical barriers are absent and climate or habitat barriers play a more important role (Geffen et al., 2004; Bai et al., 2016; Cab-Sulub and Álvarez-Castañeda, 2021). In a given area, co-distributed species may exhibit common phylogeographic breaks because they have a shared biogeographic history (Arbogast and Kenagy, 2001; Soltis et al., 2006), but diverse patterns are more frequently observed reflecting the complex history affected by not only a few barriers (Soltis et al., 2006; Shafer et al., 2010; Fan et al., 2016).

East Asia harbors greater species diversity due to its extreme topographic complexity and physiographical heterogeneity (Qian and Ricklefs, 2000; Yin et al., 2021). The absence of continental glaciation and relatively low climate change velocity contribute to the preservation of high plant endemism (Feng et al., 2016). These factors also shape the geographic distributions of intraspecific genealogical lineages across the landscape, allowing long-term refugial isolation and *in situ* survival of local populations in both subtropical and warm-temperate areas (e.g., Wang et al., 2009, 2015a,b; Sakaguchi et al., 2012; Kou et al., 2016; Li et al., 2019), even in cool-temperate regions (e.g., Hu et al., 2008; Zeng et al., 2015; Zhang et al., 2015; Ye et al., 2017b). Nevertheless, climate cooling since the mid-Miocene, the uplift of the Qinghai-Tibetan Plateau (QTP), the intensification of the East Asian monsoon, and repeated sea-level changes during the Pleistocene were shown to have strongly influenced the evolutionary history of local plants (e.g., Gao et al., 2007; Chen et al., 2012; Luo et al., 2016; Jiang et al., 2021; Li et al., 2022). These contexts determine that East Asian plants display distinct phylogeographic patterns in comparison with those on other continents (Qiu et al., 2011, 2017).

Previous phylogeographic studies have confirmed that both north-south and east-west genetic discontinuities are common patterns for temperate and subtropical plants in East Asia (Qiu et al., 2017; Ye et al., 2017b). A clear north-south phylogeographic split has been reported for widespread tree species such as *Acer mono* (Guo et al., 2014; Liu et al., 2014), *Juglans* spp. (Bai et al., 2016), and *Lindera obtusiloba* (Ye et al., 2017b). This break is closely associated with an east-west orientated arid belt that has existed during the Paleogene and redeveloped during the late Miocene (Guo et al., 2008). The belt was inferred to have acted as a climate barrier that impeded the migration across it and resulted in the late Miocene diversification of Tertiary relict plants (Bai et al., 2016; Ye et al., 2017a). Furthermore, the boundary between the subtropical and tropical regions was found to have shaped the north-south patterns of dominant species in evergreen broadleaved subtropical forests (e.g., *Lindera aggregata*, Ye and Li, 2021).

East-west phylogeographic splits were more commonly observed in East Asian plants (Qiu et al., 2017; Ye et al., 2017a). Previous studies have identified several phylogeographic breaks, coinciding with the East China Sea (ECS; e.g., *Platycrater arguta*, Qi et al., 2014; *Euptelea* spp., Cao et al., 2016) or with the boundary between the Sino-Himalayan and Sino-Japanese Forest subkingdoms (e.g., *Taxus wallichiana*, Gao et al., 2007; *Davidia involucrata*, Luo et al., 2011; and *Sophora davidii*, Fan et al., 2013). More interestingly, a recent study demonstrated that the stepped landforms of mainland China, together with the ECS, play a more important role in shaping the distinct phylogeographic structure of a widespread shrub, *Kerria japonica* (Luo et al., 2021). The geomorphology of China is characterized by three giant "steps": the high (average ~4,000 m, e.g., the QTP), middle (average ~2,000 m, e.g., the Yunnan-Guizhou Plateau and the Qinling Mountains), and low (average <500 m, e.g., the plains and hills in eastern China) "steps" spanning from the west to the east (Jiang and Wu, 1993; Wan, 2012). The mountain ranges between these three areas may have served as geographic barriers that prevent gene flow and further facilitate population differentiation (Li et al., 2015). Indeed, limited chloroplast haplotype sharing between the middle and low "steps" has been reported for several woody and herbaceous plants in subtropical China (e.g., *Juglans cathayensis* Bai et al., 2014; *Boea clarkeana*, Wang et al., 2018; *Liriodendron chinense*, Yang et al., 2019) as well as a few widespread tree species in East Asia (e.g., *Kalopanax septemlobus*, Sakaguchi et al., 2012).

In this study, we used chloroplast DNA (cpDNA) to explore the role of stepped geomorphology in shaping the phylogeographic pattern of *Toxicodendron vernicifluum* (Stokes) F. A. Barkley, a deciduous and dioecious tree widely distributed in temperate and subtropical areas of East Asia. This species belongs to the family Anacardiaceae and is commonly known as lacquer tree (also called "qishu" in Chinese, "urushi" in Japanese, and "otnamu" in Korean; Hashida et al., 2014; Suzuki et al., 2014; Kim et al., 2015; Wang et al., 2020a). It has been cultivated in East Asian countries (China, Korea, and Japan) for thousands of years, whose toxic sap is traditionally used as a highly durable lacquer to make lacquerware (Noshiro and Suzuki, 2004; Walker et al., 2008; Zhao et al., 2013; Suzuki et al., 2014; Wu et al., 2018; Li et al., 2021). In western China, wild lacquer trees usually grow in mountain forests at an altitude between 800 and 2,800 m. It is mainly distributed in the mountainous areas surrounding the Sichuan Basin (e.g., the Qinling Mountains, the Daba Mountains, the Wuling Mountains, the Dalou Mountains, and the Wumeng Mountains). In eastern China, natural forests of lacquer trees are scattered in hilly areas at an altitude less than 600 m, such as those in Liaoning and Shandong Provinces (Suzuki et al., 2007; Zhang et al., 2007). It is believed that lacquer trees in Japan and Korea are introduced from mainland China (Noshiro and Suzuki, 2004). However, a fossil wood dating back to the

incipient Jomon period (~12,600 years ago) suggested the species is likely to be native to Japan (Suzuki et al., 2014).

A previous study has used two cpDNA fragments to examine the phylogenetic relationships between lacquer trees sampled from mainland China and Japan (Suzuki et al., 2014). They found that populations in eastern China and Japan shared a haplotype, while those in western China harbored another haplotype, suggesting that the stepped landforms in China may have shaped the present phylogeographic patterns of *T. vernicifluum*. However, this study only sampled a limited number of lacquer trees from China. Other researchers also used various molecular markers (e.g., amplified fragment length polymorphisms and nuclear microsatellites) to investigate the genetic variation patterns of *T. vernicifluum* at different scales (Wei et al., 2010; Bai et al., 2017; Bui et al., 2017; Vu et al., 2018; Guo et al., 2019; Watanabe et al., 2019), but their conclusions are mainly restricted by small sample size and a narrow sampling range. Here, we used the two cpDNA makers of Suzuki et al. (2014) to sequence the samples from 39 sites that encompass the entire natural range of *T. vernicifluum* in China. The obtained dataset was combined with that of Suzuki et al. (2014) to investigate the phylogeographic structure of *T. vernicifluum* throughout East Asia. We aimed to: (1) examine range-wide patterns of cpDNA variation and identify potential phylogeographic breaks; (2) explore the role of stepped geomorphology in shaping the present phylogeographic patterns; (3) infer the possible migration routes and dispersal corridors across the landscape of East Asia.

MATERIALS AND METHODS

Sampling, DNA Extraction, PCR Amplification, and Sequencing

Based on specimens and records in local flora, we selected 39 sampling sites that encompass the entire natural range of *T. vernicifluum* in China (**Figure 1**; **Table 1**). Between August 2018 and August 2021, we sampled leaf tissue from 357 individuals at these sites. Among those, 308 individuals from 36 sites were wild or semi-wild, while the remaining 49 individuals from seven sites were under cultivation (**Table 1**). At each site, three to 11 trees spaced >30 m apart were randomly sampled. Spatially explicit information was recorded for each tree using the 2bulu Outdoor Assistant app.[1] Voucher specimens were deposited in the Herbarium of Nanjing Forestry University (NF).

Total genomic DNA was extracted from silica-gel dried leaf material according to the manufacturer's protocol for the Plant Genomic DNA Kit (Tiangen, Beijing, China). The concentration of DNA samples was diluted to 10 ng/μl and stored at −20°C for PCR amplification. Following Suzuki et al. (2014), we used two chloroplast intergenic spacers (*trnL* and *trnL-F*) to sequence all the 357 samples. Polymerase chain reactions (PCRs) were performed using a Mastercycler pro Thermal Cycler (Eppendorf, Germany) in 25-μl reaction volumes as described by Zhang et al. (2015). Thermal cycling started with a denaturation step lasting 10 min at 95°C, followed by 30 cycles each comprising 30 s of denaturation at 94°C, 40 s of annealing at 50°C, and 60 s of elongation at 72°C. Amplification ended with a 10-min extension at 72°C. The PCR products were purified and sequenced using the ABI 3730XL DNA Analyzer by Sangon Biotech (Shanghai, China).

We compiled previously published sequence data of the same two cpDNA regions for 92 *T. vernicifluum* trees sampled from 30 sites in Japan ($n = 46$ trees), two sites in South Korea ($n = 6$ trees), and 12 sites in China ($n = 40$ trees; Suzuki et al., 2014). Among those, all the Japanese and Korean samples were collected from cultivated trees, while 15 (from six sites) and 25 (from 10 sites) Chinese samples were collected from cultivated and wild trees, respectively. Given that four sites in Suzuki et al. (2014) were close to our sampling locations (i.e., FOP, ZAO, YTS, and BEN), the corresponding sequence data were combined for each site. Finally, we obtained a dataset for 449 *T. vernicifluum* trees (333 wild trees and 116 cultivated trees) sampled from 79 sites, including 397 trees from 47 sites in China, 46 trees from 30 sites in Japan, and six trees from two sites in South Korea (**Table 1**). Furthermore, we also compiled the sequence data for five congeners that were used as outgroups, including *T. succedaneum* (Suzuki et al., 2014; Wang et al., 2020b), *T. sylvestre* (He et al., 2020), *T. trichocarpum* (Suzuki et al., 2014), *T. radicans* (Suzuki et al., 2014), and *T. griffithii* (Li et al., 2020). GenBank accession numbers for all the samples analyzed in this study were listed in **Supplementary Table S1**.

Chloroplast Sequence Data Analyses

All sequences were checked and aligned by BioEdit 7.2.5 (Hall, 1999). The obtained alignments were concatenated into a single matrix using FasParser 2.1.1 (Sun, 2017). A 96-bp indel detected in the *trnL-F* region was treated as a single mutation event and coded as a substitution (A/T). Chloroplast haplotypes were determined by DnaSP 5.10 (Librado and Rozas, 2009). A median-joining network was inferred with PopART 1.7 to visualize the phylogenetic relationships among haplotypes (Leigh and Bryant, 2015). ArcGIS 10.5 was employed to show the geographic distribution of haplotypes across the range of *T. vernicifluum* in East Asia.

We calculated haplotype diversity (H_d) and nucleotide diversity (π) for each sampling site using DnaSP. We estimated average gene diversity within sampling sites (h_S), total gene diversity (h_T), and two genetic differentiation coefficients G_{ST} and N_{ST} using Permut 2.0 (Pons and Petit, 1996). N_{ST} is a measure of genetic differentiation among sites considering genetic distances between haplotypes, whereas G_{ST} is an unordered measure that does not take distances among haplotypes into account. A higher N_{ST} than G_{ST} usually indicates the presence of a phylogeographic structure, i.e., closely related haplotypes are more frequently observed in the same populations than less related ones (Pons and Petit, 1996). The significance of the difference between G_{ST} and N_{ST} was tested by a permutation test ($n = 10{,}000$). In these analyses, we excluded the cultivated trees in China but included those in Japan and South Korea

[1]https://www.2bulu.com/

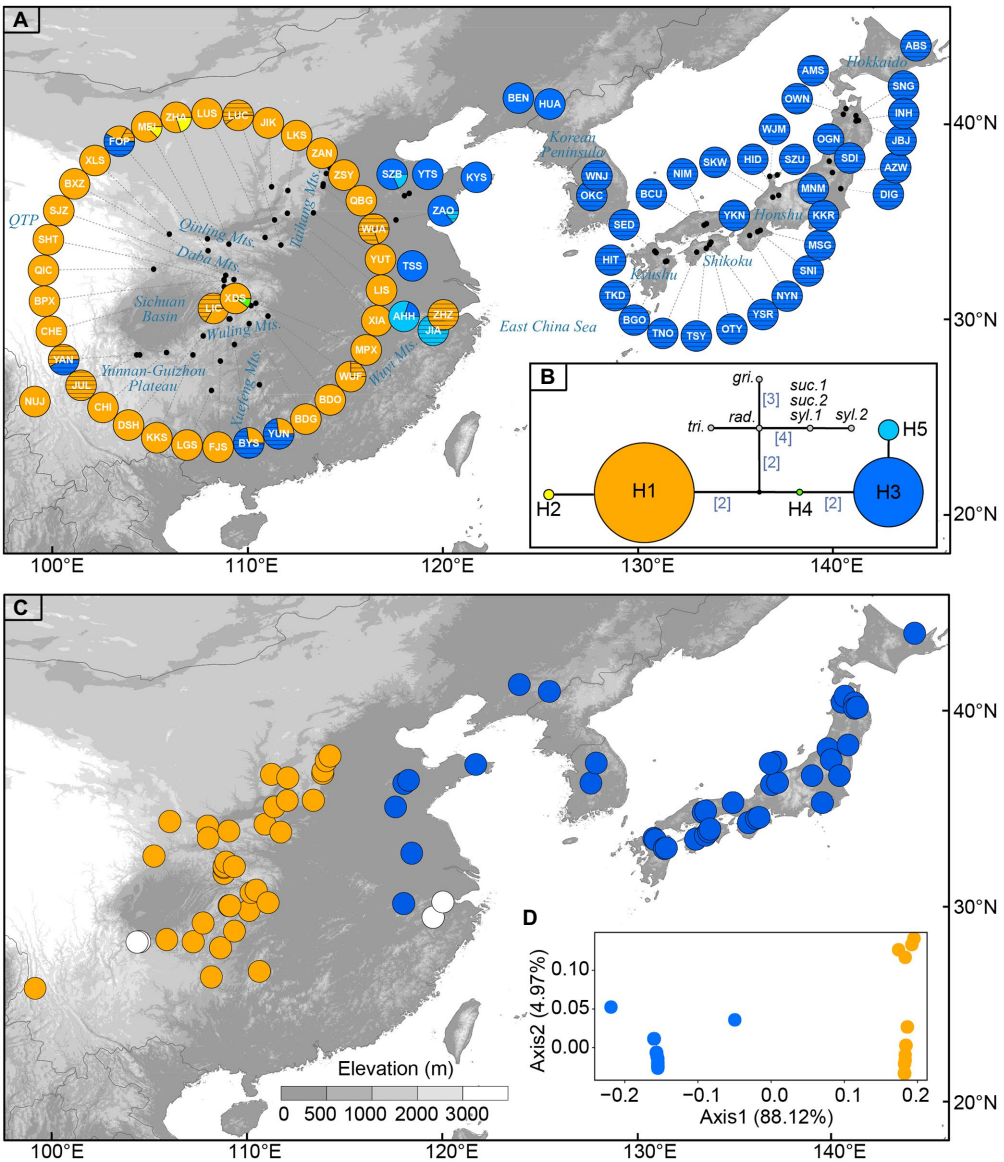


FIGURE 1 | (A) Sampling sites of *Toxicodendron vernicifluum* in East Asia and geographic distribution of the five chloroplast (cp) haplotypes identified in this study. Each pie chart represents a sampling site (see **Table 1** for site codes) and each haplotype is represented with a different color as shown in **(B)**. Sectors marked by black lines represent cultivated trees. (**B**) Median-joining network of cpDNA haplotypes estimated by POPART. Numbers in brackets on branches indicate the number of mutations between haplotypes when branches represent more than one mutation. Outgroups include *Toxicodendron griffithii* (*gri.*), *Toxicodendron radicans* (*rad.*), *Toxicodendron trichocarpum* (*tri.*), *Toxicodendron succedaneum* (*suc.*), and *Toxicodendron sylvestre* (*syl.*). **(C)** The western (orange) and eastern (blue) groups were identified by spatial analysis of molecular variance (SAMOVA) when $K = 2$. White circles represent sampling sites only including cultivated trees in mainland China. **(D)** Results of principal coordinate analysis (PCoA) based on the matrix of population pairwise F_{ST}. Orange and blue circles represent the western and eastern groups, respectively.

because they represent the distribution of *T. vernicifluum* outside China and may have a natural origin in mainland China. Given that Permut requires a minimum sample size of three for each sampling site, we first combined the adjacent sites ($n < 3$) within the same $0.625° \times 0.625°$ grid into one site and then removed the sites still with less than three wild trees. Finally, 10 sites in Japan were combined into five sites and 27 sites (including eight sites in China, one site in South Korea, and 18 sites in Japan) were excluded, resulting in a dataset comprising 353 individuals from 47 sites (including 327 trees from 39 sites in China, four trees from one site in South Korea, and 22 trees from seven sites in Japan; **Supplementary Table S2**).

To examine the phylogeographic structure of *T. vernicifluum*, we performed a spatial analysis of molecular variance (SAMOVA) using the software SAMOVA 2.0 (Dupanloup et al., 2002). This analysis used a simulated annealing procedure to maximize the proportion of total genetic variance (F_{CT}) due to differences

TABLE 1 | Locations and genetic statistics of 79 sampling sites of *Toxicodendron vernicifluum*.

Sampling sites	Location	Lon (E)	Lat (N)	E (m)	Wild trees				Cultivated trees	
					n_w	Haplotypes	H_d	π	n_c	Haplotypes
Western group										
NUJ	Nujiang, Yunnan, China	99.10	25.78	2,204	10	H1(10)	0	0	0	–
YAN	Yanjin, Yunnan, China	104.32	28.16	997	0	–	–	–	7	H1(4), H3(3)
JUL	Junlian, Sichuan, China	104.48	28.19	393	0	–	–	–	10	H1(10)
QIC	Qingchuan, Sichuan, China	105.19	32.58	1,099	10	H1(10)	0	0	0	–
CHI	Chishui, Guizhou, China	105.86	28.29	1,271	3	H1(3)	0	0	0	–
KKS	Suiyang, Guizhou, China	107.20	28.18	1,242	10	H1(10)	0	0	0	–
DSH	Daozhen, Guizhou, China	107.73	29.14	1,304	10	H1(10)	0	0	0	–
LGS	Leishan, Guizhou, China	108.16	26.37	1,204	10	H1(10)	0	0	0	–
FJS	Jiangkou, Guizhou, China	108.61	27.87	1,005	10	H1(10)	0	0	0	–
BYS	Baojing, Hunan, China	109.33	28.71	436	3	H1(3)	0	0	8	H3(8)
BDG	Sangzhi, Hunan, China	110.09	29.78	1,336	10	H1(10)	0	0	0	–
YUN	Wugang, Hunan, China	110.61	26.65	1,166	3	H1(3)	0	0	8	H3(8)
LIC*	Lichuan, Hubei, China	109.06	30.03	829	1	H1(1)	–	–	2	H1(2)
XDS	Lichuan, Hubei, China	109.10	30.03	1,019	10	H1(9), H4(1)	0.200	0.00084	0	–
BDO*	Badong, Hubei, China	110.18	30.70	1,378	2	H1(2)	0	0	0	–
MPX	Zigui, Hubei, China	110.43	30.82	1,100	10	H1(10)	0	0	0	–
WUF*	Wufeng, Hubei, China	111.05	30.17	296	3	H1(3)	0	0	1	H1(1)
CHE*	Chengkou, Chongqing, China	108.80	31.67	2,294	6	H1(6)	0	0	0	–
BPX	Chengkou, Chongqing, China	108.80	31.99	1,140	10	H1(10)	0	0	0	–
XLS	Maiji, Gansu, China	106.00	34.36	1,397	10	H1(10)	0	0	0	–
MEI	Meixian, Shaanxi, China	107.95	34.12	692	8	H1(7), H2(1)	0.250	0.00035	0	–
FOP**	Foping, Shaanxi, China	107.98	33.51	895	3	H1(3)	0	0	9	H1(2), H3(7)
SHT	Langao, Shaanxi, China	108.80	32.05	2,409	10	H1(10)	0	0	0	–
SJZ	Langao, Shaanxi, China	108.88	32.27	741	10	H1(10)	0	0	0	–
ZHA	Zhashui, Shaanxi, China	109.05	33.86	1,178	10	H1(8), H2(2)	0.356	0.00050	0	–
BXZ	Pingli, Shaanxi, China	109.31	32.03	1,637	10	H1(10)	0	0	0	–
LUS	Lushi, Henan, China	110.90	34.21	1,194	10	H1(10)	0	0	0	–
LUC	Luanchuan, Henan, China	111.71	33.80	1,542	4	H1(4)	0	0	6	H1(6)
YUT	Xiuwu, Henan, China	113.37	35.45	1,016	5	H1(5)	0	0	0	–

(Continued)

TABLE 1 | Continued

Sampling sites	Location	Lon (E)	Lat (N)	E (m)	Wild trees				Cultivated trees	
					n_w	Haplotypes	H_d	π	n_c	Haplotypes
JIK	Jiaokou, Shanxi, China	111.23	36.78	1,348	10	H1(10)	0	0	0	–
XIA	Xiaxian, Shanxi, China	111.39	35.09	936	10	H1(10)	0	0	0	–
LIS	Qinshui, Shanxi, China	112.05	35.43	1,510	10	H1(10)	0	0	0	–
LKS	Qinyuan, Shanxi, China	112.07	36.59	1,582	10	H1(10)	0	0	0	–
WUA*	Wu'an, Hebei, China	113.87	36.81	694	1	H1(1)	–	–	4	H1(4)
QBG	Wu'an, Hebei, China	113.88	36.94	752	10	H1(10)	0	0	0	–
ZSY	Zanhuang, Hebei, China	114.04	37.46	977	10	H1(10)	0	0	0	–
ZAN*	Zanhuang, Hebei, China	114.23	37.72	1,045	2	H1(2)	0	0	0	–
Eastern group										
AHH	Huangshan, Anhui, China	118.02	30.14	293	5	H3(1), H5(4)	0.400	0.00056	0	–
ZAO**	Zaozhuang, Shandong, China	117.52	35.11	229	12	H3(12)	0.264	0.00037	2	H5(2)
SZB	Boshan, Shandong, China	118.04	36.31	484	11	H3(9), H5(2)	0.327	0.00046	0	–
YTS**	Qingzhou, Shandong, China	118.28	36.46	599	11	H3(11)	0	0	0	–
KYS	Yantai, Shandong, China	121.73	37.27	209	7	H3(7)	0	0	0	–
BEN**	Benxi, Liaoning, China	123.87	41.32	269	14	H3(14)	0	0	0	–
HUA*	Huanren, Liaoning, China	125.49	41.02	230	3	H3(3)	0	0	0	–
TSS	Xuyi, Jiangsu, China	118.45	32.73	97	6	H3(6)	0	0	0	–
JIA*	Jiande, Zhejiang, China	119.52	29.44	47	0	–	–	–	4	H5(4)
ZHZ	Hangzhou, Zhejiang, China	120.03	30.22	48	0	–	–	–	3	H1(3)
OKC*	Okcheon, North Chungcheong, South Korea	127.64	36.33	194	0	–	–	–	4	H3(4)
WNJ*	Wonju, Gangwon, South Korea	127.91	37.35	156	0	–	–	–	2	H3(2)
SED*	Soeda, Fukuoka, Japan	130.87	33.49	249	0	–	–	–	1	H3(1)
HIT*	Hita, Oita, Japan	130.94	33.41	261	0	–	–	–	2	H3(2)
TKD*	Taketa, Oita, Japan	131.41	32.97	321	0	–	–	–	3	H3(3)
BGO*	Bungo-ono, Oita, Japan	131.49	32.99	175	0	–	–	–	1	H3(1)
TNO*	Tsuno, Kochi, Japan	133.02	33.45	494	0	–	–	–	2	H3(2)
TSY*	Kochi, Kochi, Japan	133.51	33.64	197	0	–	–	–	1	H3(1)
OTY*	Otoyo, Kochi, Japan	133.68	33.84	427	0	–	–	–	1	H3(1)
BCU*	Takahashi, Okayama, Japan	133.39	34.83	317	0	–	–	–	1	H3(1)
NIM*	Niimi, Okayama, Japan	133.52	34.89	281	0	_	_	_	2	H3(2)
YSR*	Miyoshi, Tokushima, Japan	133.75	33.96	216	0	_	_	_	1	H3(1)

(Continued)

TABLE 1 | Continued

Sampling sites	Location	Lon (E)	Lat (N)	E (m)	Wild trees				Cultivated trees	
					n_w	Haplotypes	H_d	π	n_c	Haplotypes
YKN*	Fukuchiyama, Kyoto, Japan	134.94	35.32	154	0	–	–	–	1	H3(1)
NYN*	Gojo, Nara, Japan	135.73	34.29	193	0	–	–	–	2	H3(2)
SNI*	Soni, Nara, Japan	136.14	34.50	592	0	–	–	–	2	H3(2)
MSG*	Tsu, Mie, Japan	136.27	34.55	227	0	–	–	–	2	H3(2)
SKW*	Shirakawa, Gifu, Japan	136.91	36.26	570	0	–	–	–	1	H3(1)
HID*	Hida, Gifu, Japan	137.22	36.34	949	0	–	–	–	2	H3(2)
WJM*	Wajima, Ishikawa, Japan	136.89	37.33	123	0	–	–	–	3	H3(3)
SZU*	Suzu, Ishikawa, Japan	137.14	37.40	213	0	–	–	–	1	H3(1)
MNM*	Minakami, Gunma, Japan	138.99	36.70	516	0	–	–	–	1	H3(1)
KKR*	Kamakura, Kanagawa, Japan	139.51	35.31	14	0	–	–	–	1	H3(1)
OGN*	Oguni, Yamagata, Japan	139.81	38.09	209	0	–	–	–	2	H3(2)
AZW*	Aizuwakamatsu, Fukushima, Japan	139.97	37.51	405	0	–	–	–	1	H3(1)
DIG*	Daigo, Ibaraki, Japan	140.40	36.70	145	0	–	–	–	1	H3(1)
OWN*	Owani, Aomori, Japan	140.53	40.49	121	0	–	–	–	1	H3(1)
AMS*	Aomori, Aomori, Japan	140.67	40.78	76	0	–	–	–	2	H3(2)
SNG*	Shingo, Aomori, Japan	141.18	40.43	158	0	–	–	–	1	H3(1)
SDI*	Sendai, Miyagi, Japan	140.85	38.26	59	0	–	–	–	1	H3(1)
JBJ*	Ninohe, Iwate, Japan	141.18	40.16	435	0	–	–	–	1	H3(1)
INH*	Ichinohe, Iwate, Japan	141.31	40.20	197	0	–	–	–	3	H3(3)
ABS*	Abashiri, Hokkaido, Japan	144.25	44.01	145	0	–	–	–	2	H3(2)

Lon, longitude; Lat, latitude; E, elevation; n_w, number of wild trees; H_d, haplotype diversity; π, nucleotide diversity; n_c, number of cultivated trees.
**Sequences were obtained by Suzuki et al. (2014); **Sequences were obtained by both Suzuki et al. (2014) and this study.*

between groups of populations. One hundred independent runs were carried out for each number of groups (K) ranging from 2 to 10 to ensure that the final configuration of the K groups is not affected by a given initial configuration. For the most likely K, the significance of variance components (overall genetic variance partitioned among groups, among populations within groups, and within populations) and their associated fixation indices (F_{CT}, F_{SC}, and F_{ST}) was assessed by 10,000 random permutations using Arlequin 3.5 (Excoffier et al., 2005). To validate the results of SAMOVA, we also performed principal coordinate analysis (PCoA) based on the matrix of population pairwise F_{ST} using GenALEx 6.5 (Peakall and Smouse, 2012). To test for the isolation by distance (IBD) pattern, we examined the correlation between population pairwise F_{ST} and the logarithm of geographic distance (km) using a Mantel test. The significance of the correlation was assessed by 9,999 permutations in GenALEx 6.5. In these analyses, we only removed the cultivated trees in China for the reason mentioned above. The final dataset comprised 385 trees from 75 sampling sites, including 333 trees from 43 sites in China, six trees from two sites in South Korea, and 46 trees from 30 sites in Japan (**Supplementary Table S2**). We used BEAST 2.6.7 (Bouckaert et al., 2019) to estimate the divergence time between the eastern and western clades of *T. vernicifluum*. The phylogenetically closest species in the median-joining network, *T. radicans*, was used as an outgroup. The best-fitting substitution model HKY was selected by ModelFinder (Kalyaanamoorthy et al., 2017). A combination of strict clock and Bayes-skyline coalescent

prior was used for node age estimation. The mean value of cpDNA substitution rate for angiosperms (2.0×10^{-9} substitutions per site per year) was employed (Wolfe et al., 1987; Sakaguchi et al., 2012). Two independent MCMC runs were performed for 100 million generations and sampled every 10,000 generations. Tree and log files were combined through LogCombiner 2.6.7, and then passed to Tracer 1.7.1 (Rambaut et al., 2018) for assessing convergence, and to TreeAnnotator 2.6.7 for constructing a maximum clade credibility tree with a posterior probability limit of 0.5 and the first 20% generations discarded as burn-in.

Ecological Niche Modeling and Niche Identity Test

We used the maximum-entropy approach in Maxent 3.4.1 (Phillips et al., 2018) to model the present distribution of *T. vernicifluum* and to reconstruct its potential distribution during the Last Interglacial [LIG; ~0.12–0.14 million years ago (mya)] and the Last Glacial Maximum (LGM; ~0.022 mya). Species occurrence data were obtained from five sources: the Global Biodiversity Information Facility (GBIF),[2] the Chinese Virtual Herbarium (CVH),[3] the Plant Photo Bank of China (PPBC),[4] literature (Wei et al., 2010; Bai et al., 2017; Guo et al., 2019), and field investigation. The Getpoint tool of Baidu Maps[5] was used to collect coordinates for the specimen records only with explicit locality information. We filtered our dataset by removing duplicate records and retaining only one observation within each $2.5' \times 2.5'$ grid to reduce the effect of spatial autocorrelation. Finally, a total of 394 presence points were obtained, of which 307, 32, and 46 records were from China, Japan, and South Korea, respectively. We retrieved the climatic data for the present and LGM from the WorldClim 1.4 database[6] at a spatial resolution of 2.5′. The LGM data were generated based on the outputs of the Community Climate System Model 4 (CCSM4). The raster layers of the LIG were obtained from the WorldClim 1.4 at a spatial resolution of 30″ and then resampled to 2.5′ *via* the nearest neighbor method as implemented in ArcGIS 10.5. We eliminated highly correlated variables ($|\text{Pearson's } r| \geq 0.8$) to prevent potential over-fitting. Finally, six of the 19 bioclimatic variables (**Supplementary Table S3**) provided by the WorldClim database were retained, including annual mean temperature (bio1), mean diurnal range (bio2; the mean of the difference of the monthly maximum and minimum temperatures over a year), isothermality (bio3), temperature seasonality (bio4), mean temperature of the wettest quarter (bio8), and annual precipitation (bio12). All the environmental layers were clipped to the same spatial range (15°–45°N, 90°–145°E) using the package "raster" 2.8-19 (Hijmans, 2019) in R 3.6.0 (R Core Team, 2018). We ran the Maxent model with default settings. Model performance was assessed using the areas under the receiver operating characteristic curve (AUC) produced by 10-fold cross-validation. The generated model was projected onto the two historical periods and the predicted suitable areas were visualized based on the logistic outputs of Maxent using ArcGIS 10.5. Principal component analysis (PCA) was performed with the 19 bioclimatic variables provided by the WorldClim database for all the presence points of *T. vernicifluum* using R 3.6.0.

We used ENMTools (Warren et al., 2010) to quantify the niche overlap between the SDMs generated for the western and eastern populations of *T. vernicifluum*. Two statistics, Schoener's *D* (Schoener, 1968) and Warren's *I* (Warren et al., 2008), were used to measure the niche overlap. These two indices are limited between 0 (the two groups have a completely discordant niche) and 1 (the two groups have an identical niche). To perform the niche identity test, first, we used the same program to create a pseudoreplicate dataset by randomly partitioning the pooled occurrence points for the two groups into two new sub-datasets with the original sample size (i.e., 457 and 54). Then, the new dataset was imported to Maxent to generate new SDMs using the default settings. Finally, we calculate Schoener's *D* and Warren's *I* for SDMs generated by 100 pseudoreplicate datasets. The niche identity was tested by comparing the observed values and the null distributions for these two statistics.

Migration Vector Analysis and Dispersal Corridors

To visualize the migration direction of *T. vernicifluum* between different periods, we performed a migration vector analysis following Gugger et al. (2013). First, the logistic outputs of Maxent were converted into presence/absence maps using the 'maximum test sensitivity plus specificity' threshold (Jiménez-Valverde and Lobo, 2007). Second, we estimated the geographic centroids of all the $0.625° \times 0.625°$ grids (i.e., each grid contained 15×15 grid cells in 2.5′) for each period using the zonal geometry function in ArcGIS 10.5. Finally, we inferred migration vectors by seeking the nearest centroids of the second period to each centroid of the first period using ArcGIS 10.5 (i.e., from LIG to LGM and from LGM to present). The obtained maps showed the potential population sources from one period to the next.

We also integrated SDMs and shared haplotype information to infer the putative dispersal corridors across the landscape (Chan et al., 2011). First, we inverted the logistic outputs of Maxent (x inverted $= 1 - x$) to create a friction layer (i.e., a dispersal cost layer), which depicted the ease of dispersal from each locality through the landscape. Second, we calculated a single least-cost path (LCP) and multiple least-cost corridors (LCCs) for each pair of sampling sites that shared haplotypes (Graves et al., 2014; Yu et al., 2015). Least-cost corridors were classified into three categories according to the percentage by which the path length was greater than that of the LCP: low (<1.0%), mid (<2%), and high (<5%). Finally, we summed all the LCCs to create a raster of the dispersal network. The low, mid, and high classes were weighted by 5, 2, and 1, respectively. Dispersal corridors were expected to be the areas where LCCs

[2]http://www.gbif.org/
[3]http://www.cvh.ac.cn/
[4]http://ppbc.iplant.cn/
[5]https://api.map.baidu.com/lbsapi/getpoint/index.html
[6]http://worldclim.org

traversed more frequently. In this analysis, we removed the records of cultivated trees in mainland China, including four sampling sites (i.e., YAN, JUL, JIA, and ZHZ) that were only composed of cultivated trees. Furthermore, only one site was chosen within each $0.625° \times 0.625°$ grid to reduce both computational cost and sampling bias. Finally, a total of 62 localities (**Supplementary Table S2**) were retained to infer the dispersal corridors of *T. vernicifluum* during the present, LGM, and LIG using SDMtoolbox 2.0 (Brown, 2014).

RESULTS

Phylogenetic Relationship and Geographic Distribution of Chloroplast Haplotypes

The combination of the *trnL* and *trnL-F* datasets resulted in an alignment with a total length of 717 bp. Based on seven substitutions, five chloroplast haplotypes (H1–H5) were identified across the 79 sampling sites of *T. vernicifluum* (**Supplementary Table S4**). The median-joining network grouped the haplotypes into two major lineages separated by three mutational steps (**Figure 1B**). Among those, the western clade was mainly detected in western China, including two haplotypes H1 and H2, while the eastern clade was mostly distributed in eastern China, South Korea, and Japan, comprising three haplotypes H3, H4, and H5 (**Figure 1A**). Exceptions were the natural occurrence of H4 in one western China site XDS, and the introduction of the trees with H3 to western China (i.e., sites JUL, BYS, YUN, and FOP) and those with H1 to eastern China (i.e., site ZHZ; **Figure 1A**).

Among the five haplotypes, H1 and H3 were at the interior of the network and were represented by 65.0 and 31.4% of the surveyed samples, respectively (**Figure 1B**). Wild trees with H1 were detected in all the 37 sampling sites of western China, while the haplotype H3 was observed in 40 of the 42 sites in eastern China, South Korea, and Japan. In contrast, H2 and H5 were found at the tips of the network, separated from H1 or H3 by one mutational step (**Figure 1B**). They had a much lower frequency, of which H2 was only detected in two sampling sites (i.e., MEI and ZHA) located at the Qinling Mountains, western China, and H3 occurred in four sites of eastern China (i.e., ZAO, SZB, AHH, and JIA). The haplotype H4 was found at an intermediate position between H1 and H3, represented by only one individual in the sampling site XDS. All five haplotypes were detected in mainland China. Most Chinese sites only had one haplotype, while 10 sites (i.e., YAN, FOP, MEI, ZHA, XDS, BYS, YUN, AHH, ZAO, and SZB) had two haplotypes. All the Korean and Japanese sites were fixed for the haplotype H3 (**Figure 1A**).

Chloroplast DNA Diversity, Differentiation, and Phylogeographic Structure

When the cultivated trees in mainland China were excluded, the number of haplotype, haplotype diversity (H_d), and nucleotide diversity (π) of each sampling site ranged from 1 to 2 (mean = 1.127), 0 to 0.4 (mean = 0.042), and 0 to 0.00084 (mean = 0.00007), respectively (**Table 1**). The highest level of haplotype diversity was observed in AHH ($H_d = 0.4$), followed by ZHA ($H_d = 0.356$) and SZB ($H_d = 0.327$), while the highest level of nucleotide diversity was observed in XDS ($\pi = 0.00084$), followed by AHH ($\pi = 0.00056$), and ZHA ($\pi = 0.0005$). The total gene diversity ($h_T = 0.484 \pm 0.049$) across all sampling sites was found to be much higher than average gene diversity within sites ($h_S = 0.033 \pm 0.014$; **Table 2**).

Genetic differentiation among sampling sites was substantial as indicated by the high values of G_{ST} and N_{ST}. Comparisons of these two measures showed that a significant phylogeographic structure occurred across the species' range ($N_{ST} = 0.983 > G_{ST} = 0.933$; $p = 0.029$) or across the sampling sites in mainland China ($N_{ST} = 0.972 > G_{ST} = 0.891$; $p = 0.027$; **Table 2**). SAMOVA revealed a high level of differentiation among groups ($F_{CT} > 0.980$) for all the K values from 2 to 10. When $K = 2$, the sampling sites of *T. vernicifluum* were divided into two groups (**Figure 1C**). The western group included the 35 sites in western China, while the eastern group included the 40 sites in eastern China, South Korea, and Japan. These two groups were separated clearly by the boundary between the middle and low units of the three-step landforms of China. Hierarchical AMOVA showed that 98.78% of the total genetic variance was partitioned between these two groups ($F_{CT} = 0.988$, $p = 0.000$), while only 0.24% and 0.97% of the variance were partitioned among sampling sites within groups ($F_{SC} = 0.199$, $p = 0.042$) and within sites ($F_{ST} = 0.990$, $p = 0.000$), respectively (**Table 3**). PCoA obtained a result consistent with that of SAMOVA ($K = 2$). The first axis explained 88.12% of the total variance (**Figure 1D**). No overlap along this axis was detected between the individuals of the western and eastern groups. IBD analyses showed that the geographic isolation effect was significant across all sampling sites of *T. vernicifluum* ($R = 0.682$, $p = 0.001$) or across those in mainland China ($R = 0.610$, $p = 0.000$). The BEAST analysis also grouped the haplotypes into western and eastern clades (posterior probability = 1). The divergence time between these two clades was estimated to be 1.36 mya (95% HPD; 0.51–2.41 mya; **Supplementary Figure S1**).

Ecological Niche Modeling and Niche Identity Test

The mean AUC value (±SD) was 0.895 ± 0.009, indicating that the SDM fitted well with the observed dataset. The predicted potential distributions of *T. vernicifluum* were shown in **Figure 2**. At present, the suitable habitats of *T. vernicifluum* are mainly distributed in western China, the Korean Peninsula, and Japan, encompassing a large number of the presence points (**Figure 2E**). The highly suitable habitats in China are mainly found in the mountainous areas of the "middle step" region, including the Yunnan-Guizhou Plateau, Qinling Mountains, and Daba mountains. During the LGM, the suitable habitats were widely distributed in western and southern China and extended to the offshore areas of the East China Sea Shelf and Japan (**Figure 2C**). During the LIG, the suitable areas were mainly restricted to southwestern China (**Figure 2A**). Both Schoener's D and Warren's I indicated that significant climatic niche divergence occurred between the western and eastern groups

TABLE 2 | Genetic statistics of *Toxicodendron vernicifluum* based on the sequence variation at two chloroplast (cp) DNA markers.

Sampling sites	Sample size	h_T (SE)	h_S (SE)	G_{ST} (SE)	N_{ST} (SE)	*p*-Value
Sites in mainland China	327	0.361 (0.081)	0.039 (0.017)	0.891 (0.044)	0.972 (0.013)	0.027*
Sites across the species' range	353	0.484 (0.049)	0.033 (0.014)	0.933 (0.028)	0.983 (0.008)	0.029*

The samples used in these analyses are described in ***Supplementary Table S2****. h_S, average gene diversity within sampling sites; h_T, total gene diversity; G_{ST}, genetic differentiation at the two cpDNA markers; N_{ST}, genetic differentiation at the two cpDNA markers taking similarities between haplotypes into account; SE, standard error.*
**$p < 0.05$, indicating that N_{ST} is significantly larger than G_{ST}.*

TABLE 3 | Hierarchical analyses of molecular variance (AMOVAs) based on chloroplast (cp) DNA haplotype frequencies of *Toxicodendron vernicifluum*.

Source of variation	df	SS	VC	Variation (%)	Fixation index
Among groups	1	413.19	2.49	98.78	$F_{CT} = 0.988$**
Among populations within groups	73	4.05	0.01	0.24	$F_{SC} = 0.199$*
Within populations	310	7.61	0.02	0.97	$F_{ST} = 0.990$**

The samples used in this analysis are described in ***Supplementary Table S2****. df, degree of freedom; SS, sum of squares; VC, variance components. p-value was obtained through 10,000 permutations in ARLEQUIN.*
***$p < 0.01$; *$p < 0.05$.*

of *T. vernicifluum*, regardless of whether the sampling sites of South Korea and Japan were included (Schoener's $D = 0.557$, $p = 0.000$; Warren's $I = 0.839$, $p = 0.000$) or not (Schoener's $D = 0.472$, $p = 0.000$; Warren's $I = 0.756$, $p = 0.000$; **Figure 3**). Similar results were also obtained for the LIG and LGM (**Figure 3**). PCA plot showed that the western group was associated with higher isothermality, higher precipitation seasonality, lower precipitation of the driest month, lower precipitation of the driest quarter, and lower precipitation of the coldest quarter (**Figure 4**).

Migration Vector Analysis and Dispersal Corridors

The putative migration directions of *T. vernicifluum* between adjacent periods were shown in **Figure 5**. Suitable areas were inferred to have expanded continuously since the LIG. From the LIG to LGM, the suitable areas in southwestern China may have expanded northward or eastward to the mountainous areas in central China (e.g., the Qinling Mountains and the Funiu Mountains). Notably, some scattered suitable habitats in southern and southeastern China (e.g., the Nanling Mountains and the Wuyi Mountains) may have also expanded northward drastically. Furthermore, both the suitable areas in the Taiwan Island and the Korean Peninsula may have extended to the East China Sea Shelf during the LGM. From the LGM to the present, *T. vernicifluum* was predicted to have spread further northward. They may have migrated from the Funiu Mountains to the Taihang Mountains, from the Shandong and Korean Peninsulas to northeastern China, and from central Japan to Hokkaido.

No population connectivity was found between western and eastern China because the wild trees of western and eastern groups did not share any haplotypes (**Figure 2**). Instead, two major dispersal routes were identified in the eastern and western parts of the range of *T. vernicifluum*. High landscape connectivity was detected among populations in northeastern China, the Korean Peninsula, and Japan. The ECS land bridge may have contributed to the migration to Japan during the LGM. In western China, dispersal corridors occurred around the mountain ranges east of the Sichuan Basin, including the Taihang Mountains, the Qinling Mountains, the Wu Mountain, and the Xuefeng Mountains (**Figure 2**).

DISCUSSION

East-West Phylogeographic Split Associated With the Stepped Geomorphology of China

Our chloroplast DNA analyses support an east-west phylogeographic split of *T. vernicifluum* (**Figure 1**). The two recognized clades were separated according to the stepped landforms of mainland China, with the wild individuals of the western clade geographically restricted to the "middle step", while those of the eastern clade mainly confined to the "low step". The two "steps" differ in geomorphology; the "middle step" is characterized by a vast extent of mountains and plateaus (average ~2,000 m), while the "low step" is mainly made up of hills and plains (average <500 m; Jiang and Wu, 1993; Wan, 2012). Moreover, a series of northeast-southwest oriented mountain ranges (e.g., the Taihang Mountains, the Wu Mountains, and the Xuefeng Mountains) occur on the border between the two "steps", further increasing the ruggedness of local terrain (Jiang and Wu, 1993; Li et al., 2015). The existence of these long-standing geographic barriers, together with the extreme physiographical heterogeneity in mainland China, may have strongly restricted the dispersal of *T. vernicifluum* across the landscape, leading to long-term isolation and allopatric genetic divergence between western and eastern lineages of *T. vernicifluum* (Qian and Ricklefs, 2000; Zhang et al., 2018; Luo et al., 2021).

We inferred that the separation between the two major lineages may have occurred during the Early Pleistocene (1.36 mya, 95% HPD: 0.51–2.41 mya; **Supplementary Figure S1**). This timing is comparable with the intraspecific divergence date of *Quercus acutissima* (1.31 mya; 95% HPD: 1.27–1.34 mya; Gao et al., 2021), a species also exhibiting an east-west

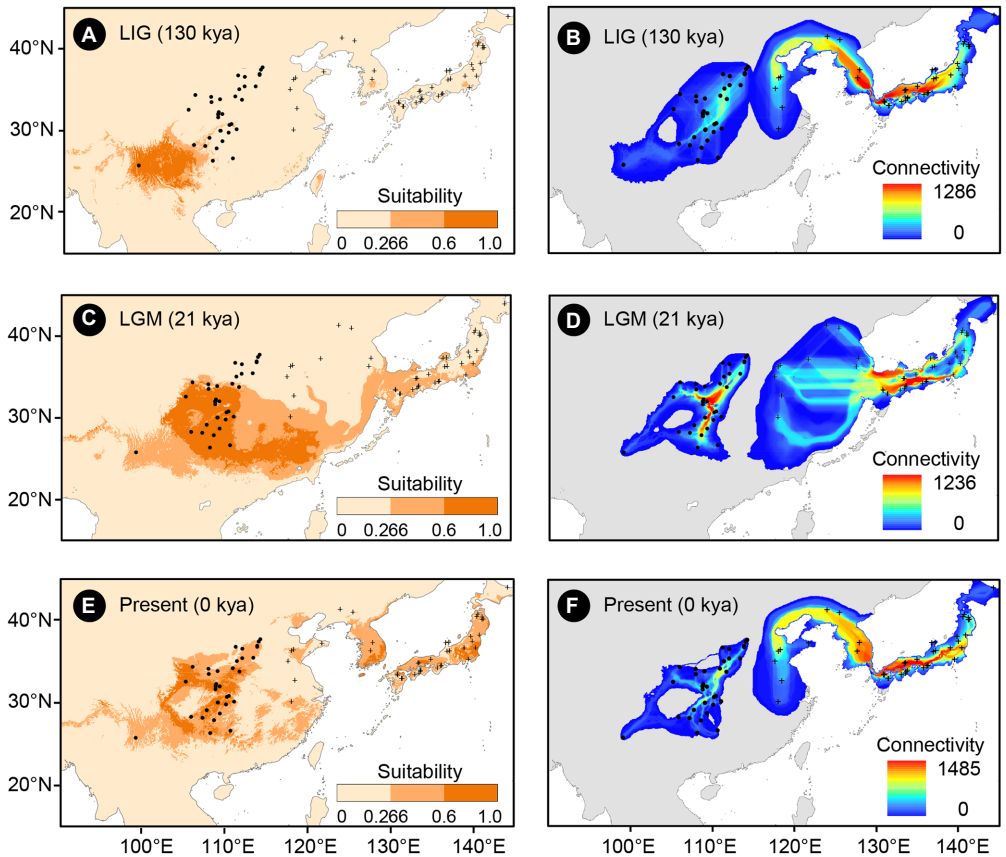


FIGURE 2 | Climatically suitable areas and dispersal corridors of *Toxicodendron vernicifluum* during the Last Interglacial (LIG; **A,B**), Last Glacial Maximum (LGM; **C,D**), and the present (**E,F**) based on the outputs of ecological niche modeling (ENM) using Maxent 3.4.1 (Phillips et al., 2018). Black dots and plus signs represent the sampling sites of the western and eastern groups, respectively.

break associated with the stepped landforms of China. These results suggest that they may have undergone a similar biogeographic history affected by the landscape features in East Asia and climatic oscillations during the Early Pleistocene. The cool climate may have caused the ancestral population to retreat to different refugia, leading to allopatric divergence in response to long-term geographic isolation. At present, the western clade of *T. vernicifluum* was mainly observed in highlands (between 800 and 2,000 m) with higher isothermality and lower precipitation of the driest quarter, while the eastern clade usually occurs in lowlands (less than 600 m) with opposite climatic conditions (**Figure 4**). The climate niche divergence was estimated to have occurred no later than the LIG (**Figure 3**), suggesting that local adaptation may have also contributed to the splitting of *T. vernicifluum*. However, our dating results must be interpreted with extreme caution because of the wide range of divergence time and uncertainty in mutation rate (Luo et al., 2021; Li et al., 2022). Nonetheless, our data depict the most likely scenario that the two major clades have diverged before the LGM, which would provide more opportunities for those lineages to adapt to distinct climatic conditions.

Previous studies have revealed an east-west phylogeographic break across the boundary between the middle and low "steps" in three widespread woody plants, including *K. septemlobus* (Sakaguchi et al., 2012), *K. japonica* (Luo et al., 2021), and *Q. acutissima* (Zhang et al., 2018). For the two former species, the pattern was confirmed by both nuclear microsatellite and chloroplast sequence markers, although limited admixture was detected near the border (Sakaguchi et al., 2012; Luo et al., 2021). For the last species, the pattern was revealed only by nuclear microsatellites (Zhang et al., 2018). A high level of genetic admixture was observed in central China, probably because wind-pollinated oak species have greater long-distance pollen-mediated gene flow than the other two insect-pollinated plants (Buschbom et al., 2011). In subtropical China, similar patterns with limited sharing of genetic variation between the two "step" were observed in several woody and herbaceous plants, such as *J. cathayensis* (Bai et al., 2014), *Castanopsis eyrei* (Shi et al., 2014), *Castanopsis fargesii* (Sun et al., 2014), *Cyclocarya paliurus* (Kou et al., 2016), *B. clarkeana* (Wang et al., 2018), and *L. chinense* (Yang et al., 2019; Zhong et al., 2019). In comparison with those species, *T. vernicifluum* exhibits a much sharper east-west phylogeographic break; no haplotypes were found to be shared among wild populations from the two sides (**Figure 1A**). Furthermore, the break not only occurs in subtropical areas (e.g., across the Xuefeng Mountain and

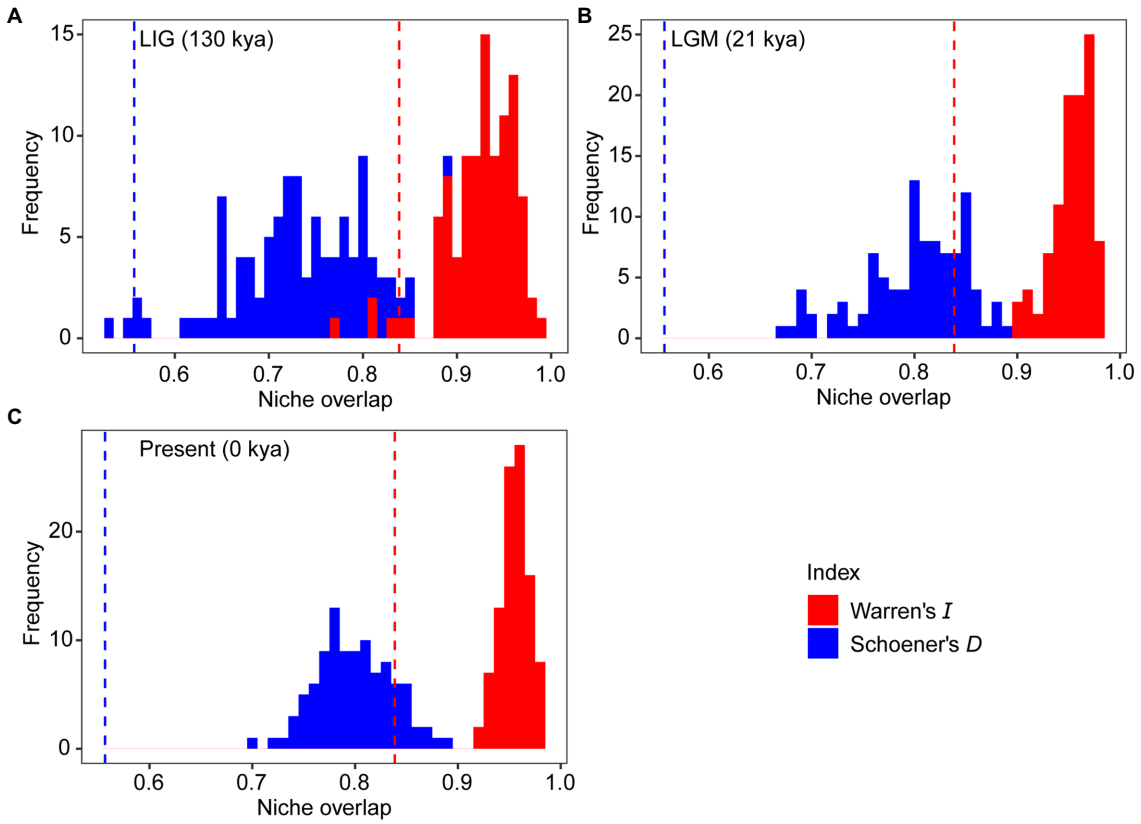


FIGURE 3 | Niche divergence between the western and eastern groups of *Toxicodendron vernicifluum* during the Last Interglacial (LIG; **A**), Last Glacial Maximum (LGM; **B**), and present **(C)**. The niche identity test was conducted to quantify niche overlap using two indices, Warren's *I* (red) and Schoener's *D* (blue). Vertical lines and histograms represent the observed values and null distributions (based on 100 pseudoreplicates) of niche overlap statistics, respectively.

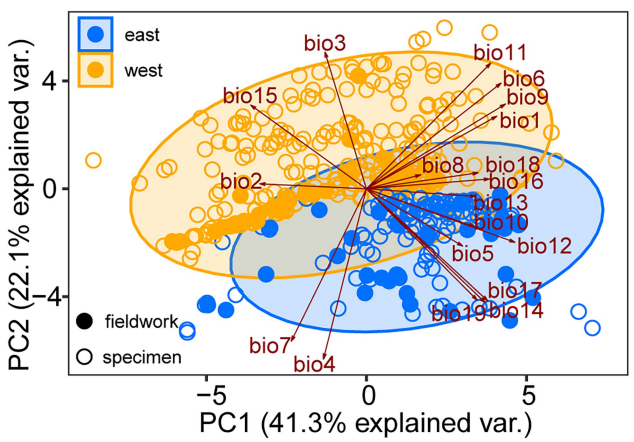


FIGURE 4 | Principal component analysis (PCA) performed with 19 bioclimatic variables (**Supplementary Table S3**) for the specimen records (open dots) and sampling sites (solid dots) of *Toxicodendron vernicifluum*. Presence points of the western and eastern groups are marked in orange and blue colors, respectively.

the Wu Mountain) but also extends to warm-temperate regions that were not fully covered by previous studies, such as the Taihang Mountains (**Figure 1A**).

Another remarkable feature of the present phylogeographic pattern is that more than 90% of the sampling sites are fixed for a single haplotype (**Figure 1A**). The near-complete fixation of two major haplotypes (H1 and H3) in western vs. eastern China resulted in an extremely low level of genetic diversity within populations ($h_S = 0.033$) and an extremely high level of genetic differentiation among populations ($G_{ST} = 0.933$; **Table 2**). Such patterns are comparable with those of previously mentioned subtropical plants such as *J. cathayensis* ($h_S = 0$, $G_{ST} = 1$; Bai et al., 2014), *C. paliurus* ($h_S = 0.066$, $G_{ST} = 0.924$; Kou et al., 2016), and *B. clarkeana* ($h_S = 0.030$, $G_{ST} = 0.964$; Wang et al., 2018), suggesting that they have experienced a common phylogeographic history of long-term isolation across fragmented mountainous habitats in subtropical China (Wang et al., 2009; Du et al., 2011; Bai et al., 2014; Li et al., 2019). *Toxicodendron vernicifluum* prefers to grow in highlands and thus exhibits a scattered distribution in mountainous areas. Within a separated region, strong forces of genetic drift, combined with low mutation and low migration rates of cpDNA sequences, would greatly reduce the genetic diversity within populations (Bai et al., 2014). Furthermore, as a dioecious plant, *T. vernicifluum* is predicted to be more sensitive to genetic drift and tend to show a higher level of differentiation as it has a smaller effective population size than hermaphrodite species (McCauley, 1994). This prediction has been verified in

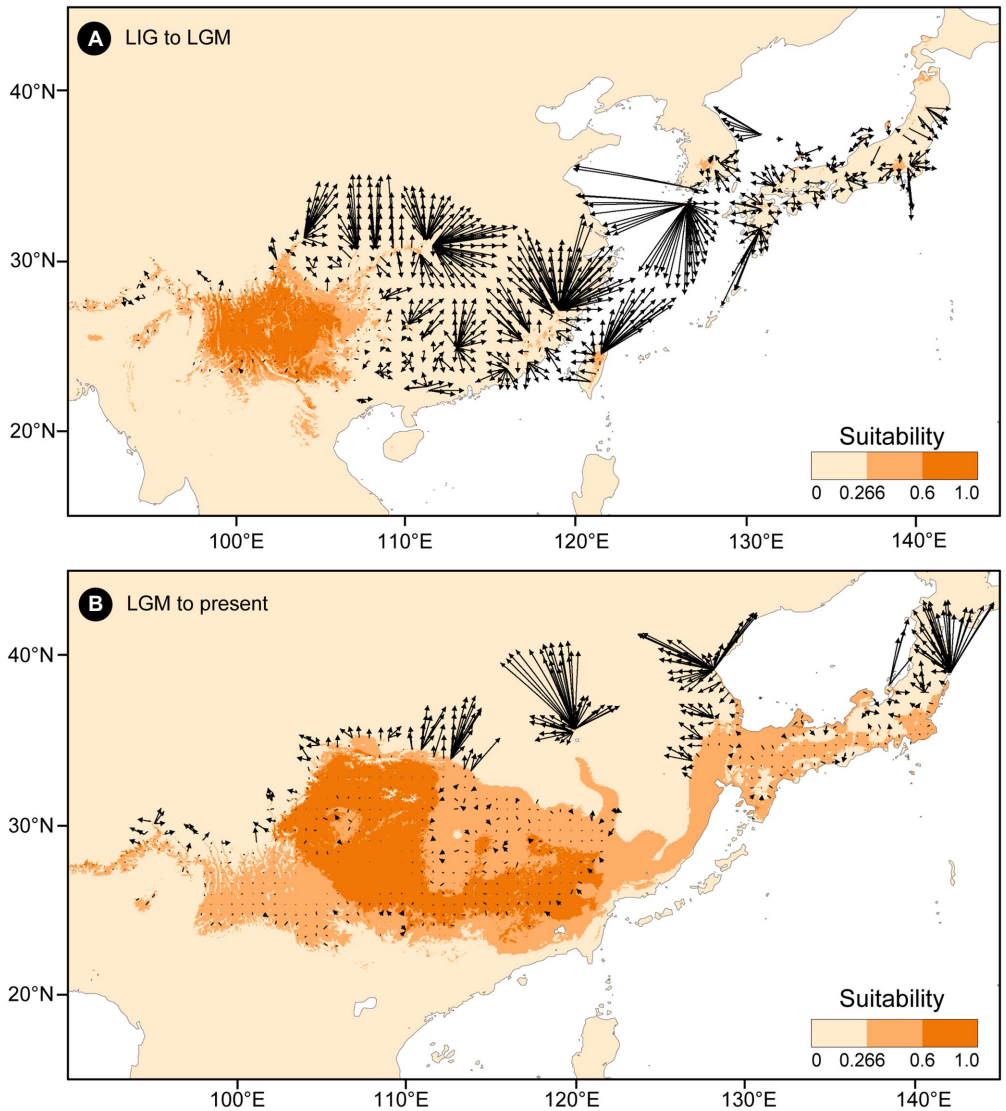


FIGURE 5 | Migration vector analysis of local changes in climatically suitable areas of *Toxicodendron vernicifluum* between the Last Interglacial (LIG) and Last Glacial Maximum (LGM; **A**), and between the LGM and present **(B)**. Black arrows represent the potential migration direction from one period to the next.

a dioecious shrub, *Ilex aquifolium*, which exhibited twice the level of differentiation in comparison with hermaphrodite species investigated over a similar geographical scale (Rendell and Ennos, 2003). Overall, we suggest that both abiotic (e.g., long-term geographic isolation) and biotic (e.g., a dioecious breeding system) factors have contributed to the occurrence of a pronounced phylogeographic structure in *T. vernicifluum*.

Glacial Refugia, Historical Migration, and Human-Aided Dispersal

The present phylogeographic structure of *T. vernicifluum* suggests that the species has experienced long-term isolation between at least two refugia in western and eastern China (Bai et al., 2014; Liao et al., 2014). However, the exact locations of refugia are uncertain because each of the two areas is dominated by an ancestral haplotype (H1 and H3) that exhibits a much wider distribution than others (Posada and Crandall, 2001). Combining the prediction of ENMs, we infer that the mountainous areas in southwestern (i.e., the Yunnan-Guizhou Plateau) and southeastern China are more likely to be glacial refugia (**Figure 2**). These regions were believed to have provided relatively stable climatic conditions for the *in situ* survival of local plants, such as *Eurycorymbus cavaleriei* (Wang et al., 2009), *J. cathayensis* (Bai et al., 2014), and *C. paliurus* (Kou et al., 2016). Furthermore, Petit et al. (2003) pointed out that populations in refugia not only have relatively high genetic diversity but also contain some unique haplotypes. In our study, it is true for the wild populations MEI, ZHA, and XDS, supporting that the Qinling Mountains, the Wuling Mountains, and the Wu Mountains may have been refugia in western China (e.g., Gong et al., 2008; Wang et al., 2009; Deng et al., 2019).

Our migration vector analyses indicate that *T. vernicifluum* may have experienced a gradual range expansion since the LIG (**Figure 5**). The northward colonization would provide opportunities for the widespread of a few ancestral haplotypes (H1 and H3) in western and eastern China (Tian et al., 2015). Numerous mountain ranges, especially those near the border between the middle and low "step", are predicted to be dispersal corridors that have increased the landscape connectivity among populations (Guan et al., 2016). Furthermore, we found that the ESC land bridge may have also acted as a dispersal corridor during the LGM (**Figure 2D**), as supported by the evidence that Japanese samples share an ancestral haplotype (H3) with those from eastern China (**Figure 1A**). Such a finding is consistent with that of previous studies showing that there is a close genetic relationship between populations from eastern China and Japan (e.g., *Quercus variabilis*, Chen et al., 2012; *K. septemlobus*, Sakaguchi et al., 2012; *Machilus thunbergii*, Jiang et al., 2021). However, this pattern does not occur in all cases because the submergence of the ECS land bridge may also lead to the divergence of the lineage between eastern China and Japan (e.g., *Ligularia hodgsonii*, Wang et al., 2013; *Platycrater arguta*, Qi et al., 2014; *Euptelea* spp., Cao et al., 2016).

It should be noted that the origin of *T. vernicifluum* in Japan is still controversial (Suzuki et al., 2014). As an economically important arbor species, the lacquer tree was used for natural lacquer collection in China 8,000 years ago (Wu et al., 2018). In Japan, the oldest lacquer was found in Hokkaido 9,000 years ago. Although evidence from fossil woods supported that *T. vernicifluum* grew in middle to northern Honshu of Japan since the Early Jomon Period, it is believed that the species was introduced from China in an earlier time because it does not grow in natural forests and is only found around human settlements (Noshiro and Suzuki, 2004; Noshiro et al., 2007). However, this view is challenged by a recent finding of a fossil wood of *T. vernicifluum* dated back to the incipient Jomon period (~12,600 years ago), from the Torihama shell midden of Fukui prefecture, Japan (Suzuki et al., 2014). This fossil wood did not show evidence of artificial processing, suggesting that *T. vernicifluum* is likely to be native to Japan during the glacial periods (Suzuki et al., 2014). If it is true, we infer that the species may have an origin from eastern China because they share an ancestral haplotype with those from eastern China, and the genetic diversity in Japan is much lower than that in eastern China. Furthermore, it should be noted that our study provides evidence for the historical transplanting of lacquer trees between eastern and western China (**Figure 1A**). This finding reminds us it is necessary to use both provenance trials and genomic analyses to assess the adaptive ability of lacquer trees to different climatic conditions, which will offer guidance for the future management of *T. vernicifluum* resources.

CONCLUSION

Toxicodendron vernicifluum exhibits a clear east-west phylogeographic break associated with the stepped geomorphology of China. This break was much sharper than previously reported because no shared haplotypes were detected among the wild trees on the two sides. Furthermore, this break occurred at a much larger scale, extending from subtropical (e.g., the Xuefeng Mountain) to warm-temperate areas (e.g., the Taihang Mountain) of China. Our study supports that the eastern and western clades may have diverged during the Early Pleistocene, suggesting a likely scenario of allopatric divergence in response to long-term geographic isolation across at least two glacial refugia. Combining the evidence from fossil records and molecular analyses, we support that *T. vernicifluum* in Japan may have an origin from eastern China, while the East China Sea may have acted as a dispersal corridor during the glacial periods.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Sequence data are available on GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under the accession numbers OL355137–OL355141.

AUTHOR CONTRIBUTIONS

LW, YF, and FZ conceived and designed this research. LW, YL, SN, MS, TA, KK, LX, and MZ collected samples. LW performed experiments and wrote the original draft. LW, YL, and NH analyzed the data. YF and FZ supervised the project. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Key Research and Development Program of China (2017YFD060130501), the China Postdoctoral Science Foundation (2020M681629), the Jiangsu Postdoctoral Research Foundation (2021K038A), Shaanxi Innovation Capability Support Plan "Chinese Lacquer Tree Resource Data Information Sharing Platform" (2022PT-05), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). LW was supported by the fund from the China Scholarship Council (202008320479).

ACKNOWLEDGMENTS

We thank Sujing Fu, Jianping He, Heng Jia, Hongli Ji, Xiaoqiang Lu, Hongshan Nie, Mingyue Zang, Baoquan Zhang, Hong Zhu, and numerous Chinese foresters and natural resource managers for their help in the sample collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.920054/full#supplementary-material

REFERENCES

Arbogast, B. S., and Kenagy, G. J. (2001). Comparative phylogeography as an integrative approach to historical biogeography. *J. Biogeogr.* 28, 819–825. doi: 10.1046/j.1365-2699.2001.00594.x

Avise, J. C. (1992). Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos* 63, 62–76. doi: 10.2307/3545516

Avise, J. C. (2000). *Phylogeography: The History and Formation of Species.* Cambridge: Harvard University Press.

Bai, G. Q., Li, W. M., Chen, H., Li, B., and Li, S. F. (2017). Characteristics of molecular evolution of *Toxicodendron vernicifluum* in Qinba Mountains by nrDNA ITS and cpDNA sequence. *Bull. Bot. Res.* 37, 579–586. doi: 10.7525/j.issn.1673-5102.2017.04.014

Bai, W. N., Wang, W. T., and Zhang, D. Y. (2014). Contrasts between the phylogeographic patterns of chloroplast and nuclear DNA highlight a role for pollen-mediated gene flow in preventing population divergence in an east Asian temperate tree. *Mol. Phylogenet. Evol.* 81, 37–48. doi: 10.1016/j.ympev.2014.08.024

Bai, W. N., Wang, W. T., and Zhang, D. Y. (2016). Phylogeographic breaks within Asian butternuts indicate the existence of a phytogeographic divide in East Asia. *New Phytol.* 209, 1757–1772. doi: 10.1111/nph.13711

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Brown, J. L. (2014). SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods Ecol. Evol.* 5, 694–700. doi: 10.1111/2041-210X.12200

Bui, T. T., Vu, D. D., Dang, Q. H., Zhang, Y., and Huang, X. H. (2017). High genetic diversity and population structure of lacquer cultivars (*Toxicodendron vernicifluum*) in Shaanxi province, China revealed by SSR markers. *Res. J. Biotechnol.* 12, 14–23.

Buschbom, J., Yanbaev, Y., and Degen, B. (2011). Efficient long-distance gene flow into an isolated relict oak stand. *J. Hered.* 102, 464–472. doi: 10.1093/jhered/esr023

Cab-Sulub, L., and Álvarez-Castañeda, S. T. (2021). Climatic dissimilarity associated with phylogenetic breaks. *J. Mammal.* 102, 1592–1604. doi: 10.1093/jmammal/gyab108

Cao, Y. N., Comes, H. P., Sakaguchi, S., Chen, L. Y., and Qiu, Y. X. (2016). Evolution of East Asia's Arcto-tertiary relict *Euptelea* (Eupteleaceae) shaped by late Neogene vicariance and quaternary climate change. *BMC Evol. Biol.* 16:66. doi: 10.1186/s12862-016-0636-x

Chan, L. M., Brown, J. L., and Yoder, A. D. (2011). Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Mol. Phylogenet. Evol.* 59, 523–537. doi: 10.1016/j.ympev.2011.01.020

Chen, D., Zhang, X., Kang, H., Sun, X., Yin, S., Du, H., et al. (2012). Phylogeography of *Quercus variabilis* based on chloroplast DNA sequence in East Asia: multiple glacial refugia and mainland-migrated island populations. *PLoS One* 7:e47268. doi: 10.1371/journal.pone.0047268

Deng, T., Abbott, R. J., Li, W., Sun, H., and Volis, S. (2019). Genetic diversity hotspots and refugia identified by mapping multi-plant species haplotype diversity in China. *Israel J. Plant Sci.* 66, 136–151. doi: 10.1163/22238980-20191083

Du, F. K., Peng, X. L., Liu, J. Q., Lascoux, M., Hu, F. S., and Petit, R. J. (2011). Direction and extent of organelle DNA introgression between two spruce species in the Qinghai-Tibetan plateau. *New Phytol.* 192, 1024–1033. doi: 10.1111/j.1469-8137.2011.03853.x

Dupanloup, I., Schneider, S., and Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.* 11, 2571–2581. doi: 10.1046/j.1365-294X.2002.01650.x

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinforma.* 1, 47–50. doi: 10.1177/117693430500100003

Fan, D., Hu, W., Li, B. O., Morris, A. B., Zheng, M., Soltis, D. E., et al. (2016). Idiosyncratic responses of evergreen broad-leaved forest constituents in China to the late quaternary climate changes. *Sci. Rep.* 6:31044. doi: 10.1038/srep31044

Fan, D. M., Yue, J. P., Nie, Z. L., Li, Z. M., Comes, H. P., and Sun, H. (2013). Phylogeography of *Sophora davidii* (Leguminosae) across the 'Tanaka-Kaiyong line', an important phytogeographic boundary in Southwest China. *Mol. Ecol.* 22, 4270–4288. doi: 10.1111/mec.12388

Feliner, G. N. (2014). Patterns and processes in plant phylogeography in the Mediterranean Basin: A review. *Perspect. Plant Ecol.* 16, 265–278. doi: 10.1016/j.ppees.2014.07.002

Feng, G., Mao, L., Sandel, B., Swenson, N. G., and Svenning, J. C. (2016). High plant endemism in China is partially linked to reduced glacial-interglacial climate change. *J. Biogeogr.* 43, 145–154. doi: 10.1111/jbi.12613

Gao, J., Liu, Z. L., Zhao, W., Tomlinson, K. W., Xia, S. W., Zeng, Q. Y., et al. (2021). Combined genotype and phenotype analyses reveal patterns of genomic adaptation to local environments in the subtropical oak *Quercus acutissima*. *J. Syst. Evol.* 59, 541–556. doi: 10.1111/jse.12568

Gao, L. M., Möller, M., Zhang, X. M., Hollingsworth, M. L., Liu, J., Mill, R. R., et al. (2007). High variation and strong phylogeographic pattern among cpDNA haplotypes in *Taxus wallichiana* (Taxaceae) in China and North Vietnam. *Mol. Ecol.* 16, 4684–4698. doi: 10.1111/j.1365-294X.2007.03537.x

Geffen, E. L. I., Anderson, M. J., and Wayne, R. K. (2004). Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Mol. Ecol.* 13, 2481–2490. doi: 10.1111/j.1365-294X.2004.02244.x

Gong, W., Chen, C., Dobeš, C., Fu, C. X., and Koch, M. A. (2008). Phylogeography of a living fossil: Pleistocene glaciations forced *Ginkgo biloba* L.(Ginkgoaceae) into two refuge areas in China with limited subsequent postglacial expansion. *Mol. Phylogenet. Evol.* 48, 1094–1105. doi: 10.1016/j.ympev.2008.05.003

Graves, T., Chandler, R. B., Royle, J. A., Beier, P., and Kendall, K. C. (2014). Estimating landscape resistance to dispersal. *Landsc. Ecol.* 29, 1201–1211. doi: 10.1007/s10980-014-0056-5

Guan, B. C., Chen, W., Gong, X., Wu, T., Cai, Q. Y., Liu, Y. Z., et al. (2016). Landscape connectivity of *Cercidiphyllum japonicum*, an endangered species and its implications for conservation. *Ecol. Inform.* 33, 51–56. doi: 10.1016/j.ecoinf.2016.04.002

Gugger, P. F., Ikegami, M., and Sork, V. L. (2013). Influence of late quaternary climate change on present patterns of genetic variation in valley oak, *Quercus lobata* Née. *Mol. Ecol.* 22, 3598–3612. doi: 10.1111/mec.12317

Guo, J. H., Liu, S. J., Wu, Y. H., Guo, Y. K., Wang, Z., and Wang, Y. L.. (2019). Genetic structure of *Toxicodendron vernicifluum* in Southern Shanxi Province based on SSR marker. *Mol. Plant Breed.* 17, 2950–2955. doi: 10.13271/j.mpb.017.002950

Guo, Z. T., Sun, B., Zhang, Z. S., Peng, S. Z., Xiao, G. Q., Ge, J. Y., et al. (2008). A major reorganization of Asian climate by the early Miocene. *Clim. Past* 4, 153–174. doi: 10.5194/cp-4-153-2008

Guo, X. D., Wang, H. F., Bao, L., Wang, T. M., Bai, W. N., Ye, J. W., et al. (2014). Evolutionary history of a widespread tree species A cer mono in East Asia. *Ecol. Evol.* 4, 4332–4345. doi: 10.1002/ece3.1278

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hashida, K., Tabata, M., Kuroda, K., Otsuka, Y., Kubo, S., Makino, R., et al. (2014). Phenolic extractives in the trunk of *Toxicodendron vernicifluum*: chemical characteristics, contents and radial distribution. *J. Wood Sci.* 60, 160–168. doi: 10.1007/s10086-013-1385-8

He, N., Wang, L., Li, Y., Fang, Y., and Zhang, F. (2020). The complete chloroplast genome sequence of *Toxicodendron sylvestre* (Anacardiaceae). *Mitochondrial DNA B Resour.* 5, 2008–2009. doi: 10.1080/23802359.2020.1756960

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. Lond. B Bio. Sci.* 359, 183–195. doi: 10.1098/rstb.2003.1388

Hickerson, M. J., Carstens, B. C., Cavender-bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., et al. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Mol. Phylogenet. Evol.* 54, 291–301. doi: 10.1016/j.ympev.2009.09.016

Hijmans, R. J. (2019). Introduction to the 'raster' package (version 2.8-19).

Hu, L. J., Uchiyama, K., Shen, H. L., Saito, Y., Tsuda, Y., and Ide, Y. (2008). Nuclear DNA microsatellites reveal genetic variation but a lack of phylogeographical structure in an endangered species, *Fraxinus mandshurica*, across north-East China. *Ann. Bot.* 102, 195–205. doi: 10.1093/aob/mcn074

Jaramillo-Correa, J. P., Beaulieu, J., Khasa, D. P., and Bousquet, J. (2009). Inferring the past from the present phylogeographic structure of north American forest trees: seeing the forest for the genes. *Can. J. For. Res.* 39, 286–307. doi: 10.1139/X08-181

Jiang, K., Tong, X., Ding, Y. Q., Wang, Z. W., Miao, L. Y., Xiao, Y. E., et al. (2021). Shifting roles of the East China Sea in the phylogeography of red nanmu in East Asia. *J. Biogeogr.* 48, 2486–2501. doi: 10.1111/jbi.14215

Jiang, F., and Wu, X. (1993). Fundamental characteristics of the stepped landform in China continent. *Mar. Geol. Quat. Geol.* 13, 15–24.

Jiménez-Valverde, A., and Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecol.* 31, 361–369. doi: 10.1016/j.actao.2007.02.001

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kim, K. H., Moon, E., Choi, S. U., Pang, C., Kim, S. Y., and Lee, K. R. (2015). Identification of cytotoxic and anti-inflammatory constituents from the bark of *Toxicodendron vernicifluum* (stokes) F. A. Barkley. *J. Ethnopharmacol.* 162, 231–237. doi: 10.1016/j.jep.2014.12.071

Kou, Y., Cheng, S., Tian, S., Li, B., Fan, D., Chen, Y., et al. (2016). The antiquity of *Cyclocarya paliurus* (Juglandaceae) provides new insights into the evolution of relict plants in subtropical China since the late early Miocene. *J. Biogeogr.* 43, 351–360. doi: 10.1111/jbi.12635

Leigh, J. W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Li, Y., Tang, Y., and Wu, T. (2020). The complete chloroplast genome of *Toxicodendron griffithii*. *Mitochondrial DNA B Resour.* 5, 2211–2212. doi: 10.1080/23802359.2020.1768931

Li, Y., Zhang, X., and Fang, Y. (2019). Landscape features and climatic forces shape the genetic structure and evolutionary history of an oak species (*Quercus chenii*) in East China. *Front. Plant Sci.* 10:1060. doi: 10.3389/fpls.2019.01060

Li, M. C., Zhang, Y. Q., Meng, C. W., Gao, J. G., Xie, C. J., Liu, J. Y., et al. (2021). Traditional uses, phytochemistry, and pharmacology of *Toxicodendron vernicifluum* (stokes) F. A. Barkley-a review. *J. Ethnopharmacol.* 267:113476. doi: 10.1016/j.jep.2020.113476

Li, Y., Zhang, X., Wang, L., Sork, V. L., Mao, L., and Fang, Y. (2022). Influence of Pliocene and Pleistocene climates on hybridization patterns between two closely related oak species in China. *Ann. Bot.* 129, 231–245. doi: 10.1093/aob/mcab140

Li, J., Zhao, M., Wei, S., Luo, Z., and Wu, H. (2015). Geologic events coupled with Pleistocene climatic oscillations drove genetic variation of Omei treefrog (*Rhacophorus omeimontis*) in southern China. *BMC Evol. Biol.* 15:289. doi: 10.1186/s12862-015-0572-1

Liao, Y. Y., Guo, Y. H., Chen, J. M., and Wang, Q. F. (2014). Phylogeography of the widespread plant *Ailanthus altissima* (Simaroubaceae) in China indicated by three chloroplast DNA regions. *J. Syst. Evol.* 52, 175–185. doi: 10.1111/jse.12065

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Liu, C., Tsuda, Y., Shen, H., Hu, L., Saito, Y., and Ide, Y. (2014). Genetic structure and hierarchical population divergence history of *Acer mono* var. *mono* in south and Northeast China. *PLoS One* 9:e87187. doi: 10.1007/s00468-011-0581-7

Luo, S., He, Y., Ning, G., Zhang, J., Ma, G., and Bao, M. (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* 25, 1063–1071. doi: 10.1007/s00468-011-0581-7

Luo, D., Xu, B., Li, Z. M., and Sun, H. (2021). Biogeographical divides delineated by the three-step landforms of China and the East China Sea: insights from the phylogeography of *Kerria japonica*. *J. Biogeogr.* 48, 372–385. doi: 10.1111/jbi.14002

Luo, D., Yue, J. P., Sun, W. G., Xu, B., Li, Z. M., Comes, H. P., et al. (2016). Evolutionary history of the subnival flora of the Himalaya-Hengduan Mountains: first insights from comparative phylogeography of four perennial herbs. *J. Biogeogr.* 43, 31–43. doi: 10.1111/jbi.12610

McCauley, D. E. (1994). Contrasting the distribution of chloroplast DNA and allozyme polymorphism among local populations of *Silene alba*: implications for studies of gene flow in plants. *Proc. Natl. Acad. Sci. U. S. A.* 91, 8127–8131. doi: 10.1073/pnas.91.17.8127

Noshiro, S., and Suzuki, M. (2004). *Rhus verniciflua* stokes grew in Japan since the early Jomon period. *Jpn. J. Hist. Bot.* 12, 3–11. doi: 10.34596/hisbot.12.1_3

Noshiro, S., Suzuki, M., and Sasaki, Y. (2007). Importance of *Rhus verniciflua* Stokes (lacquer tree) in prehistoric periods in Japan, deduced from identification of its fossil woods. *Veg. Hist. Archaeobotany* 16, 405–411. doi: 10.1007/s00334-006-0058-6

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28, 2537–2539. doi: 10.1111/j.1471-8286.2005.01155.x

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., Bittkau, C., Brewer, S., Cheddadi, R., et al. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300, 1563–1565. doi: 10.1126/science.1083264

Phillips, S. J., Dudík, M., and Schapire, R. E. (2018). Maxent software for modeling species niches and distributions. Available at: http://biodiversityinformatics.amnh.org/open_source/maxent/ (Accessed June 15, 2021).

Pons, O., and Petit, R. J. (1996). Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144, 1237–1245. doi: 10.1093/genetics/144.3.1237

Posada, D., and Crandall, K. A. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45. doi: 10.1016/S0169-5347(00)02026-7

Qi, X. S., Yuan, N., Comes, H. P., Sakaguchi, S., and Qiu, Y. X. (2014). A strong 'filter' effect of the East China Sea land bridge for East Asia's temperate plant species: inferences from molecular phylogeography and ecological niche modelling of *Platycrater arguta* (Hydrangeaceae). *BMC Evol. Biol.* 14:41. doi: 10.1186/1471-2148-14-41

Qian, H., and Ricklefs, R. E. (2000). Large-scale processes and the Asian bias in species diversity of temperate plants. *Nature* 407, 180–182. doi: 10.1038/35025052

Qiu, Y. X., Fu, C. X., and 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. doi: 10.1016/j.ympev.2011.01.012

Qiu, Y. X., Lu, Q. X., Zhang, Y. H., and Cao, Y. N. (2017). Phylogeography of East Asia's tertiary relict plants: current progress and future prospects. *Biodivers. Sci.* 25, 24–28. doi: 10.17520/biods.2016292

R Core Team (2018). R: A language and environment for statistical computing.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rendell, S., and Ennos, R. A. (2003). Chloroplast DNA diversity of the dioecious European tree *Ilex aquifolium* L. (English holly). *Mol. Ecol.* 12, 2681–2688. doi: 10.1046/j.1365-294X.2003.01934.x

Sakaguchi, S., Qiu, Y. X., Liu, Y. H., Qi, X. S., Kim, S. H., Han, J., et al. (2012). Climate oscillation during the quaternary associated with landscape heterogeneity promoted allopatric lineage divergence of a temperate tree *Kalopanax septemlobus* (Araliaceae) in East Asia. *Mol. Ecol.* 21, 3823–3838. doi: 10.1111/j.1365-294X.2012.05652.x

Schaal, B. A., Hayworth, D. A., Olsen, K. M., Rauscher, J. T., and Smith, W. A. (1998). Phylogeographic studies in plants: problems and prospects. *Mol. Ecol.* 7, 465–474. doi: 10.1046/j.1365-294x.1998.00318.x

Schoener, T. W. (1968). The Anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology* 49, 704–726. doi: 10.2307/1935534

Shafer, A. B., Cullingham, C. I., Cote, S. D., and Coltman, D. W. (2010). Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Mol. Ecol.* 19, 4589–4621. doi: 10.1111/j.1365-294X.2010.04828.x

Shi, M. M., Michalski, S. G., Welk, E., Chen, X. Y., and Durka, W. (2014). Phylogeography of a widespread Asian subtropical tree: genetic east–west differentiation and climate envelope modelling suggest multiple glacial refugia. *J. Biogeogr.* 41, 1710–1720. doi: 10.1111/jbi.12322

Soltis, D. E., Gitzendanner, M. A., Strenge, D. D., and Soltis, P. S. (1997). Chloroplast DNA intraspecific phylogeography of plants from the Pacific northwest of North America. *Plant Syst. Evol.* 206, 353–373. doi: 10.1007/BF00987957

Soltis, D. E., Morris, A. B., McLachlan, J. S., Manos, P. S., and Soltis, P. S. (2006). Comparative phylogeography of unglaciated eastern North America. *Mol. Ecol.* 15, 4261–4293. doi: 10.1111/j.1365-294X.2006.03061.x

Sun, Y. B. (2017). FasParser: a package for manipulating sequence data. *Zool. Res.* 38, 110–112. doi: 10.24272/j.issn.2095-8137.2017.017

Sun, Y., Hu, H., Huang, H., and Vargas-Mendoza, C. F. (2014). Chloroplast diversity and population differentiation of *Castanopsis fargesii* (Fagaceae): a

dominant tree species in evergreen broad-leaved forest of subtropical China. *Tree Genet. Genomes* 10, 1531–1539. doi: 10.1007/s11295-014-0776-3

Suzuki, M., Noshiro, S., Tanaka, T., Kobayashi, K., Wang, Y., Liu, J. Q., et al. (2014). Origin of Urushi (*Toxicodendron vernicifluum*) in the Neolithic Jomon period of Japan. *Bull. Natl. Mus. Jpn. Hist.* 187, 49–70. doi: 10.34596/hisbot.15.1_58

Suzuki, M., Yonekura, K., and Noshiro, S. (2007). Distribution and habitat of *Toxicodendron vernicifluum* (stokes) F. A. Barkl. (Anacardiaceae) in China. *Jpn. J. Hist. Bot.* 15, 58–62. doi: 10.34596/hisbot.15.1_58

Taberlet, P., Fumagalli, L., Wust-Saucy, A. G., and Cosson, J. F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.* 7, 453–464. doi: 10.1046/j.1365-294x.1998.00289.x

Tian, S., Lei, S. Q., Hu, W., Deng, L. L., Li, B. O., Meng, Q. L., et al. (2015). Repeated range expansions and inter-/postglacial recolonization routes of Sargentodoxa cuneata (Oliv.) Rehd. et Wils.(Lardizabalaceae) in subtropical China revealed by chloroplast phylogeography. *Mol. Phylogenet. Evol.* 85, 238–246. doi: 10.1016/j.ympev.2015.02.016

Vu, D. D., Bui, T. T., and Nguyen, T. H. (2018). Isolation and characterization of polymorphic microsatellite markers in *Toxicodendron vernicifluum*. *Czech J. Genet. Plant Breed.* 54, 17–25. doi: 10.17221/183/2016-CJGPB

Walker, S., Williams, J., Lear, J., and Beck, M. (2008). FS11.5 *Toxicodendron* dermatitis in the United Kingdom. *Contact Dermatitis* 50:163. doi: 10.1111/j.0105-1873.2004.0309cy.x

Wan, T. F. (2012). *The Tectonics of China: Data, Maps and Evolution*. Berlin: Springer Science & Business Media.

Wang, J., Gao, P., Kang, M., Lowe, A. J., and Huang, H. (2009). Refugia within refugia: the case study of a canopy tree (*Eurycorymbus cavaleriei*) in subtropical China. *J. Biogeogr.* 36, 2156–2164. doi: 10.1111/j.1365-2699.2009.02165.x

Wang, J. F., Gong, X., Chiang, Y. C., and Kuroda, C. (2013). Phylogenetic patterns and disjunct distribution in *Ligularia hodgsonii* hook. (Asteraceae). *J. Biogeogr.* 40, 1741–1754. doi: 10.1111/jbi.12114

Wang, L., He, N., Li, Y., Fang, Y. M., and Zhang, F. L. (2020a). Complete chloroplast genome sequence of Chinese lacquer tree (*Toxicodendron vernicifluum*, Anacardiaceae) and its phylogenetic significance. *Biomed. Res. Int.* 2020:9014873. doi: 10.1155/2020/9014873

Wang, L., He, N., Li, Y., Fang, Y. M., and Zhang, F. L. (2020b). The complete chloroplast genome sequence of *Toxicodendron succedaneum* (Anacardiaceae). *Mitochondrial DNA B Resour.* 5, 1956–1957. doi: 10.1080/23802359.2020.1756956

Wang, Y. H., Jiang, W. M., Comes, H. P., Hu, F. S., Qiu, Y. X., and Fu, C. X. (2015a). Molecular phylogeography and ecological niche modelling of a widespread herbaceous climber, *Tetrastigma hemsleyanum* (Vitaceae): insights into Plio–Pleistocene range dynamics of evergreen forest in subtropical China. *New Phytol.* 206, 852–867. doi: 10.1111/nph.13261

Wang, Y., Liu, K., Bi, D., Zhou, S., and Shao, J. (2018). Molecular phylogeography of east Asian *Boea clarkeana* (Gesneriaceae) in relation to habitat restriction. *PLoS One* 13:e0199780. doi: 10.1371/journal.pone.0199780

Wang, W., Tian, C. Y., Li, Y. H., and Li, Y. (2015b). Molecular data and ecological niche modelling reveal the phylogeographic pattern of *Cotinus coggygria* (Anacardiaceae) in china's warm-temperate zone. *Plant Biol.* 16, 1114–1120. doi: 10.1111/plb.12157

Warren, D. L., Glor, R. E., and Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62, 2868–2883. doi: 10.1111/j.1558-5646.2008.00482.x

Warren, D. L., Glor, R. E., and Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* 33, 607–611. doi: 10.1111/j.1600-0587.2009.06142.x

Watanabe, A., Tamura, M., Izumi, Y., Yamaguchi, R., Iki, T., and Tabata, M. (2019). Evaluation of genetic diversity of *Toxicodendron vernicifluum* planted in Japan using EST-SSR and genetic SSR markers. *J. Jpn For. Soc.* 101, 298–304. doi: 10.4005/jjfs.101.298

Wei, S., Zhao, X., Tian, M., Li, L., and Hu, Z. (2010). Application of plant morphology and AFLP molecular markers to identify *Toxicodendron vernicifluum* varieties of Shaanxi. *Xi Bei Zhi Wu Xue Bao* 30, 665–671.

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Wu, M., Zhang, B., Jiang, L., Wu, J., and Sun, G. (2018). Natural lacquer was used as a coating and an adhesive 8000 years ago, by early humans at Kuahuqiao, determined by ELISA. *J. Archaeol. Sci.* 100, 80–87. doi: 10.1016/j.jas.2018.10.004

Yang, A., Zhong, Y., Liu, S., Liu, L., Liu, T., Li, Y., et al. (2019). New insight into the phylogeographic pattern of *Liriodendron chinense* (Magnoliaceae) revealed by chloroplast DNA: east-west lineage split and genetic mixture within western subtropical China. *PeerJ* 7:e6355. doi: 10.7717/peerj.6355

Ye, J. W., Bai, W. N., Bao, L., Wang, T. M., Wang, H. F., and Ge, J. P. (2017b). Sharp genetic discontinuity in the aridity-sensitive *Lindera obtusiloba* (Lauraceae): solid evidence supporting the tertiary floral subdivision in East Asia. *J. Biogeogr.* 44, 2082–2095. doi: 10.1111/jbi.13020

Ye, J. W., and Li, D. Z. (2021). Distinct late Pleistocene subtropical-tropical divergence revealed by fifteen low-copy nuclear genes in a dominant species in south-East China. *Sci. Rep.* 11:4147. doi: 10.1038/s41598-021-83473-w

Ye, J. W., Zhang, Y., and Wang, X. J. (2017a). Phylogeographic breaks and the mechanisms of their formation in the Sino-Japanese floristic region. *Chin. J. Plant Ecol.* 41, 1003–1019. doi: 10.17521/cjpe.2016.0388

Yin, X., Qian, H., Sui, X., Zhang, M., Mao, L., Svenning, J. C., et al. (2021). Effects of climate and topography on the diversity anomaly of plants disjunctly distributed in eastern Asia and eastern North America. *Glob. Ecol. Biogeogr.* 30, 2029–2042. doi: 10.1111/geb.13366

Yu, H. B., Zhang, Y. L., Liu, L. S., Qi, W., Li, S. C., and Hu, Z. J. (2015). Combining the least cost path method with population genetic data and species distribution models to identify landscape connectivity during the late quaternary in Himalayan hemlock. *Ecol. Evol.* 5, 5781–5791. doi: 10.1002/ece3.1840

Zeng, Y. F., Wang, W. T., Liao, W. J., Wang, H. F., and Zhang, D. Y. (2015). Multiple glacial refugia for cool-temperate deciduous trees in northern East Asia: The *Mongolian* oak as a case study. *Mol. Ecol.* 24, 5676–5691. doi: 10.1111/mec.13408

Zhang, X. W., Li, Y., Liu, C., Xia, T., Zhang, Q., and Fang, Y. M. (2015). Phylogeography of the temperate tree species *Quercus acutissima* in China: inferences from chloroplast DNA variations. *Biochem. Syst. Ecol.* 63, 190–197. doi: 10.1016/j.bse.2015.10.010

Zhang, X. W., Li, Y., Zhang, Q., and Fang, Y. M. (2018). Ancient east-west divergence, recent admixture, and multiple marginal refugia shape genetic structure of a widespread oak species (*Quercus acutissima*) in China. *Tree Genet. Genomes* 14:88. doi: 10.1007/s11295-018-1302-9

Zhang, F. L., Zhang, W. Q., and Wei, S. N. (2007). Research and fine application of lacquer tree resources in China. *J. Chin. Lacquer* 2, 36–50. doi: 10.19334/j.cnki.issn.1000-7067.2007.02.005

Zhao, M., Liu, C., and Zheng, G. (2013). Comparative studies of bark structure, lacquer yield and urushiol content of cultivated *Toxicodendron vernicifluum* varieties. *N. Z. J. Bot.* 51, 13–21. doi: 10.1080/0028825X.2012.731005

Zhong, Y., Yang, A., Liu, S., Liu, L., Li, Y., Wu, Z., et al. (2019). RAD-Seq data point to a distinct split in *Liriodendron* (Magnoliaceae) and obvious east–west genetic divergence in *L. chinense*. *Forests* 10:13. doi: 10.3390/f10010013

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

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

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fpls.2022.829888

OPEN ACCESS

EDITED BY
Rong Wang,
East China Normal University, China

REVIEWED BY
Alexander Andrew Myburg,
University of Pretoria,
South Africa
Wen-Bin Yu,
Xishuangbanna Tropical Botanical Garden (CAS), China

*CORRESPONDENCE
Deqiang Zhang
deqiangzhang@bjfu.edu.cn

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 06 December 2021
ACCEPTED 13 July 2022
PUBLISHED 29 July 2022

CITATION
Song F, Zhou J, Quan M, Xiao L, Lu W, Qin S, Fang Y, Wang D, Li P, Du Q, El-Kassaby YA and Zhang D (2022) Transcriptome and association mapping revealed functional genes respond to drought stress in *Populus*. *Front. Plant Sci.* 13:829888. doi: 10.3389/fpls.2022.829888

Transcriptome and association mapping revealed functional genes respond to drought stress in *Populus*

Fangyuan Song[1,2], Jiaxuan Zhou[1,2], Mingyang Quan[1,2], Liang Xiao[1,2], Wenjie Lu[1,2], Shitong Qin[1,2], Yuanyuan Fang[1,2], Dan Wang[1,2], Peng Li[1,2], Qingzhang Du[1,2], Yousry A. El-Kassaby[3] and Deqiang Zhang[1,2]*

[1]National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [2]Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [3]Department of Forest and Conservation Sciences, Faculty of Forestry, Forest Sciences Centre, University of British Columbia, Vancouver, BC, Canada

Drought frequency and severity are exacerbated by global climate change, which could compromise forest ecosystems. However, there have been minimal efforts to systematically investigate the genetic basis of the response to drought stress in perennial trees. Here, we implemented a systems genetics approach that combines co-expression analysis, association genetics, and expression quantitative trait nucleotide (eQTN) mapping to construct an allelic genetic regulatory network comprising four key regulators (*PtoeIF-2B*, *PtoABF3*, *PtoPSB33*, and *PtoLHCA4*) under drought stress conditions. Furthermore, Hap_01PtoeIF-2B, a superior haplotype associated with the net photosynthesis, was revealed through allelic frequency and haplotype analysis. In total, 75 candidate genes related to drought stress were identified through transcriptome analyses of five *Populus* cultivars (*P. tremula* × *P. alba*, *P. nigra*, *P. simonii*, *P. trichocarpa*, and *P. tomentosa*). Through association mapping, we detected 92 unique SNPs from 38 genes and 104 epistatic gene pairs that were associated with six drought-related traits by association mapping. eQTN mapping unravels drought stress-related gene loci that were significantly associated with the expression levels of candidate genes for drought stress. In summary, we have developed an integrated strategy for dissecting a complex genetic network, which facilitates an integrated population genomics approach that can assess the effects of environmental threats.

KEYWORDS

association genetics, co-expression, eQTN, epistasis, drought tolerance, *Populus*

Introduction

Drought is an inevitable and recurring feature of global climate change, it is increasing in frequency and intensity. Forest trees constitute ~45% of global terrestrial carbon stocks and have key roles in ecosystem stability (Rogers et al., 2018). Extreme drought is greatly harmful to forest trees, it causes substantial productivity losses, affects ecosystem security, and threatens human survival (Vitasse et al., 2019). Thus, there is a need to explore the genetic architecture and regulatory mechanisms of drought stress in forest tree populations. Drought tolerance is a complex trait that involves several mechanisms, including escape, avoidance, and tolerance (Gupta et al., 2020). Plants drought stress mechanisms are related to hydraulic signals, reactive antioxidants, osmotic regulation, and phytohormone movements (Attipalli et al., 2004; Ahmad et al., 2010). Under drought stress, rapid expression of the *P5CS* gene in barley led to proline accumulation (Frimpong et al., 2021). Additionally, the over-expression of *PeCHYR1* (CHY-type/CTCHY-type/RING-type zinc finger protein) significantly improved drought tolerance in poplar trees by enhancing hyperoxide production and reducing the stomatal aperture (He et al., 2018). However, naturally occurring drought stress variation and the effects of drought stress adaptation at the species and population levels have not been systematically investigated.

Association genetics using molecular marker-based technologies, enables decryption of the genetic basis of phenotypic variation in forest trees. Specifically, population genomics promotes genetic improvement of drought tolerance and the development of diagnostic tools for the conservation and management of forest tree natural populations (Neale and Kremer, 2011). Additionally, association mapping is a widely used approach to investigate the allelic variants that underpin complex traits, it is particularly powerful in forest trees because of the high levels of nucleotide diversity and low linkage disequilibrium in perennial woody plants (Wegrzyn et al., 2010; Beaulieu et al., 2011; Guerra et al., 2013). In particular, association studies concerning additive, dominant and epistatic gene effects have provided insights into the genetic architecture that underlies plant phenotypic variation (Du et al., 2015; Deng et al., 2017). This strategy enables dissection of the genetic effects of multi-gene networks in *Populus*, allowing clarification of the genetic regulation of complex traits in trees (Quan et al., 2019). Expression quantitative trait nucleotide (eQTN) mapping, defined as associations between SNPs and expression level of candidate genes, is used to decipher the allelic variations that contribute to phenotypes at the transcriptional level, thus facilitating investigation of the effects of candidate genes (Lu et al., 2021). Co-expression network analysis allows the integration of transcriptome data types and clustering of genes with correlated expression patterns into co-expression modules, these capabilities permit exploration of the functional connections between candidate genes involved in the same or shared biological pathways (Serin et al., 2016). Thus, the combination of co-expression network analysis, association genetics, and eQTN mapping will provide insights into the genetic architecture that underlies the response of trees to drought stress.

Populus is a major fast-growing plantation tree genus used for bioenergy, timber, and pulp manufacturing; it also is an excellent model system of long-lived forest trees for biological studies related to environmental changes (Jansson and Douglas, 2007; Lu et al., 2021; Zhao et al., 2021). *Populus* comprises >30 species and is geographically distributed throughout the northern hemisphere (Taylor, 2002). However, most fast-growing poplar varieties have poor drought stress tolerance (Tuskan et al., 2006; Lüttschwager et al., 2015). The construction of a systematic network and identification of candidate genes related to the drought stress response would improve our understanding of drought stress in *Populus*. Here, we firstly used transcriptome data of five representative poplar species (*P. tremula* × *P. alba*, *P. nigra*, *P. simonii*, *Populus trichocarpa*, and *P. tomentosa*) to identify differentially expressed genes (DEGs) under drought conditions. Through weighted gene co-expression network analysis (WGCNA), and based on DEGs, we detected three important modules and 75 candidate genes related to drought stress. Next, we performed association mapping to identify the significant associated loci and genes for six drought-responsive traits in an association population of 300 *P. tomentosa* accessions under drought stress. Based on these findings, we proposed the genetic networks in the drought stress response pathway, which will be useful for molecular marker-assisted breeding of drought tolerant individuals in poplar. Expression quantitative trait nucleotide (eQTN) mapping combined with the analysis of six drought-responsive traits aided our interpretation of candidate genes related to drought stress. Our method will enable the exploration of the candidate genes related to drought tolerance for molecular marker-assisted selection (MAS) of drought-tolerant varieties of poplar.

Materials and methods

Plant materials and drought stress treatment

The association population consisted of 300, one-year-old *P. tomentosa* accessions with three ramets of each genotype, which were asexually propagated *via* root segments in 2018 in Guan Xian County, Shandong Province, China (36°23′N, 115°47′E); this area represents most of the species' natural distribution range. The distribution of these individuals was divided into southern (S, $n = 94$), northwestern (NW, $n = 108$), and northeastern (NE, $n = 108$) geographical regions (Huang, 1992). All individuals were well-watered by an automatic irrigation system three times per week and subjected to a well-watered (WW) period for 20 days to ensure their root development. Water deficit (WD) treatment began when leaf 6 (L6) was initiated on the apex, according to visual inspection (Boyes, 2001). The drought stress treatment was

as follows: (1) 20 days well-watered (WW); (2) followed by 30 days water deficit (WD) period until 70% of the leaves became wilted and yellow; and (3) then, a re-watering (RW) period for 20 days three times per week (Supplementary Figure S1). The volumetric soil water content (SWC) was measured using a model 4,300 neutron attenuation soil moisture meter and used to evaluate the degree of drought (the soil water contents were ~10% for WD and 40% for both WW and RW) (Grote et al., 2010). The daily mean minimum and maximum temperatures from WW to WD were 32.3°C and 37.6°C, respectively; these minimum and maximum temperatures were 34.2°C and 39.0°C, respectively, from WD to RW. Plants were exposed to the ambient mean relative air humidity (67.5%), with minimal precipitation. Functional leaves (i.e., the fourth to sixth leaves from the top of the stem) were collected from three biological replications separately from the 300 individuals, then three technical replications of each sample were conducted separately. The same experiment was conducted at well-watered (WW) condition, water deficit (WD) condition and re-watering (RW) condition, separately. The samples were immediately immersed in liquid nitrogen and stored at −80°C prior to vacuum freeze-drying. The leaf materials were used for subsequent drought stress index measurement.

Phenotype analysis

Photosynthesis, proline content (PRO), and catalase activity (CAT) are highly sensitive to changes in environmental factors, including drought stress. We measured six drought stress-related traits under water deficit (WD) and well-watered (WW) conditions. The photosynthetic traits were net photosynthesis (Pn), stomatal conductance (Cond), transpiration rate (Trmmol), and relative chlorophyll content (Chl). The proline content (PRO) and catalase activity (CAT) were also measured. The phenotypic variation of the traits is provided in Supplementary Table S1.

Photosynthetic traits were measured from fully expanded leaves (three functional leaves, the top fourth to sixth leaves) using a portable photosynthesis system (LI-6400, LI-COR, Lincoln, NE, Unites States) in accordance with the manufacturer's instructions. Each genotype was measured on sunny days between 9:00 and 11:30 a.m. under a fixed light intensity of 1,200 $\mu mol\, m^{-2}\ s^{-1}$ during the drought treatment. All measurements were performed using three replicates per individual genotype. Next, we used a portable chlorophyll meter (SPAD-502, Konica-Minolta, Japan) to measure the leaf chlorophyll concentration, which is presented as the SPAD value. For each leaf, the chlorophyll content was estimated as the mean of 10 SPAD values at different positions of the leaf middle section excluding the leaf midrib.

After the photosynthetic characteristics had been measured, the same functional leaves were immediately collected from the 300 accessions (separately from three ramets of each accession) and frozen in liquid nitrogen for PRO and CAT measurement. Proline content was extracted from 1.0 g of fresh leaves using 10 ml of 3% sulfosalicylic acid at 100°C for 10 min. A 4-ml aliquot of the extract was then mixed with 4 ml of ninhydrin reagent containing glacial acetic acid, then incubated at 95°C for 30 min. The reaction mixture was quickly cooled with running tap water. The colored reaction product was extracted with 8 ml of toluene, and the absorbance of the toluene phase was measured at 520 nm using a spectrophotometer (Shimadzu, Model UV 1800, Kyoto, Japan). To determine CAT activity in leaf extracts, 30 µl of extract were added to 50 mM K-phosphate buffer (pH 7.0) and 2% H_2O_2 for a total volume of 3 ml. Enzyme activity was calculated based on the absorbance at 240 nm recorded for 2 min using a spectrophotometer (see above).

Coefficient of variation (CV) values defined as the ratio of the standard deviation (SD) to the mean of each drought stress-related trait in the population, were independently calculated using the mean of the biological replicates of the untransformed drought stress-related traits data. The Pearson correlation coefficient (r) for each drought stress-related trait pair was calculated using the R package psych (Revelle, 2017).

Transcriptome data processing

Fully expanded leaves of 10 *P. tomentosa* genotypes, which covered three different regions were collected at WW (well-watered) and WD (water deficit) time points, respectively. The collected leaves were immediately immersed in liquid nitrogen, and stored at −80°C. Total RNA was extracted using a Qiagen RNeasy Kit (Qiagen China, Shanghai, China), in accordance with the manufacturer's instructions. In addition, DNase digestion was performed using an RNase Free DNase Kit (Qiagen). Detailed descriptions of the methods used for processing transcriptome data are provided in the supplemental materials (Supplementary Methods S1; Xiao et al., 2019).

Populus tremula × *P. alba*, *P. nigra*, *P. simonii*, and *P. trichocarpa* transcriptome data under different drought conditions were obtained from the NCBI SRA database (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/sra), and saved in FASTQ format using the SRA Toolkit. The quality control method is described in Method S1. In total, we obtained an expression data set composed of 58 RNA-seq samples and 27,644 genes (Supplementary Tables S3, S4). Transcript expression level was normalized by calculating the Z-score based on fragments per kilobase of transcript per million fragments (FPKM) method (Supplementary Table S4).

The linear model LIMMA package in Bioconductor[1] was used to perform differential gene expression analysis for the five species (Ritchie et al., 2015; genes with $|\mathrm{Log_2}\ (\text{fold-change})| > 1$ ($p < 0.05$) for DEGs). Gene ontology (GO) analysis was performed *via*

1 https://www.bioconductor.org/

AgriGO,[2] based on *P. trichocarpa* v3.0 annotation. Pathway enrichment analysis was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and a hypergeometric statistical test.[3]

Weighted gene co-expression network analysis analyses of DEGs

Weighted gene co-expression network analysis to identify key modules and hub genes is increasingly used in bioinformatics analyses in various biological contexts (Smita et al., 2013). An expression matrix based on 1,236 genes differentially expressed in more than three species in response to drought stress were used to construct a weighted gene co-expression network using the WGCNA package (Langfelder and Horvath, 2008). Weighted gene co-expression network analysis network construction and module detection were conducted using an unsigned type of topological overlap matrix (TOM), a power β of 8, a minimal module size of 30, and a branch merge cut height of 0.25. The adjacency matrix dissimilarity was 0.2. We then obtained several key network properties such as the edge weight and node connectivity. To identify the hub genes of a module, genes with edge weight ≥ 0.5 and the node connectivity ≥10 in the network were considered to be hub genes. Then Cytoscape (v.2.8.3) was used to visualize the correlation relationships between specified genes (Cline et al., 2007).

Reverse transcription-quantitative polymerase chain reaction

Five genes were selected for validation of their expression profiles in 10 individuals with different genotypes using RT-qPCR. Three leaves (the fourth to sixth from the top of the stem) were collected from 10 one-year-old *P. tomentosa* seedlings and immediately immersed in liquid nitrogen. Total RNA was extracted from each leaf and reverse transcribed into cDNA using the Reverse Transcription System (Promega Corporation, Madison, WI, United States). Reverse transcription-quantitative PCR (RT-qPCR) was performed on the 7,500 Fast Real-Time PCR System using SYBR Premix Ex Taq (TaKaRa, Dalian, China), in accordance with the manufacturer's protocol. Specific primer pairs for each gene were designed using Primer-BLAST software (Ye et al., 2012; Supplementary Table S8). All reactions were performed with triplicate technical and triplicate biological repetitions, with actin (EF145577) as the internal control, in accordance with the PCR program described by Xiao et al. (2019).

2 http://systemsbiology.cau.edu.cn/agriGOv2/index.php

3 https://www.kegg.jp/

Genome re-sequencing and SNP/InDel calling of *Populus tomentosa* association population

We used a Plant DNeasy Mini kit (Qiagen, Shanghai, China) to isolate the total genomic DNA of the 300 *P. tomentosa* unrelated individuals, in accordance with the manufacturer's instructions. Total genomic DNA was re-sequenced at a depth > 15 × (raw data) using the Illumina GA2 sequencing platform. The clean reads were mapped to the *P. trichocarpa* reference genome v.3.0; they were used to perform SNP calling. SNP calling as described by Xiao et al. (2019). VCFtools was used to extract the gene-derived biallelic SNPs/InDels within the genes, including their 1,000 bp upstream and 1,000 bp downstream sequences. Finally, 5,553 SNPs of 75 candidate genes from the 300 accessions were used for association analysis (Supplementary Table S12).

Association analysis

Single SNP-based association

The mixed linear model (MLM) in Tassel 5.0 was used to test the statistical associations between SNPs and the drought stress-related traits which were normalized based on the Z-score (Bradbury et al., 2007), after accounting for the population structure (Q) and pairwise kinship coefficients (K). The K matrix was derived by SPAGeDiv1.3 (Hardy and Vekemans, 2002) and the Q matrix was determined *via* STRUCTURE v.2.3.4 based on significant sub-populations ($k = 3$) (Evanno et al., 2005). The QVALUE package in R was used to correct for multiple testing with the positive false discovery rate (FDR) method (Storey, 2003). SNPs were considered significantly associated at $p < 0.001$ and $q < 0.05$ were identified. Manhattan plots and Q-Q plots were created using the qqman package in R v.3.0.2 (Mukrimin et al., 2018). Haplotype analysis was performed through Haploview v.4.2 software with default parameters (Barrett et al., 2005). Superior haplotypes were identified in accordance with the method established by Lv et al. (2021).

Multi-SNP epistasis association analysis

The EPISNP package in the epiSNP v.4.2 software suite was used to analyze epistatic effects (Ma et al., 2008). The SNP–SNP interaction effect with phenotypic traits was partitioned into four epistatic effects based on the extended Kempthorne model: additive × additive, additive × dominant, dominant × additive, and dominant × dominant epistatic effects. The significance level was defined as $p < 0.001$. Only the SNPs that demonstrated significance ($p < 0.01$) in SNP-based association mapping were used for epistasis analysis. A multifactor dimensionality reduction (MDR) algorithm was conducted to investigate the genotype combination effects in our studies (Hahn et al., 2003).

eQTN mapping

eQTN mapping was performed using Tassel v.5.0 software, with a method identical to the SNP-based association analysis. eQTNs were considered significantly associated at $p < 0.001$ and $q < 0.05$. RNA-seq was used to measure the transcript levels of genes from the functional leaves of the 300 *P. tomentosa* individuals. RNA library construction and sequencing were performed by Beijing Biomarker Technology Cooperation (Beijing, China). The FPKM values (i.e., gene expression levels) were calculated as described in Supplementary Methods S1. Gene expression traits with missing data >20% and expression levels <0.1 (FPKM < 0.1) in >95% of the 300 individuals were removed. The detected eQTNs located in the 10-kb window around the expressed gene were defined as cis-eQTNs, and the remaining eQTNs were regarded as trans-eQTNs.

Sequence analysis and phylogenetic tree construction

Amino acid sequences were obtained from the NCBI database,[4] which were aligned and used to infer their phylogenetic relationships. Multiple sequence alignment was performed with MEGA v.6.0. The phylogenetic tree was constructed using MEGA v.6.0 with the neighbor-joining (NJ) algorithm and Kimura two-parameter model. Genetic distance was calculated using sequence pairwise alignments. The reliability of nodes on the neighbor-joining tree was estimated using a bootstrap analysis with 1,000 replicates.

Results

Construction of co-expression networks in *Populus* under drought stress

Comparative analyses were conducted for the five poplar species to identify drought-responsive genes. 1,236 drought-responsive DEGs (i.e., genes differentially expressed in >3 species in response to drought stress) were identified among the five species; these comprised 261 (35.23%) up- and 975 (64.77%) down-regulated genes (Supplementary Table S4). GO analysis of the 1,236 DEGs revealed the enrichment of 78 significant terms ($p < 0.05$) related to biological processes such as photosynthesis, multiple hormone-mediated regulations, and energy metabolism (Supplementary Table S6). KEGG analysis indicated that the DEGs were enriched in pathways such as photosynthesis and oxidative phosphorylation (Supplementary Table S7). Transcriptome analysis suggested that photosynthesis was the process most susceptible process in response to drought stress. We validated the expression levels of five randomly selected DEGs by RT-qPCR (Supplementary Figure S2). The expression patterns of the five genes according to RT-qPCR were similar to the patterns identified by RNA-seq, thereby validating the RNA-seq results.

A weighted co-expression network was constructed using 1,236 DEGs in the five-poplar species. The blue, brown, and turquoise modules contained 376, 168, and 692 genes, respectively, these results implied highly similar expression patterns among candidate genes (see the dendrogram in Figures 1A–F; Supplementary Table S5). Of the three modules, 75 hub genes with edge weight ≥ 0.5 and node connectivity ≥10 were selected. There were 16, 17, and 42 hub genes included in the blue, brown, and turquoise modules, respectively (Figures 1D–F; Supplementary Table S9). Among the 75 hub genes, most were known to be involved in drought stress. For example, the defense response gene *PtoRGP* (reduction in growth and productivity) regulates cellular processes that are involved in growth and abiotic stress responses (Lee et al., 2014); *PtoVHA-c* (encodes hydrolysis of the V-ATPase c subunit) confers stress tolerance through enhancing superoxide dismutase and peroxidase activities under drought stress (Cheng et al., 2013), and *PtoHDA15* (histone deacetylase) inhibits abscisic acid (ABA) signaling genes (Liu et al., 2013). Additionally, several novel high-degree hub gene signatures were identified in our analysis, such as *PtoRPN5A* (26S proteasome regulatory protein) and *PtoNRP1* (nodulin-related protein 1), a DNA-binding protein. To assess the roles of these 75 hubs in the network, we used these hub genes from each module and conducted a GO analysis. The genes were significantly enriched in hormone-mediated biosynthesis or antioxidant process (hydrogen peroxide) and photosynthetic components (cellular component; Supplementary Figure S3; Supplementary Table S6), which is consistent with the notion that hub genes typically play roles in the integration of other genes within a module (Ravasz, 2002).

Allelic variation significantly associated with drought stress-related traits in *Populus tomentosa*

To further explore the genetic effects of candidate genes for drought stress-related traits in the co-expression network, six drought stress-related traits (Pn, Cond, Trmmol, Chl, PRO, and CAT) in functional leaves were measured in the 300 *P. tomentosa* individuals under WD conditions. Four photosynthetic traits were decreased under drought stress, while PRO, and CAT were increased (Supplementary Table S1). All six drought stress-related traits exhibited high genetic variation, with the coefficient of variation (CV) values ranging from 0.15 (Cond) to 7.59 (PRO; Supplementary Table S1). Estimates of heritability showed that four of six drought stress-related traits were with the broad-sense heritability (H^2) > 0.6 (Supplementary Table S1); thus, illustrated they were presumed to be controlled by genomic variants. Pearson

4 http://www.ncbi.nlm.nih.gov/

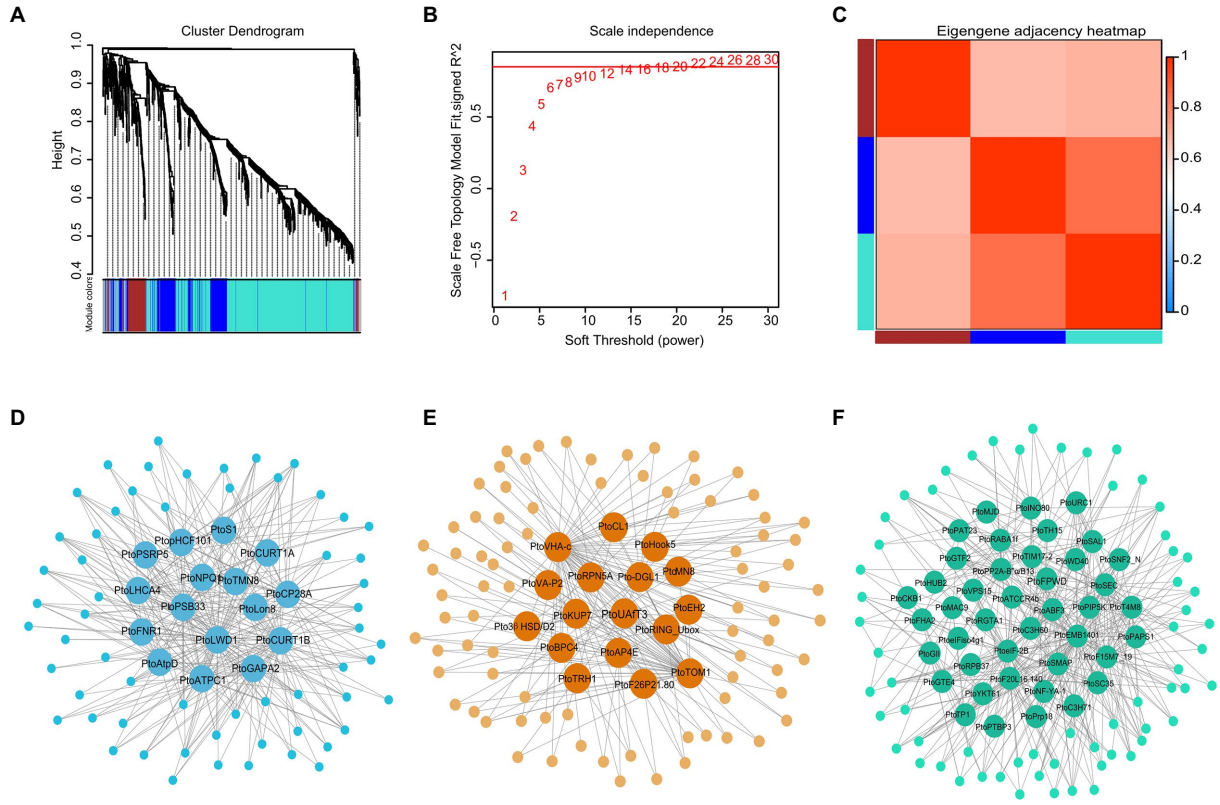


FIGURE 1
Co-expression network of differentially expressed genes (DEGs) in five poplar species. **(A–C)** Clustering dendrograms of gene expression patterns. Dissimilarity was based on topological overlap, along with assigned module colors. The three co-expression modules are shown in different colors. **(D–F)** Network relationship among the modules. In total, 16, 17, and 42 hub genes included in the blue, brown, and turquoise modules, respectively. Hub genes with edge weight ≥0.5 and node connectivity ≥10 were used to construct the network in Cytoscape.

correlation analysis showed that traits within the same category were often closely correlated (Supplementary Table S2). These results indicated that the association population possessed significant genetic variability and could be used for population genetics analysis of the response to drought stress.

We conducted an association analysis concerning the genetic associations of 5,553 SNPs in 75 candidate genes with six drought stress-related traits (Supplementary Figure S4; Supplementary Table S12). The model identified 92 unique SNPs from 38 genes that showed significant associations with the six drought-related traits ($p < 0.001$, $q < 0.05$); the mean explained phenotypic variation (R^2) was 9.40% (range: 0.13%–28.24%; Supplementary Table S13). Of these associations, 11 showed a combination of additive and dominant effects (Supplementary Table S13). Furthermore, three SNPs were simultaneously associated with two traits, indicating that they had pleiotropic effects on different drought related traits. For example, PtoLHCA4_SNP4 (T/A), located in the 3'UTR region of *PtoLHCA4* (encodes chlorophyll a-b binding protein), was simultaneously associated with Chl ($R^2 = 21.11\%$), and PRO ($R^2 = 9.68\%$). PtoABF3_SNP31 (A/G) was simultaneously associated with Cond ($R^2 = 23.27\%$), and Pn ($R^2 = 8.02\%$). Finally, PtoWD40_SNP25 (T/A), located in the intron region of *PtoWD40* (transducin family protein/WD-40 repeat family protein), was associated with Cond ($R^2 = 13.32\%$), and Pn ($R^2 = 7.49\%$; Supplementary Table S13).

We also detected multiple SNPs that were associated with the same trait (Supplementary Table S13). Notably, there were 44 SNPs associated with PRO in 23 annotated genes. Genes in the turquoise module were mainly enriched in photosynthetic components (Supplementary Figure S3; Supplementary Table S6). Three candidate genes (*PtoWD40*, *PtoRGP*, and *PtoeIF-2B*) in the turquoise module were selected (Supplementary Tables S9, S11). The SNPs (PtoWD40_SNP7, PtoRGP_SNP28, and PtoeIF-2B_SNP116) were significantly associated with PRO (Figure 2A; Supplementary Table S13). Distinct genotypes of the three SNPs contributed differently to PRO; there were three possible common genotypic combinations (frequency > 5%, $p < 0.01$) for PRO (Figures 2B,C). The genotypic combinations of three SNPs led to PRO phenotypic differences, in which TT-AG-CT and TT-AA-CC combinations represented the maximum (63.20 μg/g) and minimum (37.44 μg/g) phenotypic values (Figure 2C).

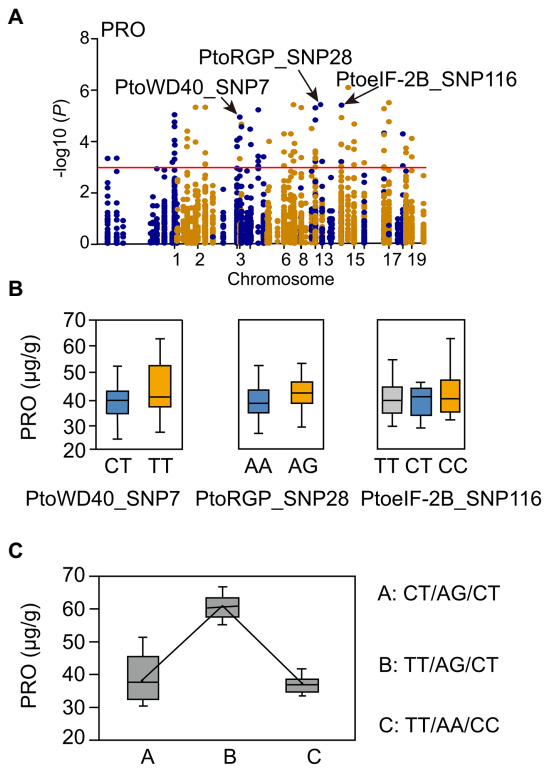


FIGURE 2
Proposed multi-SNP-based genotypic combinations for proline content (PRO). **(A)** Manhattan plot for PRO marked with candidate genes. **(B)** Genotypic effects of causal SNPs for PRO. **(C)** Three possible genotypic combinations with a frequency of ≥5% from the three allelic variations, and the genotypic combination effects for proline content in the *P. tomentosa* association population. SNPs in each genotypic combination were ordered according to **(B)**.

Pairwise epistasis of candidate loci revealed complex genetic networks under drought stress

Epistasis is a critical component of the genetic basis of quantitative traits because it defines the non-additive interactions between variants or genes (Mackay, 2013). To decipher the genetic networks in the response to drought stress, epiSNP was used to assess the epistatic effects of SNP–SNP pairs (Ma et al., 2008). In total, 104 significant pairwise associations ($p < 0.001$) were identified; these associations involved six drought stress-related traits with 95 unique SNPs from 21 genes (Supplementary Table S14). Kempthorne partitioned Fisher's epistasis effect into four components—additive × additive, additive × dominance, dominance × additive, and dominance × dominance—with the genetic interpretation of allele × allele, allele × genotype, genotype × allele, and genotype × genotype interactions, respectively (Mao et al., 2006). These interactions were partitioned into additive × additive (19 pairs), additive × dominant or dominant × additive (75 pairs), and dominant × dominant (10 pairs; Supplementary Table S14). Additionally, 11 significantly associated genes were repeatedly found to exhibit epistatic effects (104 pairwise), including 10 genetic variants that showed additive/dominant effects. For example, PtoSEC_SNP62 and PtoSEC_SNP75 showed combined additive and dominant effects for Chl. Moreover, they displayed epistatic interactions with PtoeIF-2B_SNP116 on CAT and Cond (Supplementary Table S14). In a total of 43 SNPs showed epistatic interactions with 2-20 SNPs and 11 SNP–SNP pairs were associated with more than one trait. For example, epistatic effects were detected for PtoPSB33_SNP20 (A/T) and PtoeIF-2B_SNP78 (C/T) on CAT and PtoABF3_SNP3 (A/G) with PtoeIF-2B_SNP78 (C/T) for Cond. Additionally, PtoeIF-2B_SNP78 (C/T) interacted with PtoPSB33_SNP35 (T/A) and PtoPSB33_SNP19 (T/G), both of which interacted with Cond (Figures 3A,B). The different allelic interactions showed distinct effects under drought stress. These results suggested that *PtoeIF-2B* (putative translation initiation factor eIF-2B epsilon subunit) had pleiotropic effects on several traits in response to drought stress.

Different genotypic combinations of SNP–SNP pairs had distinct epistatic effects. For instance, Chl varied across different genotypic interactions of PtoABF3_SNP19 (A/G) and PtoLHCA4_SNP3 (A/T); their mean differences in phenotypic values ranged from 18.83 mg/g (AG-TT) to 40.74 mg/g (GG-AA) (Figure 3D). Notably, PtoeIF-2B_SNP4 (C/T) and PtoLHCA4_SNP3 (A/T) also showed epistatic interaction on Chl, but the mean phenotypic values of each genotypic interaction were distinct, ranged from 19.06 mg/g (CC-AT) to 51.27 mg/g (CC-AA) (Figure 3E). The phenotypic values of various genotypic combinations differed from the values of single SNP effects (Figures 3C–E). These epistatic networks of significant drought stress-responsive genetic factors provide alternative effect models for photosynthetic and enzyme activity traits in *P. tomentosa*.

Genetic regulation of gene expression explains a substantial proportion of the phenotypic variations in response to drought stress of *Populus*

To explore the regulatory interactions between allelic variants and expression levels of candidate genes, eQTN mapping was conducted between 5,553 common SNPs (minor allele frequencies >0.05 and missing data <20%) and the expression levels of 75 candidate genes under drought stress. At the threshold of $p < 0.001$ and $q < 0.05$, 319 SNP-gene pairs were identified; thus, 194 unique SNPs in 35 candidate trans-acting factors were associated with the expression levels of 45 candidate genes (Supplementary Table S15). In a total of 52 SNPs were associated with the expression levels of 2–11 genes, suggesting that the expression levels of these candidate genes are under complex genetic regulation. For example, the trans-eQTN *PtoPSB33* (photosystem II protein 33, PtoPSB33_SNP4) was significantly associated with Pn and determined the expression levels of four genes: *PtoPAT23* (protein S-acyl transferases), *PtoPTBP3* (polypyrimidine tract-binding protein), *PtoVHA-c* (vacuolar adenosine triphosphate synthase family

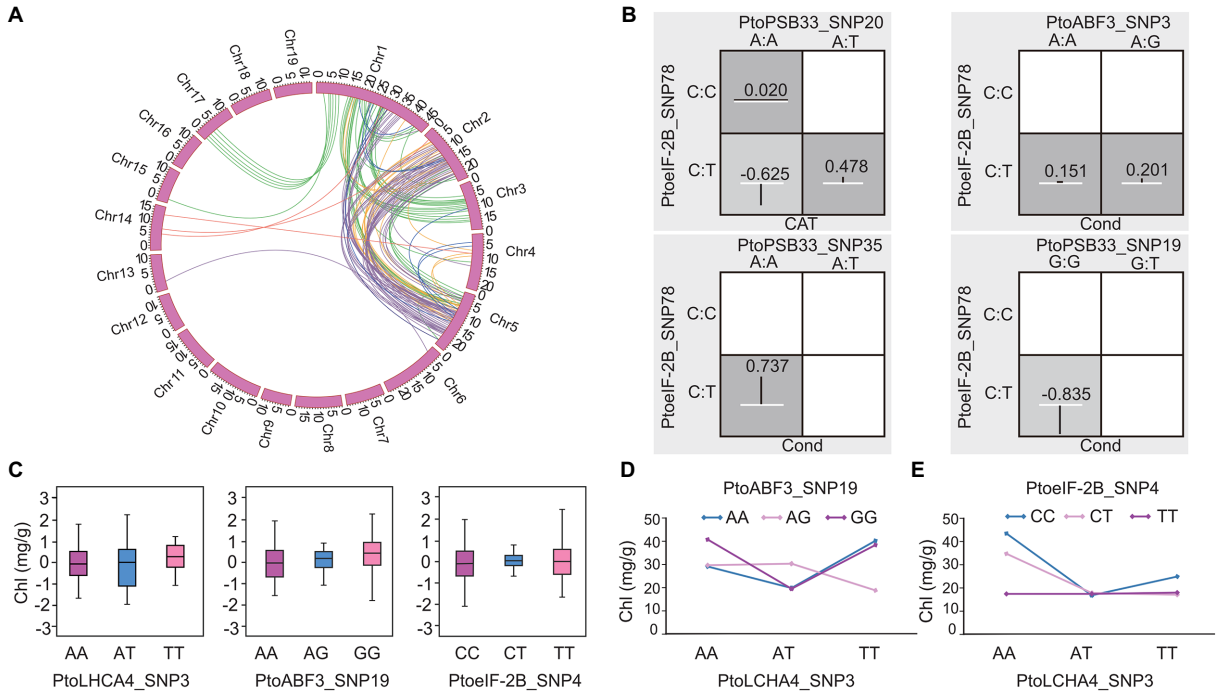


FIGURE 3
Allelic interactions between significant pairwise SNPs in candidate genes related to the co-expression network of drought stress traits. **(A)** Circos plot showing 104 pairwise interactions for drought stress-related traits ($p<0.001$). The 19 *P. tomentosa* chromosomes (Chr1-19) are shown in a circle. Interior lines represent the pairwise interactions that underlie six drought stress-related traits; colored lines represent different categories [green, purple, orange, red, dark blue, and light blue indicate relative chlorophyll content (Chl), stomatal conductance (Cond), net photosynthesis (Pn), proline content (PRO), transpiration rate (Trmmol), and catalase activity (CAT), respectively]. **(B)** Epistatic effects of different genotypic combinations for drought stress-related traits. Dark-shaded cells represent high-risk genotype combinations, while light-shaded cells represent low-risk genotype combinations. Values in boxes are individual information gains and positive values along the line indicate positive interactions. The white line in the middle of each box represents the mean phenotypic value of the population, while the vertical line represents the difference between the mean phenotypic value of each genotype combination and the overall mean. The width of the vertical line in the box indicates the number of individuals in this genotype combination. The negative values can be explained as negative interaction/redundancy (i.e., the amount of information shared by the attributes). **(C)** Genotypic effects of chlorophyll content (Chl) causal SNPs. **(D,E)** Epistatic effects for chlorophyll content between PtoLHCA4_SNP3 with PtoABF3_SNP19 and PtoeIF-2B_SNP4.

protein), and *PtoHDA15* (histone deacetylase). The *PtoHDA15* and *PtoVHA-c* expression levels were negatively correlated with Pn (Figure 4I, Supplementary Tables S12, S14), suggesting that *PtoPSB33* indirectly serves as a master regulator or mediates the leaf physiological response to drought stress.

In addition, a total of 17 SNPs detected by single SNP-based association studies (Supplementary Table S13), and 29 SNPs detected by the epistasis model overlapped with SNPs identified by eQTN analysis (Supplementary Table S14). For example, PtoPSB33_SNP1 and PtoeIF-2B_SNP2 formed an epistatic interaction with Pn and Cond (Figures 5A,B). Moreover, both were associated with the expression level of *PtoHDA15* (Figure 5C), which could contribute to the photosynthetic traits. Therefore, epistatic effects in the genetic architecture of complex quantitative traits have been overlooked. PtoLHCA4_SNP1 was associated with Chl and the expression level of *PtoVHA-c* (Figure 5D), suggesting that *PtoLHCA4* probably contributed to the phenotypic variation. Alternatively, these factors might also affect photosynthetic traits by regulating the expression levels of other genes. Therefore, the combination of association mapping and eQTN mapping enabled the evaluation of the genetic interactions and regulatory networks of the leaf physiological response to drought stress.

Determination of the four conserved genes that respond to drought stress in *Populus*

By combining the results of association mapping, eQTN, and co-expression analyses, we evaluated the putative functions of the genes involved in drought stress. We detected the candidate gene *PtoeIF-2B* in two major variants (PtoeIF-2B_SNP4 and PtoeIF-2B_InDel1). PtoeIF-2B_SNP4 (C/T) was a significant SNP for Chl ($p=7.20\times10^{-5}$, $q<0.05$, $R^2=21.94\%$; Figures 4A,B). The frequency of allele (TT) from PtoeIF-2B_SNP4 for higher Chl increased from the S (5.6%) geographical region to the NE (14.8%) and NW (20.6%) geographical regions, suggesting that *PtoeIF-2B* is subjected to adaptive selection in response to the local environment (Figure 4D; Supplementary Table S16). To assess the functional roles

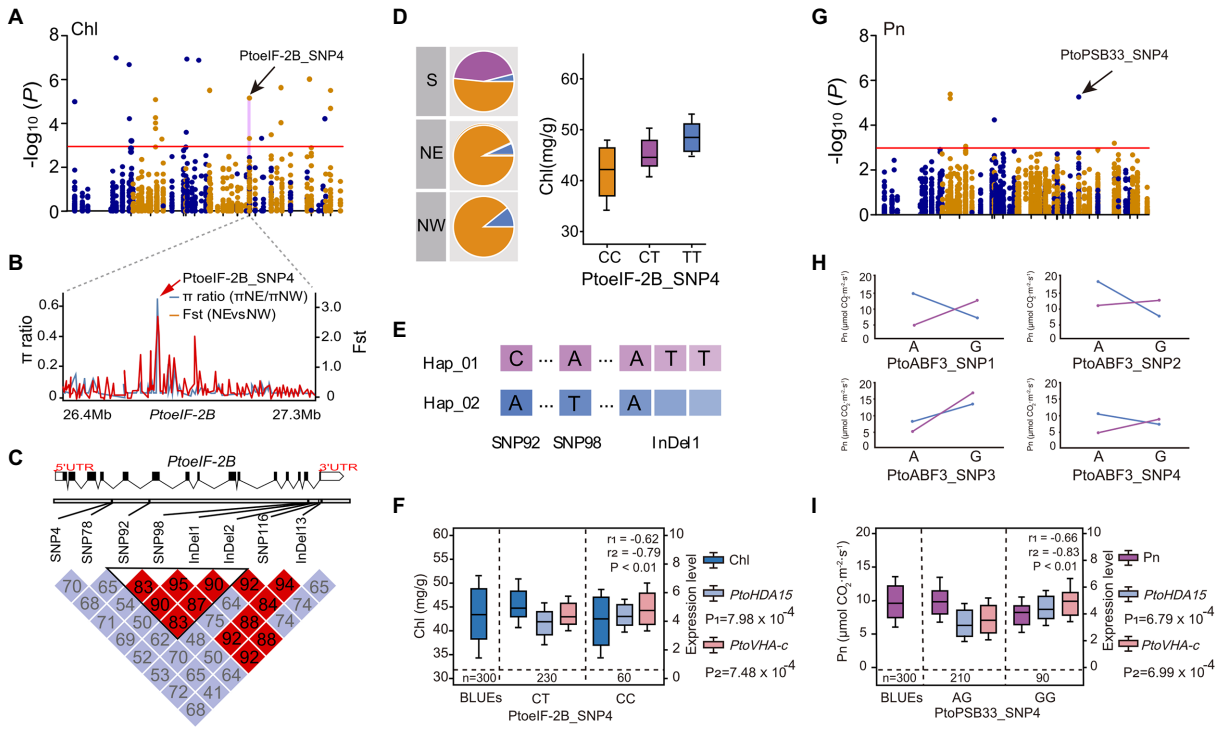


FIGURE 4
PtoPSB33 and *PtoeIF-2B* are implicated in the response to drought stress in *P. tomentosa*. **(A)** Manhattan plot for Chl. **(B)** Identification of the selection sweep signature of PtoeIF-2B_SNP4. **(C)** The genome structure and haplotype block of *PtoeIF-2B* gene. **(D)** Genotypic frequencies of significant SNP of *PtoeIF-2B* among the three regions. **(E)** Three significant loci detected by single-gene association analysis, which constituted a conserved haplotype. **(F)** Box plots for Chl (dark-blue), *PtoHDA15* expression (blue) and *PtoVHA-c* expression (pink) plotted as an effect of genotype at PtoeIF-2B_SNP4. **(G)** Manhattan plot for Pn marked with the focal SNP of PtoPSB33_SNP4. **(H)** The epistatic effects of PtoPSB33_SNP4 with four SNPs in *PtoABF3*. **(I)** Box plots for Pn (purple), *PtoHDA15* expression (blue) and *PtoVHA-c* expression (pink) plotted as an effect of genotype at PtoPSB33_SNP4.

of *PtoeIF-2B*, we conducted a haplotype analysis and the results revealed PtoeIF-2B_InDel1 and two other SNPs (PtoeIF-2B_SNP92 and PtoeIF-2B_SNP98) in *PtoeIF-2B* that constitute the haplotype were significantly associated with drought stress-related traits. Hap_01PtoeIF-2B (C-A-ATT) was identified as the superior haplotype associated with Pn; the mean value for Hap_01PtoeIF-2B that was 12.0% higher than the mean value for Hap_02PtoeIF-2B ($p < 0.05$; Figures 4C–E; Supplementary Table S17). Furthermore, PtoeIF-2B_SNP4 (C/T) was a trans-eQTN for the expression levels of *PtoHDA15* and *PtoVHA-c* which were negatively correlated with Chl ($r_1 = -0.62$ and $r_2 = -0.79$, respectively, $p < 0.01$; Figure 4F). Notably, we detected seven strong eQTN signals that were associated with the expression levels of *PtoHDA15* and *PtoVHA-c*. Of them, three eQTNs (PtoeIF-2B_SNP4, PtoABF3_SNP19, and PtoLHCA4_SNP3) showed significant epistatic interactions with Chl; different combinations of these three SNPs had distinct contributions to the expression levels of *PtoHDA15* and *PtoVHA-c* (Figure 3D), implying that the three candidate genes might indirectly affect Chl. Additionally, the expression levels of *PtoHDA15* and *PtoVHA-c* were associated with PtoABF3_SNP1 and PtoLHCA4_SNP4 which displayed potential epistatic interactions for Pn (Supplementary Table S14). These results suggest that *PtoeIF-2B*, *PtoABF3*, and *PtoLHCA4* may associate with the expression levels of *PtoHDA15* and *PtoVHA-c*, which, in turn, may affect Pn by regulating Chl (Figure 5E; Supplementary Table S14).

Association mapping showed that PtoPSB33_SNP4 (G/A) was significantly associated with Pn ($p = 7.01 \times 10^{-5}$, $q < 0.05$, $R^2 = 12.15\%$) and epistatic interact with four SNPs (PtoABF3_SNP1, PtoABF3_SNP2, PtoABF3_SNP3, and PtoABF3_SNP4) in *PtoABF3* for Pn (Figures 4G,H). Moreover, *PtoPSB33* and *PtoABF3* both had eQTNs associated with the expression levels of *PtoHDA15* and *PtoVHA-c*, which were negatively correlated with Pn variation ($r_1 = -0.66$, $r_2 = -0.83$, respectively, $p < 0.01$) (Figure 4I). In total, four candidate genes (*PtoeIF-2B*, *PtoABF3*, *PtoPSB33*, and *PtoLHCA4*) were identified by both association mapping and co-expression network analysis; they were considered hub genes for regulating the drought stress response in poplar. A phylogenetic tree based on the protein sequences of the four genes showed that all clustered in the same group in the five poplar species, indicating that they were highly conserved in poplar (sequence similarities 95.93%–99.42%; Supplementary Figure S5). In addition, the expression patterns of these four potential hub genes were similar in the five poplar species (Supplementary Figure S6). Finally, we identified four hub candidate genes (*PtoeIF-2B*, *PtoABF3*, *PtoPSB33*, and *PtoLHCA4*) that formed a conserved network with *PtoHDA15* and *PtoVHA-c* during the response to drought stress (Supplementary Figure S7). *PtoeIF-2B* is

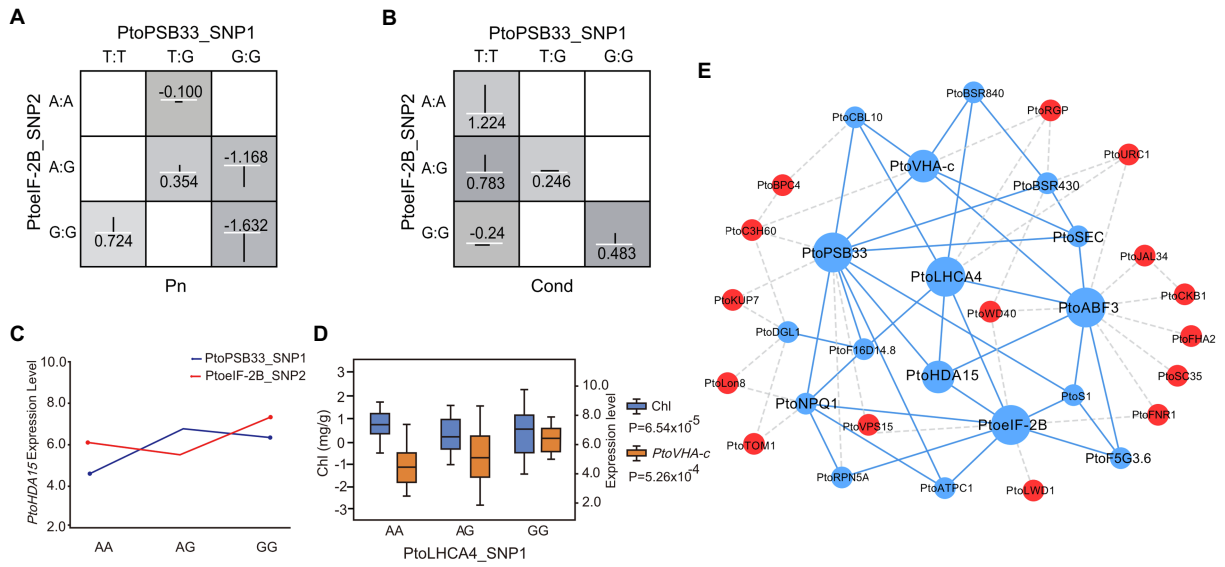


FIGURE 5
Integration of epistasis and eQTN analyses identifies factors involved in the response to drought stress. **(A)** and **(B)** Photosynthetic-trait epistatic effects of different genotypic combinations. **(C)** Pairwise interactions between PtoPSB33_SNP1 and PtoeIF-2B_SNP2 associate with the expression of *PtoHDA15* with different genotypic combinations at the two loci. **(D)** Box plots for Chl (blue) and *PtoVHA-c* expression (orange) plotted as an effect of genotype at PtoLHCA4_SNP1. **(E)** Proposed network of interactions among genetic factors. Putative regulatory network of candidate genes constructed by association mapping and co-expression analysis. Blue circles represent genes with epistatic interactions while red circles represent association mapping-verified regulators. Dotted lines indicate potential interactions, including co-expression interactions; solid lines indicate interactions verified by association mapping and co-expression analysis.

a potential regulator associated with the expression levels of *PtoHDA15* and *PtoVHA-c* (Figure 4F). Moreover, *PtoeIF-2B* showed significant epistatic interactions with *PtoABF3* and *PtoLHCA4* (Figures 3B,E), different genotypic combinations had distinct contributions to the expression levels of *PtoHDA15* and *PtoVHA-c*. *PtoPSB33* and *PtoABF3* formed epistatic interaction networks for Pn (Figures 4G,H); they jointly effected the expression levels of *PtoHDA15* and *PtoVHA-c*. Our results suggested that these hub genes have conserved genetic effects in *Populus* on transcription role and phenotypic variations.

Discussion

Drought stress is one of the most drastic abiotic stresses in plants, and regulatory factors that mediate the response to drought stress have been identified by reverse genetics (Zhou et al., 2007; Bang et al., 2018). Here, we used forward genetics to identify four important genes in the response to drought stress in *Populus*; our work provides an important theoretical foundation for the genetic improvement of drought stress tolerance in forest trees. The drought stress response is a complex multi-gene process. Thus, we used co-expression analysis, association genetics, and eQTN mapping to explore the genetic regulatory network of the drought stress response in *Populus*. We identified a conserved network and four key genes that are putatively involved in the drought response in *Populus*. This strategy enabled the identification of modules of co-expressed genes across multiple spatial, temporal, and environmental conditions, thereby providing insights into the co-expression network and candidate genes potentially responsible for the plant drought stress response.

Physiological and transcriptional regulation responses of *Populus* to drought stress

The drought stress response involves hydraulic signals, reactive antioxidants, osmotic regulation, and phytohormone movement processes (Attipalli et al., 2004; Ahmad et al., 2010). Compared with annual crops, perennial forest trees are exposed to long-term, complex external environmental conditions; they have evolved adaptive traits to manage drought stress (Lu et al., 2021). In this study, Pn, Cond, and Trmmol decreased under drought conditions (Supplementary Table S1). In contrast, we found that plant leaves had reduced Chl under drought stress, which might explain their lower rate of photosynthesis (Guo et al., 2018). The correlations among the six drought stress-related traits suggest that these traits jointly response to drought stress. In addition, the phenotypes of six drought stress-related traits displayed considerable variation, with the coefficient of variation values that ranged from 0.15 to 7.59 and H^2 values that ranged from 0.29 to 0.87 (Supplementary Table S1). Collectively, the above findings demonstrated that the six drought stress-related traits were suitable for investigating the genetic control of poplar drought tolerance by association mapping.

Transcriptionally co-regulated and functionally related genes have been identified by co-expression analysis. The transcriptomic profile of the genus *Populus* under drought stress has been investigated (Street et al., 2006; Viger et al., 2016; Garcia et al., 2019). We evaluated the transcriptome profiles of five *Populus* species under multiple spatial, temporal, and drought stress conditions to identify drought stress response genes that were conserved during the evolution of angiosperms. Our co-expression network analysis revealed 75 candidate genes that were implicated in the response to drought stress, with edge weight ≥0.5 and node connectivity ≥10 in the network (Supplementary Table S9). Some of these genes were reported to participate in the drought stress response, consistent with the observation that drought stress involves a complex regulatory network (Georgii et al., 2019). For example, *ABF3* plays an important role in the regulation of the drought response by interacting with the ABA-independent proteins DREB2A, DREB1A, and DREB2C in the *Arabidopsis* (Liu et al., 2018). *PtoVHA-c* confers stress tolerance by enhancing superoxide dismutase and peroxidase activities under drought stress (Feng et al., 2015). In addition, many novel high-degree hub gene signatures were identified in our analysis. For example, *RPN5A* is a 26S proteasome subunit, which degrades a wide range of intracellular proteins (Book et al., 2009). These findings underscore the ability of co-expression analysis to identify genes implicated in the responses to abiotic stresses.

Multi-omics analysis as a high-confidence strategy to assess drought stress

The drought stress response in trees involves multiple interconnected molecular pathways, which modulate various cellular functions (Gupta et al., 2020). Unlike most crops, trees form large continuous natural populations in highly heterogeneous environments that harbor significant genetic diversity, thus promoting phenotypic response to environmental (Di Filippo et al., 2015). Therefore, the construction of a drought stress systematic regulatory network and identification of potential regulatory genes would provide insights into the evolutionary background of perennial trees and promote plant breeding on the basis of specific regional climate (Lu et al., 2021). We constructed three co-expression network modules, and identified 75 hub genes based on edge weight and node connectivity; our results enabled the identification of a putative master regulator of the drought stress response (Supplementary Table S9). General regulator network based on 75 hub genes was constructed by systematic integration of association mapping analysis in drought stress. The most likely candidate genes in the drought stress response were *PtoeIF-2B*, *PtoPSB33*, *PtoABF3*, and *PtoLHCA4*. This integrative strategy has facilitated the functional interpretation of complex trait-associated signals in other studies and enables the identification of target traits and functionally associate genes (Nica et al., 2010). Furthermore, epistasis analysis allows the identification of the functional allele pairs that contribute to drought stress traits (Xiao et al., 2019). *PtoeIF-2B*, *PtoABF3*, and *PtoLHCA4* showed significant epistatic interactions affecting Chl. Different genotypic combinations of *PtoeIF-2B*, *PtoABF3* had distinct contributions to the expression levels of *PtoHDA15* and *PtoVHA-c* expression (Figure 4; Supplementary Table S14). Indeed, *eIF-2B* and *Lhca4* were down-regulated under drought stress in a drought sensitive wheat cultivar (Abbasi et al., 2021). These findings improve our understanding of the role of epistasis in drought stress adaptation in trees (Du et al., 2019).

The analysis of naturally occurring allelic variance and allele frequency in different climatic regions provides insights into adaptive evolution (Mitchell-Olds and Schmitt, 2006). Here, we found PtoeIF-2B_SNP4 was annotated as the important drought stress-responsive gene *PtoeIF-2B*, located in a selective sweep region; the frequency of allele PtoeIF-2B_SNP4 (increased chlorophyll content) increased from S, to NE, to NW. This suggested that genetic loci related to chlorophyll content might be associated with drought stress adaptation in *Populus* (Supplementary Table S16). Therefore, these selected loci showed significant regional differentiation, highlighting the potential roles in the response to drought stress (Kurasch et al., 2017). Haplotype data from population samples contain information regarding the history of allelic associations, which may aid in forest tree conservation (Leitwein et al., 2019). In this study, the haplotype frequency was higher in individuals in the NW than in the S or NE, implying that the lower annual rainfall led to selection for Hap_01. Combining co-expression analysis, association analysis, and eQTN mapping enabled analysis of quantitative traits in complex regulatory networks. Using this strategy, we constructed a genetic network of genes in trees under drought stress, which will promote forest tree genetic improvement programs and provide diagnostic tools for the conservation and management of natural populations (Neale and Kremer, 2011).

Functional interpretation of four conserved candidate genes associated with drought stress

Based on a multi-omics strategy, *PtoeIF-2B*, *PtoPSB33*, *PtoABF3*, and *PtoLHCA4* were identified as the four candidate genes response for drought stress in *Populus*. Notably, *PtoPSB33*, *PtoeIF-2B*, and *PtoABF3* showed epistatic interactions occur between allelic variation at different loci, which in turn can have an effect on the traits (Figures 3B, 4H). *PtoPSB33*, *PtoeIF-2B* affected the expression levels of *PtoHDA15* and *PtoVHA-c*, which were negatively correlated with Chl and Pn (Figures 4F,I). Moreover, different genotypic combinations of *PtoLHCA4* had distinct contributions to the expression levels of *PtoVHA-c* (Figure 5D). Therefore, *PtoeIF-2B*, *PtoPSB33*, *PtoABF3*, and *PtoLHCA4* jointly affected the expression levels of *PtoHDA15* and *PtoVHA-c*; they also affected Pn by regulating the Chl content under drought stress, suggesting that these hub genes have conserved genetic effects in five poplar species. Future studies should investigate the allelic coordination between drought stress and physiological functions. *PtoHDA15* and *PtoVHA-c* have important roles in the plant stress response and tolerance (Nakashima and

Yamaguchi-Shinozaki, 2013; Kato et al., 2017). Liu et al. (2013) reported that *AtHDA15*, a homolog of *PtoHDA15* in *Arabidopsis thaliana*, acted as a transcriptional repressor and negatively regulated levels of genes linked to chlorophyll biosynthesis and photosynthesis. Our genetic analysis indicated that *PtoeIF-2B*, *PtoPSB33*, *PtoABF3*, and *PtoLHCA4* were key genes in the genetic association network of the response to drought stress in poplar (Figure 5E). We propose the following mechanisms for this association. *PtoeIF-2B*, *PtoABF3*, *PtoLHCA4*, and *PtoPSB33* presumably constitute the initial defense response, which also induces the expression of downstream genes essential for adaptation to environmental stress. In addition, these four candidate genes exhibited significant differentiation during the evolution of dicots and monocots, and they were highly conserved in poplar (Supplementary Figure S5). It is unclear that four candidate genes ancestral by purifying selection, or independent positive selection in different lineages. However, this conservation enables the mining of drought stress-related genes in plant species. The drought stress-related genes in *Populus* identified in our study can serve as a useful resource for other species.

In summary, co-expression analysis, association genetics, and eQTN mapping enable dissection of the complex genetic networks of quantitative traits, such as drought tolerance. This integration strategy is a considerable advancement from the approach to facilitate an integrated conservation genomics approach for assessment of the effects of genetics and environment on adaptive traits (Ouborg et al., 2010). Our method allowed exploration of the association relationships of drought stress genes and provided a basis for understanding the complex genetic regulation involved. However, because the mechanisms that underlie the interactions between epistasis and eQTNs are unclear, functional analyses are required (Ingvarsson and Street, 2010). Continued use of genome-editing techniques in *Populus* will decipher the functions of candidate genes (e.g., *PtoeIF-2B*, *PtoABF3*, *PtoLHCA4*, and *PtoPSB33*) involved in drought stress in *Populus* (Zhou et al., 2015). Further analysis of single nucleotide substitutions will provide insights into the mechanism of allelic interactions in poplar.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found at: https://ngdc.cncb.ac.cn/, CRA005557.

Author contributions

DZ designed the experiments, obtained the funding, and is responsible for this article. FS, QD, MQ, LX, and WL performed the experiments. JZ, SQ, YF, DW, and PL collected and analyzed the data. FS wrote the manuscript. YE-K revised the manuscript and provided valuable suggestions concerning the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The present study was supported by The Major Science and Technology Projects of Inner Mongolia Autonomous Region (2021ZD0008), the Project of the National Natural Science Foundation of China (nos. 31872671 and 32170370), and the 111 Project (no. B20050).

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.829888/full#supplementary-material

References

Abbasi, A. Z., Bilal, M., Khurshid, G., Yiotis, C., Zeb, I., Hussain, J., et al. (2021). Expression of cyanobacterial genes enhanced CO_2 assimilation and biomass production in transgenic *Arabidopsis thaliana*. *PeerJ.* 9:27. doi: 10.7717/peerj.11860

Ahmad, P., Jaleel, C. A., Salem, M. A., Nabi, G., and Sharma, S. (2010). Roles of enzymatic and nonenzymatic antioxidants in plants during abiotic stress. *Crit. Rev. Biotechnol.* 30, 161–175. doi: 10.3109/07388550903524243

Attipalli, R. R., Kolluru, V. C., and Munusamy, V. (2004). Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. *J. Plant Physiol.* 161, 1189–1202. doi: 10.1016/j.jplph.2004.01.013

Bang, S. W., Lee, D.-K., Jung, H., and Chung, P. J. (2018). Overexpression of *OsTF1L*, a rice HD-zip transcription factor, promotes lignin biosynthesis and stomatal closure that improves drought tolerance. *Plant Biotechnol. J.* 17, 118–131. doi: 10.1111/pbi.12951

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Beaulieu, J., Doerksen, T., Boyle, B., Clement, S., Deslauriers, M., Beauseigle, S., et al. (2011). Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. *Genetics* 188, 197–214. doi: 10.1534/genetics.110.125781

Book, A. J., Smalle, J., Lee, K.-H., Yang, P., Walker, J. M., Casper, S., et al. (2009). The *RPN5* subunit of the 26s proteasome is essential for gametogenesis, sporophyte

development, and complex assembly in *Arabidopsis*. *Plant Cell* 21, 460–478. doi: 10.1105/tpc.108.064444

Boyes, D. C. (2001). Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* 13, 1499–1510. doi: 10.1105/tpc.13.7.1499

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Cheng, M. C., Liao, P. M., Kuo, W. W., and Lin, T. P. (2013). The *Arabidopsis* ETHYLENE RESPONSEFACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *Plant Physiol.* 162, 1566–1582. doi: 10.1104/pp.113.221911

Cline, M. S., Smoot, M., and Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* 2, 2366–2382. doi: 10.1038/nprot.2007.324

Deng, M., Li, D., Luo, J., Xiao, Y., Liu, H., Pan, Q., et al. (2017). The genetic architecture of amino acids dissection by association and linkage analysis in maize. *Plant Biotechnol. J.* 15, 1250–1263. doi: 10.1111/pbi.12712

Di Filippo, A., Pederson, N., Baliva, M., Brunetti, M., Dinella, A., Kitamura, K., et al. (2015). The longevity of broadleaf deciduous trees in northern hemisphere temperate forests: insights from tree-ring series. *Front. Ecol. Evol.* 3. doi: 10.3389/fevo.2015.00046

Du, Q., Tian, J., Yang, X., Pan, W., Xu, B., Li, B., et al. (2015). Identification of additive, dominant, and epistatic variation conferred by key genes in cellulose biosynthesis pathway in *Populus tomentosa*. *DNA Res.* 22, 53–67. doi: 10.1093/dnares/dsu040

Du, Q., Yang, X., Xie, J., Quan, M., Xiao, L., Lu, W., et al. (2019). Time-specific and pleiotropic quantitative trait loci coordinately modulate stem growth in Populus. *Plant Biotechnol. J.* 17, 608–624.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: asimulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Feng, L., Ding, H., Wang, J., Wang, M., Xia, W., and Zang, S. (2015). Molecular cloning and expression analysis of *RrNHX1* and *RrVHA-c* genes related to salt tolerance in wild *Rosa rugosa*. *Saudi J Biol Sci.* 22, 417–423. doi: 10.1016/j.sjbs.2015.01.008

Frimpong, F., Windt, C. W., van Dusschoten, D., Naz, A. A., Frei, M., and Fiorani, F. (2021). A wild allele of Pyrroline-5-carboxylate synthase1 leads to proline accumulation in spikes and leaves of barley contributing to improved performance under reduced water availability. *Front. Plant Sci.* 12:633448. doi: 10.3389/fpls.2021.633448

Garcia, B. J., Labbé, J. L., Jones, P., Abraham, P. E., Hodge, I., Climer, S., et al. (2019). Phytobiome and transcriptional adaptation of *Populus deltoides* to acute progressive drought and cyclic drought. *Phytobiomes J.* 2, 249–260. doi: 10.1094/PBIOMES-04-18-0021-R

Georgii, E., Kugler, K., Pfeifer, M., Vanzo, E., Block, K., Domagalska, M. A., et al. (2019). The systems architecture of molecular memory in poplar after abiotic stress. *Plant Cell* 31, 346–367. doi: 10.1105/tpc.18.00431

Grote, E. E., Belnap, J., Housman, D. C., and Sparks, J. P. (2010). Carbon exchange in biological soil crust communities under differential temperatures and soil water contents: implications for global change. *Glob. Chang. Biol.* 16, 2763–2774. doi: 10.1111/j.1365-2486.2010.02201.x

Guerra, F. P., Wegrzyn, J. L., Sykes, R., Davis, M. F., Stanton, B. J., and Neale, D. B. (2013). Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytol.* 197, 162–176. doi: 10.1111/nph.12003

Guo, Z., Yang, W., Chang, Y., Ma, X., Tu, H., Xiong, F., et al. (2018). Genome-wide association studies of image traits reveal the genetic architecture of drought resistance in rice. *Mol. Plant* 11, 789–805. doi: 10.1016/j.molp.2018.03.018

Gupta, A., Rico-Medina, A., and Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science* 368, 266–269. doi: 10.1126/science.aaz7614

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19, 376–382. doi: 10.1093/bioinformatics/btf869

Hardy, O. J., and Vekemans, X. (2002). Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

He, F., Wang, H.-L., Li, H.-G., Yanyan, S., Li, S., Yang, Y., et al. (2018). *PeCHYR1*, a ubiquitin E3 ligase from *Populus euphratica*, enhances drought tolerance via ABA-induced stomatal closure by ROS production in *Populus*. *Plant Biotechnol. J.* 16, 1514–1528. doi: 10.1111/pbi.12893

Huang, Z. H. (1992). The study on the climatic regionalization of the distributional region of *Populus tomentosa*. *J Beijing Forestry Univ.* 14, 26–32.

Ingvarsson, P. K., and Street, N. R. (2010). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x

Jansson, S., and Douglas, C. J. (2007). *Populus*: a model system for plant biology. *Annu. Rev. Plant Biol.* 58, 435–458. doi: 10.1146/annurev.arplant.58.032806.103956.peng

Kato, Y., Yokono, M., Akimoto, S., Takabayashi, A., Tanaka, A., and Tanaka, R. (2017). Deficiency of the Stroma-lamellar protein *LIL8/PSB33* affects energy transfer Around PSI in *Arabidopsis*. *Plant Cell Physiol.* 58, 2026–2039. doi: 10.1093/pcp/pcx124

Kurasch, A. K., Hahn, V., Leiser, W. L., Vollmann, J., Schori, A., Bétrix, C.-A., et al. (2017). Identification of mega-environments in Europe and effect of allelic variation at maturity E loci on adaptation of European soybean. *Plant Cell Environ.* 40, 765–778. doi: 10.1111/pce.12896

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

Lee, J., Han, C.-T., Kim, H. R., and Hur, Y. (2014). A MORN-domain protein regulates growth and seed production and enhances freezing tolerance in Arabidopsis. *Plant Biotechnol. Rep.* 8, 229–241. doi: 10.1007/s11816-014-0315-6

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P.-A., and Bernatchez, L. (2019). Using Haplotype Information for Conservation Genomics. *Trends Ecol. Evol.* 35. doi: 10.1016/j.tree.2019.10.012

Liu, X., Chen, C. Y., Wang, K. C., Luo, M., Tai, R., Yuan, L., et al. (2013). PHYTOCHROME INTERACTING FACTOR3 associates with the histone deacetylase *HDA15* in repression of chlorophyll biosynthesis and photosynthesis in etiolated *Arabidopsis* seedlings. *Plant Cell* 25, 1258–1273. doi: 10.1105/tpc.113.109710

Liu, S., Lv, Z., Liu, Y., Li, L., and Zhang, L. (2018). Network analysis of ABA-dependent and ABA-independent drought responsive genes in *Arabidopsis thaliana*. *Genet. Mol. Biol.* 41, 624–637. doi: 10.1590/1678-4685-gmb-2017-0229

Lu, W., Du, Q., Xiao, L., Lv, C., Quan, M., Li, P., et al. (2021). Multi-omics analysis provides insights into genetic architecture of flavonoid metabolites in *Populus*. *Ind. Crop. Prod.* 168:113612. doi: 10.1016/j.indcrop.2021.113612

Lüttschwager, D., Ewald, D., and Alía, L. A. (2015). Comparative examinations of gas exchange and biometric parameters of eight fast-growing poplar clones. *Acta Physiol. Plant.* 37, 214. doi: 10.1007/s11738-015-1968-7

Lv, C., Lu, W., Quan, M., Xiao, L., Li, L., Zhou, J., et al. (2021). Pyramiding superior haplotypes and epistatic alleles to accelerate wood quality and yield improvement in poplar breeding. *Ind. Crop. Prod.* 171:113891. doi: 10.1016/j.indcrop.2021.113891

Ma, L., Runesha, H. B., Dvorkin, D., Garbe, J. R., and Da, Y. (2008). Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics.* 9, 315. doi: 10.1186/1471-2105-9-315

Mackay, T. F. C. (2013). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* 15, 22–33. doi: 10.1038/nrg3627

Mao, Y., London, N. R., Ma, L., Dvorkin, D., and Da, Y. (2006). Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. *Physiol. Genomics* 28, 46–52. doi: 10.1152/physiolgenomics.00096.2006

Mitchell-Olds, T., and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature* 441, 947–952. doi: 10.1038/nature04878

Mukrimin, M., Kovalchuk, A., Neves, L. G., Jaber, E. H. A., Haapanen, M., Kirst, M., et al. (2018). Genome-wide exon-capture approach identifies genetic variants of Norway spruce genes associated with susceptibility to *Heterobasidion parviporum* infection. *Front. Plant Sci.* 9:793. doi: 10.3389/fpls.2018.00793

Nakashima, K., and Yamaguchi-Shinozaki, K. (2013). ABA signaling in stress-response and seed development. *Plant Cell Rep.* 32, 959–970. doi: 10.1007/s00299-013-1418-1

Neale, D. B., and Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* 12, 111–122. doi: 10.1038/nrg2931

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895. doi: 10.1371/journal.pgen.1000895

Ouborg, N., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., and Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends Genet.* 26, 177–187. doi: 10.1016/j.tig.2010.01.001

Quan, M., Du, Q., Xiao, L., Lu, W., Wang, L., Xie, J., et al. (2019). Genetic architecture underlying the lignin biosynthesis pathway involves noncoding RNA s and transcription factors for growth and wood properties in *Populus*. *Plant Biotechnol. J.* 17, 302–315. doi: 10.1111/pbi.12978

Ravasz, E. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Revelle, W. R. (2017). psych: Procedures for Personality and Psychological Research.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rogers, B. M., Solvik, K., Hogg, E. H., Junchang, J., Masek, J. G., Michaelian, M., et al. (2018). Detecting early warning signals of tree mortality in boreal North America using multiscale satellite data. *Glob. Chang. Biol.* 24, 2284–2304. doi: 10.1111/gcb.14107

Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M., and Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Smita, S., Katiyar, A., Pandey, D. M., Chinnusamy, V., Archak, S., and Bansal, K. C. (2013). Identification of conserved drought stress responsive gene-network across tissues and developmental stages in rice. *Bioinformation* 9, 72–78. doi: 10.6026/97320630009072

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31, 2013–2035. doi: 10.1214/aos/1074290335

Street, N. R., Skogström, O., Sjödin, A., Tucker, J., Rodríguez-Acosta, M., Nilsson, P., et al. (2006). The genetics and genomics of the drought response in *Populus*. *Plant J.* 48, 321–341. doi: 10.1111/j.1365-313X.2006.02864.x

Taylor, G. (2002). *Populus*: *Arabidopsis* for forestry. Do we need a model tree? *Ann. Bot.* 90, 681–689. doi: 10.1093/aob/mcf255

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Viger, M., Smith, H. K., Cohen, D., Dewoody, J., Trewin, H., Steenackers, M., et al. (2016). Adaptive mechanisms and genomic plasticity for drought tolerance identified in European black poplar (*Populus nigra L.*). *Tree Physiol.* 36, 909–928. doi: 10.1093/treephys/tpw017

Vitasse, Y., Bottero, A., Cailleret, M., Bigler, C., Fonti, P., Gessler, A., et al. (2019). Contrasting resistance and resilience to extreme drought and late spring frost in five major European tree species. *Glob. Chang. Biol.* 25, 3781–3792. doi: 10.1111/gcb.14803

Wegrzyn, J. L., Eckert, A. J., Choi, M., Lee, J. M., Stanton, B. J., Sykes, R., et al. (2010). Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, salicaceae) secondary xylem. *New Phytol.* 188, 515–532. doi: 10.1111/j.1469-8137.2010.03415.x

Xiao, L., Liu, X., Wenjie, L., Chen, P., Quan, M., et al. (2019). Genetic dissection of the gene co-expression network underlying photosynthesis in *Populus*. *Plant Biotechnol. J.* 18, 1015–1026. doi: 10.1111/pbi.13270

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13, 134. doi: 10.1186/1471-2105-13-134

Zhao, T., Yang, X., Rao, P., An, X. M., and Chen, Z. (2021). Identification of key flowering-related genes and their seasonal expression in *Populus tomentosa* reproductive buds suggests dual roles in floral development and dormancy. *Ind. Crop. Prod.* 161:113175. doi: 10.1016/j.indcrop.2020.113175

Zhou, X., Jacobs, T. B., Xue, L.-J., Harding, S. A., and Tsai, C.-J. (2015). Exploiting SNPs for biallelic CRISPR mutations in the outcrossing woody perennial *Populus* reveals 4-coumarate: CoA ligase specificity and redundancy. *New Phytol.* 208, 298–301. doi: 10.1111/nph.13470

Zhou, J., Wang, X., Jiao, Y., Qin, Y., Liu, X., He, K., et al. (2007). Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle. *Plant Mol. Biol.* 63, 591–608. doi: 10.1007/s11103-006-9111-1

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 September 2022
DOI 10.3389/fpls.2022.970023

OPEN ACCESS

EDITED BY
Fang Du,
Beijing Forestry University, China

REVIEWED BY
Ye Ai,
Fujian Agriculture and Forestry University, China
Ruidong Jia,
Institute of Vegetables and Flowers (CAAS), China
Yunxing Zhang,
Henan Polytechnic University, China

*CORRESPONDENCE
Zhiqiang Wu
wuzhiqiang@caas.cn
Jin Ma
majinzjl@163.com.
Cuihua Gu
gucuihua@zafu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 15 June 2022
ACCEPTED 22 August 2022
PUBLISHED 07 September 2022

CITATION
Hong S, Wang J, Wang Q, Zhang G, Zhao Y, Ma Q, Wu Z, Ma J and Gu C (2022) Decoding the formation of diverse petal colors of *Lagerstroemia indica* by integrating the data from transcriptome and metabolome.
Front. Plant Sci. 13:970023.
doi: 10.3389/fpls.2022.970023

Decoding the formation of diverse petal colors of *Lagerstroemia indica* by integrating the data from transcriptome and metabolome

Sidan Hong[1,2,3†], Jie Wang[1,2,3,4,5,6†], Qun Wang[1,2,3†], Guozhe Zhang[1,2,3], Yu Zhao[1,2,3], Qingqing Ma[1,2,3], Zhiqiang Wu[4,5*], Jin Ma[1,2,3*] and Cuihua Gu[1,2,3*†]

[1]College of Landscape and Architecture, Zhejiang A&F University, Hangzhou, China, [2]Zhejiang Provincial Key Laboratory of Germplasm Innovation and Utilization for Garden Plants, Zhejiang A&F University, Hangzhou, China, [3]Key Laboratory of National Forestry and Grassland Administration on Germplasm Innovation and Utilization for Southern Garden Plants, Zhejiang A&F University, Hangzhou, China, [4]Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, [5]Kunpeng Institute of Modern Agriculture, Foshan, China, [6]College of Science, Health, Engineering and Education, Food Futures Institute, Murdoch University, Murdoch, WA, Australia

Lagerstroemia indica has great economic value due to its ecological, medicinal, and ornamental properties. Because its bloom color is one of the most essential characteristics, research into its color development is a hot topic. In this study, five representative colored cultivars were chosen, each representing a different color, such as white, red, pink, violet, and purple. Fully bloomed flowers were used to detect flavonoids in the petals. Anthocyanin is the main factor for the color formation of *L. indica*. 14 anthocyanins were discovered among the 299 flavonoids. Among 14 anthocyanins, malvidin-3,5-di-*O*-glucoside varied greatly among four colored samples and is the main contributor to color diversity. Transcriptome sequencing revealed that compared to white flowers, Anthocyanin pathway genes appear to be more active in colored samples. Analyzing the correlation network between metabolites and differential expressed genes, 53 key structural genes, and 24 TFs were detected that may play an essential role in the formation of color in *L. indica* flowers. Among these, the differential expression of *F3′5′H* and *F3′H* between all samples are contributors to color diversity. These findings lay the foundation for discovering the molecular mechanism of *L. indica* flower color diversity.

KEYWORDS

anthocyanin, breeding, color diversity, flower color, *Lagerstroemia indica*, metabolome

Introduction

Flavonoids/anthocyanins, carotenoids, and betalains are the most common pigments that provide plants their colors (Tanaka et al., 2008). Carotenoids are lipid-soluble chemicals that contribute to yellow to red color and are extensively distributed in seed plants that engage in photosynthesis (Szabolcs, 1989). Betalains are water-soluble chemicals that give an orange to red-purple coloration and are exclusively found in the order Caryophyllales (Timoneda et al., 2019). In addition, anthocyanins that belong to Flavonoids are the most common essential pigment and play a key role in flower color formation (Winkel Shirley, 2001). There are six major forms of anthocyanidins in the plant world based on the substituents of the benzene ring: delphinidin, petunidin, cyanidin, malvidin, pelargonidin, and peonidin (Zhao and Tao, 2015; Yonekura Sakakibara et al., 2019). Different types and content of anthocyanins in plant tissues result in the unique color of the flowers ranging from orange/red to purple/blue. Besides, anthocyanins are vital for heredity in plants because of attracting pollinators and animals that can help with broadcasting pollen and seeds (Landi et al., 2015). Moreover, it can also defend plants from biotic or abiotic stresses (Dao et al., 2011; Liang and He, 2018). Therefore, plants with abundant anthocyanins are more likely to be favored by consumers.

Pathway genes and transcription factors (TF) that regulate anthocyanin biosynthesis are relatively conserved in many plants (Hichri et al., 2011). One molecule of *p*-coumaroly-CoA and three molecules of malonyl-CoA are converted to naringenin chalcone by catalysis of chalcone synthase (CHS) (Cain et al., 1997). Then, chalcone isomerase (CHI) rapidly converts naringenin chalcone to naringenin. Subsequently, naringenin is converted to eriodictyol or pentahydroxy flavone by the catalysis of F3′H or F3′5′H respectively (Koes et al., 2005). Following catalysis of a series of downstream enzymes, dihydroflavonol reductase (DFR), anthocyanin synthase (ANS), *O*-methyltransferases (OMT), and UDP glycosyltransferase (UGT) various types of anthocyanins are produced (Irmisch et al., 2019). Thus, the flow direction of the carbon flux in the pathway branch is very important, especially in the downstream branch from naringenin. F3H has a direct link to the production of pelargonidin, which can result in red flowers (Lukačin et al., 2000). In cornflower, the varied expression of gene *F3′H* and *DFR* decides the different concentrations of cyanidin and pelargonidin respectively, and thus influences the formation of flower color (Deng et al., 2019). According to some studies, delphinidin is a major pigment of most blue flowers, F3′5′H is the key enzyme of delphinidin synthesis. Therefore, F3′5′H may have strong relevance in forming the blue-hued flowers (Tanaka, 2006). When compared to wild peas with violet-colored flowers, the absence of expression of *F3′5′H* in pink flowers leads to a reduction in the delphinidin content (Moreau et al., 2012). There is a similar issue in *Lyceum* and chrysanthemums (He et al., 2013). MYB-bHLH-WD40 (MBW) transcription complexes are critical for controlling the flavonoid pathway. The control mechanism of these three genes is critical for efficiently regulating the expression of key structural genes. R2R3-myeloblastosis (MYB), basic helix-loop-helix (bHLH), and WD40 protein make up the MBW. bHLH is a protein that is generally conserved. *AtTT8* can regulate the expression of DFR and ANS in the capsule of seeding in *Arabidopsis* which contributes to the increased accumulation of anthocyanins (Nesi et al., 2000). Moreover, WD40 is related to the synthesis of floral pigments (Brueggemann et al., 2010; Payyavula et al., 2013). Of all three types of TFs, MYB is the most significant. MYBs, as essential transcription factors, bind directly to others to stimulate gene expression. Four MYBs (*AtPAP1*, *AtPAP2*, *AtMYB113*, and *AtMYB114*) are identified in *Arabidopsis* that influence the accumulation of anthocyanins (Borevitz et al., 2000; Teng et al., 2005; Stracke et al., 2007; Gonzalez et al., 2008). *VvMYBA1* and *VvMYBA2* in grape (Walker et al., 2007), *MdMYB10,* and *MdMYB110a* in apple (Espley et al., 2007; Chagné et al., 2012), and *PpMYB10.1* are reported to be responsible for anthocyanin synthesis (Rahim et al., 2014). In addition, 49 candidate MYBs were identified in leaves of *L. indica* and it is shown that they had a favorable influence on the regulation of the flavonoid-anthocyanin pathway (Qiao et al., 2019).

Lagerstroemia indica is a species of the *Lagerstroemia* genus that originated in China and is now widespread in most Chinese cities (Liu et al., 2008). It is a type of garden tree with high decorative value and many wonderful ornamental characteristics, such as a long flowering period, distinctive floral shapes, and lovely features (Pounders et al., 2007). Among others, the color of its petals with a wide hue (white, pink, red, purple, violet, and their combined colors) is most prominent. However, consumers increasingly want flowers with more unique colors such as real blue, and current cultivars are no longer able to meet market demand (Cabrera, 2004). Some research has been conducted on pigment composition and a preliminary molecular investigation. Four types of anthocyanins are detected in the flowers (delphinidin 3-*O*-glucoside, petunidin 3-*O*-glucoside, malvidin 3-*O*-glucoside, and cyanidin 3-*O*-glucoside) (Zhang et al., 2008) and seven R2R3-MYB transcription factors (TFs) were identified as probable regulators of flower color (Yu et al., 2021). However, neither the metabolic nor molecular foundation for the *L. indica* color diversity is entirely clear, and the control of the anthocyanin pathway requires further investigation.

In summary, the key anthocyanins and the molecular mechanism causing the color diversity of crape myrtle are unclear. In this study, we performed metabolomics and transcriptomics analyses of five typical colored petal tissue of crape myrtle cultivars at full bloom stages. Differentially expressed metabolites (DEMs) and differentially expressed genes (DEGs) were identified and analyzed. In addition, the

molecular mechanisms of color diversity of *L. indica* petals would be discussed.

Materials and methods

Plant materials

In this investigation, five different *L. indica* cultivars were employed. The petal colors of the five cultivars were "Natchez" (white, WH), "Lihongtianyuan" (pink, PK), and "Honghuaguifei" (red, RD), "Zihuaguifei" (purple, PP), and "Lanzi" (violet, VT). All plant materials used were acquired from the Germplasm Resource Nursery of Crape Myrtle in the School of Landscape and Architecture, Zhejiang Agriculture and Forestry University (located at long. 118°51′ to 119°52′ E, lat. 29°56′ to 30°23′ N). Petal samples of five cultivars from the full bloom stage were collected (**Figure 1A**) from 8 am to 10 am. The color of petals were compared using CIELAB analysis by spectrophotometer. Five different flowers were chosen for analysis indoors, every flower was tested three times. Fresh petals were kept under −80°C for flavonoid and the total RNA extraction.

Content measurement of total flavonoids and total anthocyanins

About 0.1 g of samples were used for the determination of total anthocyanin and total flavonoids in each variety. After grinding 0.1 g petals with liquid nitrogen, 1mL extraction solution (0.1% formic acid methanol) was added immediately, and the sample was rapidly and fully mixed with the extraction solution by rapid shock for 30 s. After 24 h extraction at 4∘C under dark conditions, add the extraction solution to a constant volume of 3 mL. The absorbance was determined at 524and 325 nm by UV spectrophotometer (Shimadzu, UV2700), and methanol 0.1% formate was used as blank control. The absorbance was determined three times. The standard curve was prepared with 0.1, 0.05, 0.025, and 0.0125 mg·mL^{-1} chlorinated centathrin and rutin standard solution as the standard. The anthocyanin content in the sample was calculated as X = C × B/M. X is the anthocyanin content in the sample (mg.g^{-1}), C is the cornflower pigment content in the extraction liquid (mg·mL^{-1}), B is the volume of extraction liquid (mL), and M is the sample mass (g).

Qualitative and quantitative analysis of flavonoid

Fifteen different colored *L. indica* samples were used for the measurement of (Five cultivars, three biological repetitions each). First, a mixer mill was used to grind the freeze-dried samples (MM 400, Retsch). Then, a solution of 70% methanol in 1.2 mL was required to dissolve 100 mg of flower powder for the entire flavonoid extraction. The sample was rapidly and fully mixed with the extraction solution by rapid shock for 30 s every 30 min for a total of 6 times. After the samples were treated, they were stored overnight at 4°C. Finally, a nylon syringe filter was used to filter the supernatant (SCAA-104, 0.22 μm particle size; ANPEL, Shanghai, China[1]), and remove the sediment by centrifuging for 10 min at 12000 rpm. After sediment was removed and the liquid was clear, flavonoid analysis could be conducted both qualitatively and quantitatively (Fu et al., 2021). The UPLC-ESI-MS/MS equipment (UPLC, SHIMADZU Nexera X2; MS, Applied Biosystems 4500 Q TRAP) was used to examine all extraction solutions. **Supplementary Methods 1** and **2** would show the detailed condition of UPLC and ESI-Q TRAP-MS/MS.

Multiple reaction monitoring (MRM) was performed to identify the flavonoids in *L. indica* flowers based on the standard Metabolites Database, which is commercially accessible (Metware Biotechnology Co., Ltd., Wuhan, China). In accordance with secondary spectrum information collected by UPLC based on the system of MRM, signals were collected and analyzed using Analyst 1.6.3. (Metware Biotechnology Co., Ltd., Wuhan, China) to quantify and identify the flavonoids. PCA (principal component analysis) was performed for analyzing the biological repetition condition of fifteen samples (five cultivars, three biological replicates) (Jiang et al., 2020). Significantly differential accumulated flavonoids were analyzed by R package MetaboAnalystR (Chen et al., 2009) based on OPLS_DA (orthogonal partial least squares discriminant analysis) analysis (Thévenot et al., 2015). The standard of screening: VIP ≥ 1 and $|\log_2 FC| \geq 1$.

RNA extraction, sequencing, and analysis

Freeze-dried petals were grounded for RNA extraction. A Trizol reagent kit (Invitrogen, Carlsbad, CA, United States) was used to isolate total RNA from petals from five different cultivars. After monitoring total RNA degradation and agarose gel contamination, total RNA purity was determined using a NanoPhotometer® and spectrophotometer (IMPLEN, CA, United States). RNA Nano 6000 Assay Kit for the Bioanalyzer 2100 system was used to quantify total RNA concentration. The integrity of total RNA was detected by Qubit® RNA Assay Kit in the Qubit®2.0 Fluorometer. The cDNA library was made by Illumina's NEBNext® Ultra TM RNA Library Prep Kit and 1 μg of total RNA. Oligo(dT) magnetic beads were used to enrich RNA with polyA and disrupt it randomly. It was used to synthesize the first cDNA, and then used dNTPs as raw

1 http://www.anpel.com.cn/

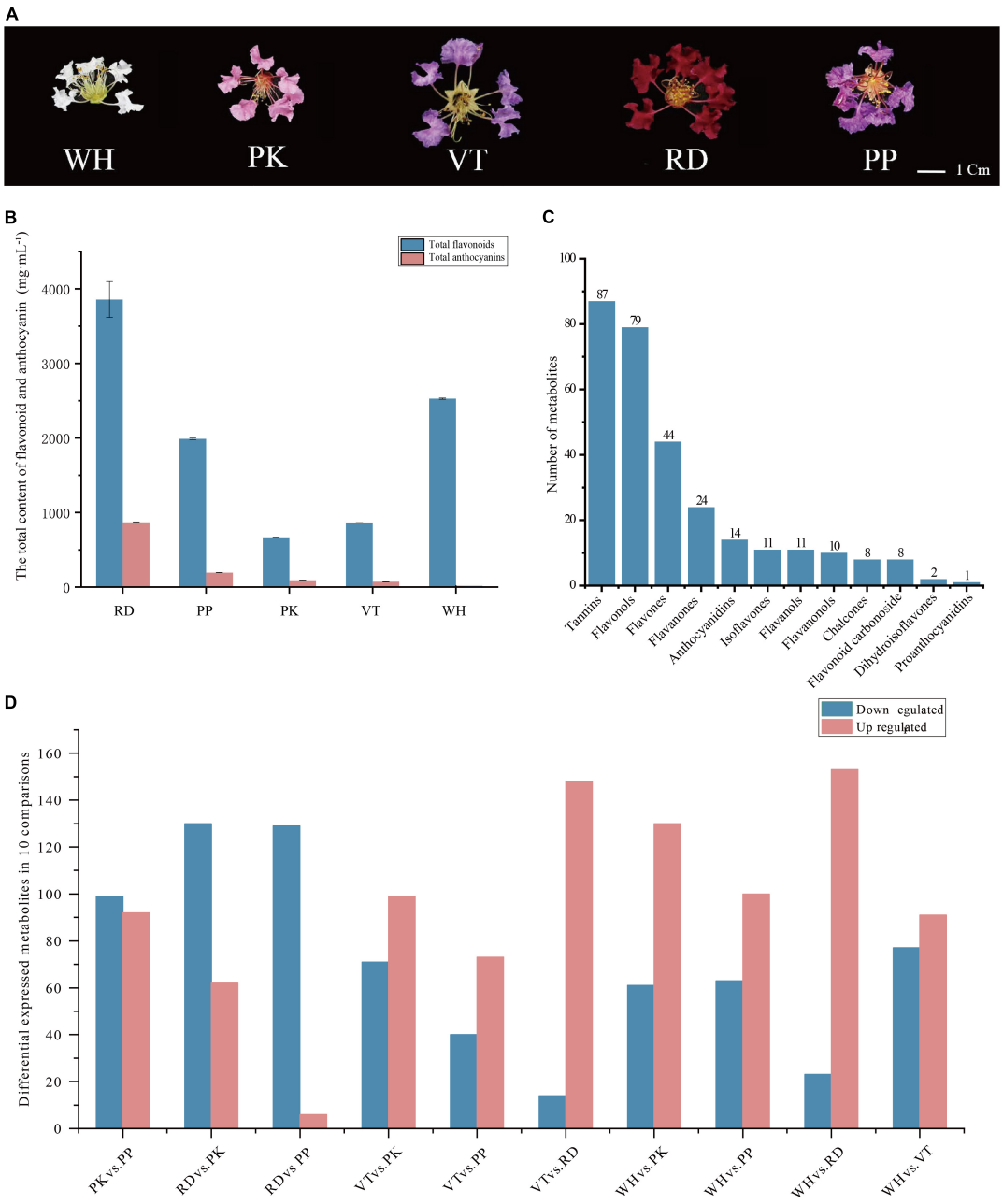


FIGURE 1
Phenotypes of comparisons, composition, and total content of flavonoids and anthocyanins among five *Lagerstroemia indica* petals. **(A)** Phenotypes of *L. indica* petals. **(B)** The total content of flavonoid and anthocyanin in five samples. **(C)** Classification and number of all flavonoids detected. **(D)** Number of differential accumulated metabolites between all comparisons.

materials and mRNA fragments as templates to synthesize the second cDNA. After screening cDNA with AMPure XP Beads, the PCR products were purified again using AMPure XP Beads to obtain the library, which was around 200 bp in length. After examining the quality of the library, all products were sequenced by Illumina.

Fastp v 0.19.3 was used to remove adapters, ploy-N, and low-quality reads from the original data for acquiring clean reads. Trinity was used to assemble the clean reads (v2.11.0) (Grabherr et al., 2011), and the assembled transcripts were clustered to eliminate redundancy using Corset. TransDecoder[2] was used to identify candidate coding regions within transcript sequences generated by *de novo* RNA-Seq transcript assembly

2 https://github.com/TransDecoder/

using Trinity. Subsequently, to acquire the annotation of unigenes, As a first step toward obtaining unigene annotation, DIAMOND BLASTX (Buchfink et al., 2015) was performed for functional gene prediction based on the Nr database (NCBI non-redundant), Swiss-prot protein database (Bairoch and Boeckmann, 1991), TrEMBL, KEGG database (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2007), as well as the COG (Clusters of Orthologous Groups of proteins) (Tatusov et al., 2000) and GO database (Gene Ontology) (Ashburner et al., 2000). Then, the amino acid sequence was aligned with the Pfam database by HMMER (Finn et al., 2013). The level of gene expression was estimated by RSEM (Li and Dewey, 2011), and the FPKM was calculated based on the gene length of each gene. With the standard of $|\text{fold change}| \geq 2$ as well as false discovery rate (FDR) < 0.5, DESeq2 (Love et al., 2014) detected DEGs. The enrichment analysis was performed based on the KEGG pathway and GO term. The PCA and sample correlation analysis are shown here in **Supplementary Figure 1**, and the data from PP2 were eliminated.

Based on iTAK (Zheng et al., 2016), hmmscan (Jin et al., 2013) was used to identify and annotate candidate transcription factors (TFs) that would regulate anthocyanin synthesis. Using DESeq2 to identify differentially expressed TFs using the same standard of pathway genes.

Interaction network between genes and metabolites

We calculated Pearson's correlation coefficients (PCC) based on the Metware Cloud[3] and displayed them using Cytoscape (Shannon et al., 2003) afterward to discover the interconnections between candidate genes, including TFs and pathway genes, and anthocyanin components (v.3.9.1). $|\text{PCC}| > 0.8$ as well as p-value 0.05 are the correlation analysis standards.

Results

Flower pigment of *Lagerstroemia indica*

The content of total anthocyanins and total flavonoids in petals of different flower colors was measured. RD had the highest total anthocyanins content (868.07 mg·mL^{-1}), followed by PP (194.19 mg·mL^{-1}), PK (92.63 mg·mL^{-1}), VT (70.1 mg·mL^{-1}), WH had the lowest anthocyanins content. RD had the highest total flavonoid content, WH was lower than RD, and PK had the lowest (**Figure 1B**). These results indicated that anthocyanins were the main factors to form the color of crape myrtle, and flavonoids were the main pigment of WH. It should be noted that the color lightness of PP (37.774 ± 1.27) was not significantly different from that of RD (22.188 ± 1.27) (**Supplementary Table 1**), While the content of anthocyanin in RD was nearly five times that of PP (**Figure 1B**). Flavonoids are important types of co-pigment, and co-pigment can make the pigment more stable, indicating that flavonoids in PP can better maintain the stability of pigment and make purple more obvious.

To better understand the metabolic changes between five different colored petals, we performed flavonoid analysis based on LC-ESI-MS/MS, and five typical cultivars were selected for this study. PCA was used to measure the trend of metabolic separation between all groups. The result of PCA showed that all samples were separated into five groups on the PC1 × PC2 score plot (**Supplementary Figure 1**). The results showed that the biological repeatability among the five groups was good, and intragroup correlations were high.

A total of 299 compounds were identified in total across all samples (**Figure 1C**). All metabolites can be categorized into 12 classes further, including tannins (87), flavonols (79), flavones (44), flavanones (24), anthocyanidins (14), isoflavones (11), flavanols (11), flavanonols (10), chalcones (eight), flavonoid carbonoside (eight), dihydroisoflavones (two), and proanthocyanidins (one).

From PCA of all samples, significant differences were shown between groups. OPLS_DA was performed for five groups of samples to analyze differential expression. Non-repetitive comparisons had Q2 values greater than 0.9, and the p-values of models were all below the threshold of 0.05, which indicated that models built by each group were capable of making accurate predictions (**Supplementary Figure 2**).

Differential accumulated metabolites were screened based on OPLS_DA model results (**Figure 1D**). There were 191 DEMs in PK vs. PP (99 down-regulated, 92 up-regulated), 192 DEMs in RD vs. PK (130 down-regulated, 62 up-regulated), 135 DEMs in RD vs. PP (129 down-regulated, 6 up-regulated), 170 DEMs in VT vs. PK (71 down-regulated, 99 up-regulated), 113 DEMs in VT vs. PP (40 down-regulated, 73 up-regulated), 162 DEMs in VT vs. RD (14 down-regulated, 148 up-regulated), 191 DEMs in WH vs. PK (61 down-regulated, 130 up-regulated), 163 DEMs in WH vs. PP (63 down-regulated, 100 up-regulated), 176 DEMs in WH vs. RD (23 down-regulated, 153 up-regulated), and 168 DEMs in WH vs. VT (77 down-regulated, 91 up-regulated).

Anthocyanins among the petal of *Lagerstroemia indica* cultivars

Anthocyanins were significant contributors to color formation. To know the certain type of anthocyanin that have a great influence on the color diversity of *L. indica*, the detected anthocyanins were analyzed further. 14 types of anthocyanin were detected from five cultivars (**Figure 2A**).

3 https://cloud.metware.cn/#/tools/tool-list

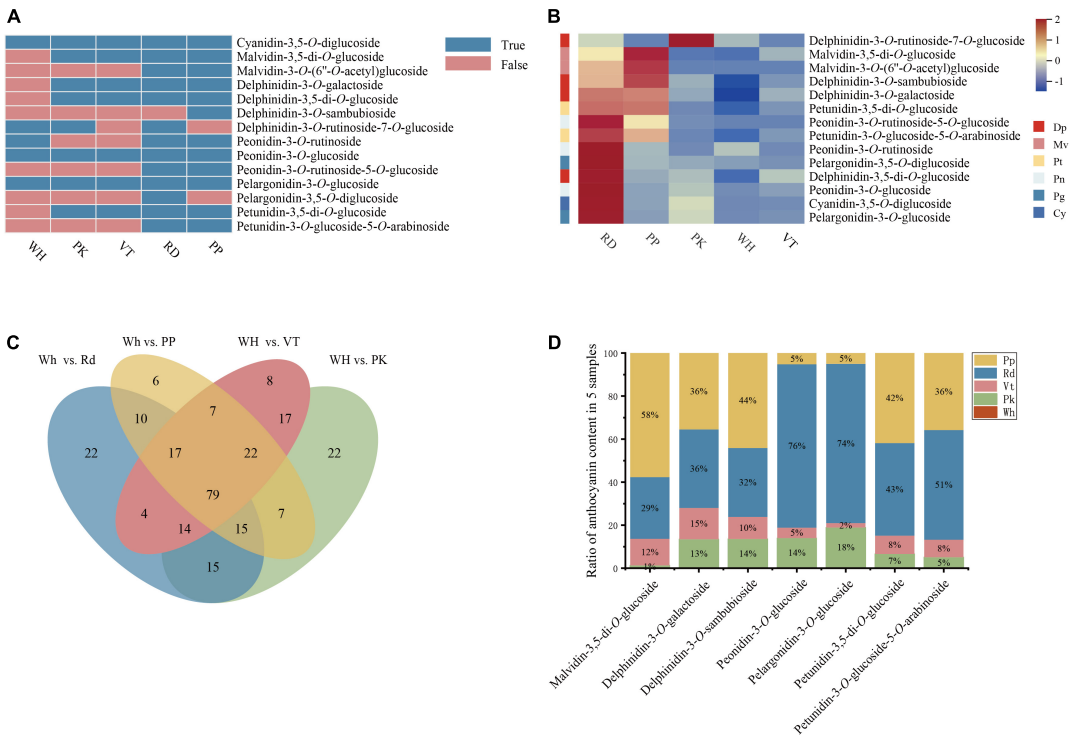


FIGURE 2
Composition and content of anthocyanin among five *L. indica* petals. **(A)** Type of anthocyanin detected in *L. indica* petals. **(B)** Heatmap of all anthocyanins. **(C)** Venn analysis among WH vs. RD, WH Vs. PP, WH vs. VT, and WH vs. PK. **(D)** The ratio of relative content of overlapped DAAs in five samples.

There were four delphinidin derivatives, two malvidin derivatives, three peonidin derivatives, two pelargonidin derivatives, two petunidin derivatives, and one cyanidin derivative. From the heatmap of all detected anthocyanins (**Figure 2B**), derivatives of peonidin, pelargonidin, and Cyanidin had higher accumulation levels in RD and PK than in other samples. These pigments can form red color. While accumulation level of delphinidin derivatives and malvidin derivatives was higher in PP and VT, and those two types of anthocyanin had high relation to purple-blue-hued flower color. Besides, malvidin-3-*O*-(6″-*O*-acetyl) glucoside and petunidin-3-*O*-glucoside-5-*O*-arabinoside were only found in PP and RD. Pelargonidin-3,5-*O*-diglucoside and delphinidin-3-*O*-sambubioside were specifically accumulated in RD and PP respectively. Delphinidin-3-*O*-rutinoside-7-*O*-glucoside was only detected in PK, RD, and WH, and it was accumulated higher in PK than in RD and WH. Malvidin-3-*O*-(6″-*O*-acetyl) glucoside, pelargonidin-3,5-*O*-diglucoside, delphinidin-3-*O*-sambubioside, and delphinidin-3-*O*-rutinoside-7-*O*-glucoside might have an important role in forming the red color, purple color, and pink color.

To know the formation of petal colors, a Venn diagram was built with four comparisons (WH vs. RD, WH vs. PK, WH vs. PP, WH vs. VT), and there were 79 overlapped differential accumulated metabolites (**Figure 2C**). Among these, there were seven differential accumulated anthocyanins (DAAs) up-regulated in PK, RD, PP, and VT compared with WH, respectively, including delphinidin-3-*O*-galactoside, delphinidin-3-*O*-sambubioside, petunidin-3-*O*-glucoside-5-*O*-arabinoside, petunidin-3,5-di-*O*-glucoside, malvidin-3,5-di-*O*-glucoside, peonidin-3-*O*-glucoside, pelargonidin-3-*O*-glucoside. Compared with the fold change value of overlapped differential accumulated metabolites between four colored sample comparisons, malvidin-3,5-di-*O*-glucoside had a great value (**Supplementary Table 2**), which indicated that it might be the key pigment that influences color diversity in crape myrtle flowers (**Figure 2D**).

Analysis of the RNA-seq data

Flower petals of *L. indica* were used to perform RNA-seq for potential molecular mechanism analysis. After filtering low-quality reads from the raw data, fifteen samples (5 cultivars and three biological replicates each) were used to establish libraries. A total of 99.54 Gb clean reads were generated, and 6.09−8.89 Gb for each cultivar was obtained by sequence The proportion of both Q20 and Q30 bases in each library was

greater than 90% respectively. And the GC content was ranging from 49.1% to 52.02% (**Supplementary Table 3**). 149,235 Transcripts were assembled by Trinity as a reference sequence. After eliminating redundancy, 141,673 unigenes were generated with an N50 length of 1,701 bp, an N90 length of 430 bp, and the mean length of 1,045 bp (**Supplementary Table 4**). The length of unigene around 300−400 bp (21,341, 15.06%) was the most abundant followed by 200−300 bp (21,211, 14.97%), 400−500 bp (14,348, 10.13%), over 2,000 bp (20349, 14.36%). While the unigene with a length of 1,800−1,900 bp accounted for the smallest proportion (**Supplementary Table 5**). To better know the gene function, all unigenes were compared to seven functional database; 141,672 unigenes were annotated. Among them, NR database and Trembl database annotated the most unigenes, and the ratio of annotation was 61.18% (86,670) and 61.12% (86,585) respectively. The number of unigenes compared to GO database was 71,292 (50.32%) after the NR database and the Trembl database. In addition, the number of annotations in the KEGG database, SwissProt database, and Pfam database was 62,812 (44.34%), 62,732 (44.28%), and 62,036 (43.79%), respectively (**Supplementary Table 6**). Besides, the ratio successfully annotated to the KOG database was the least (52,140, 36.8%). 70.85% of unigenes of *L. indica* shared the highest similarity (70.85%) with *Punica granatum* based on the NR database (**Supplementary Figure 3**). The results of GO annotation showed that all unigenes were successfully categorized into three major categories, including biological process, cellular component, and molecular function. And GO functions involved 60 subcategories. Among them, most of unigenes were enriched in the subcategories cell (53,405, 74.91%), cell part (53,282, 74.74%), cellular process (45,240, 63.46%), binding (42,942, 60.23%), organelle (40,749, 57.16%), and metabolic process (38,792, 54.41%; **Supplementary Figure 4**). In addition, the KOG database was used for orthologous protein annotation. As a result, 25 KOG functional categories were annotated, general function prediction only enriched the majority of genes, and the number of it is 9,948. The number of genes annotated in posttranslational modification, protein turnover, chaperones followed subsequently (**Supplementary Figure 5**).

Differentially expressed genes between different colored flowers

By analyzing transcripts obtained from the transcriptome, DEGs in different colored samples were identified based on their FPKM values. There were 17,131 (WH vs. VT), 14,940 (WH vs. RD), 21,570 (WH vs. PP), 19,223 (WH vs. PK), 19,011 (VT vs. RD), 23,096 (VT vs. PP), 16,348 (VT vs. PK), 17,038 (RD vs. PP), 22,691 (RD vs. PK), and 25,028 (RD vs. PP) DEGs in 10 comparison groups, respectively (**Figure 3A**). To know the gene function of DEGs, gene functional annotation was performed. From GO annotation of all DEGs. A total of 33,534 DEGs were distributed into 58 terms, including 13 molecular functions, 16 cellular components, 28 biological processes, and 55.31% in biological processes. The greatest abundance terms were cellular process, metabolic process, response to stimulus, and biological regulation. Under cellular component, cell, cell part, organelle, and membrane were the most abundant terms. Binding and catalytic activity contained the most DEGs within the biological process (**Figure 3B**).

To further investigate the metabolic pathway of the DEGs, KEGG pathway enrichment of DEGs between five cultivars, including WH vs. VT, WH vs. RD, WH vs. PP, WH vs. PK, VT vs. RD, VT vs. PP, VT vs. PK, RD vs. PP, RD vs. PK, and PP vs. PK, respectively. The top 20 enriched pathways were shown in all comparisons (**Figure 3C** and **Supplementary Figure 6**). From the results, the pathway of biosynthesis of secondary metabolites, metabolic pathways, phenylalanine metabolism, and flavonoid biosynthesis, which were connected to anthocyanin synthesis, were enriched in the VT, PP, RD, and PK.

The differential expressed structural genes regulating the flavonoid biological synthesis

By analyzing KEGG relationships and gene annotations, DEGs related to the biological synthesis of flavonoids were identified. As a result, 53 DEGs on the flavonoid pathway were discovered, including eight *CHS*, four *CHI*, one *F3H*, three *F3′5′H*, four *F3′ H*, and one *DFR*, one *ANS*, *three OMT*, nine *UGT*, six *FLS*, three *FNS*, and 10 *ANR*. All these 53 DEGs were identified in 10 unrepetitive comparison pairs of PK, RD, PP, VT, and WH. From the heatmap (**Figure 4**), *CHS* and *CHI* gene expression of PP was relatively lower than others. For the *F3H* gene, the expression of VT was higher than PK and RD, while the upstream genes (*CHS* and *CHI*) of the PK and RD had higher expression. *FNS* genes of PK and RD had higher expression. Enzyme FNS and F3H catalyzed the same substrate naringenin, which indicated that when naringenin was transformed into Dihydrokaempferol (DHK) and flavones, more substrate flowed to the synthesis of flavones. At the same time, the *FNS* gene of the WH sample was also highly expressed, which indicated that the flavonoid accumulation of the WH sample was higher than that of the other samples.

When it comes to the expression of the *F3′5′H* and *F3′H* genes in five cultivars, there were some interesting findings. The expression of the *F3′5′H* gene in VT was much higher than in other samples, but the expression of the *F3′H* gene was relatively low. *F3′5′H* gene and *F3′H* gene co-catalyzed the substrate DHK, indicating that there was competition for substrate allocation in VT, and ultimately more substrate flows to the synthesis of dihydromyricetin (DHM). This tributary led

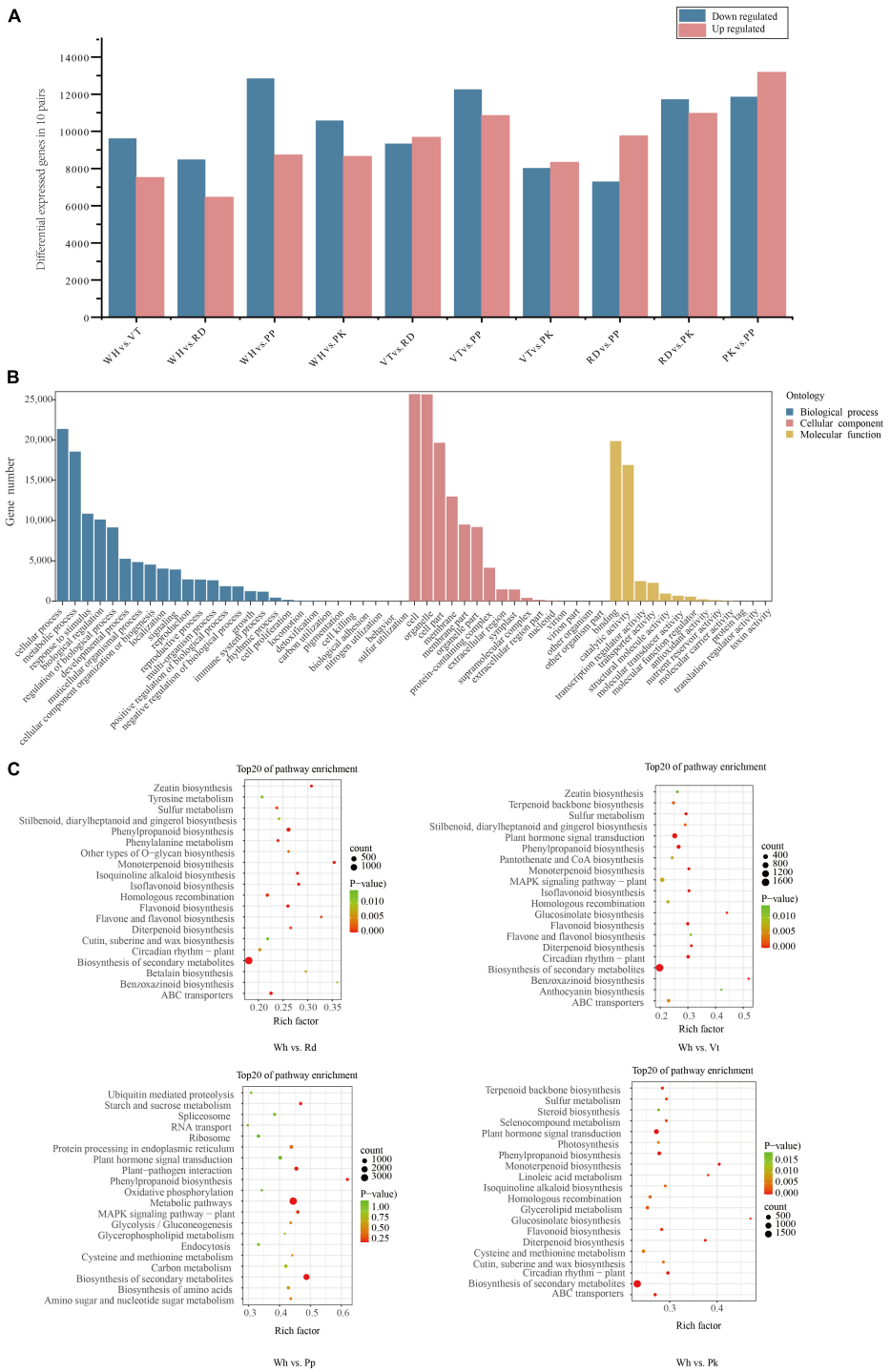


FIGURE 3
Statistical analysis of DEGs, GO, and KEGG enrichment analysis of all DEGs. **(A)** Number of DEGs among all comparison units. **(B)** Go enrichment of all DEGs. **(C)** KEGG enrichment analysis among WH vs. RD, WH Vs. PP, WH vs. VT, and WH vs. PK. The top 20 enriched pathways are shown.

to the synthesis of delphinidin derivatives, malvidin derivatives, and peonidin derivatives. From the metabolome analysis of VT, VT contained four such compounds, accounting for 57.14% of all anthocyanin types. The results of gene expression in VT can also be consistent with the metabolome analysis. The expression levels of the *F3′H* gene in RD were higher than in others, which indicated that the substrate DHK flowed more toward the synthesis of dihydroquercetin (DHQ) than

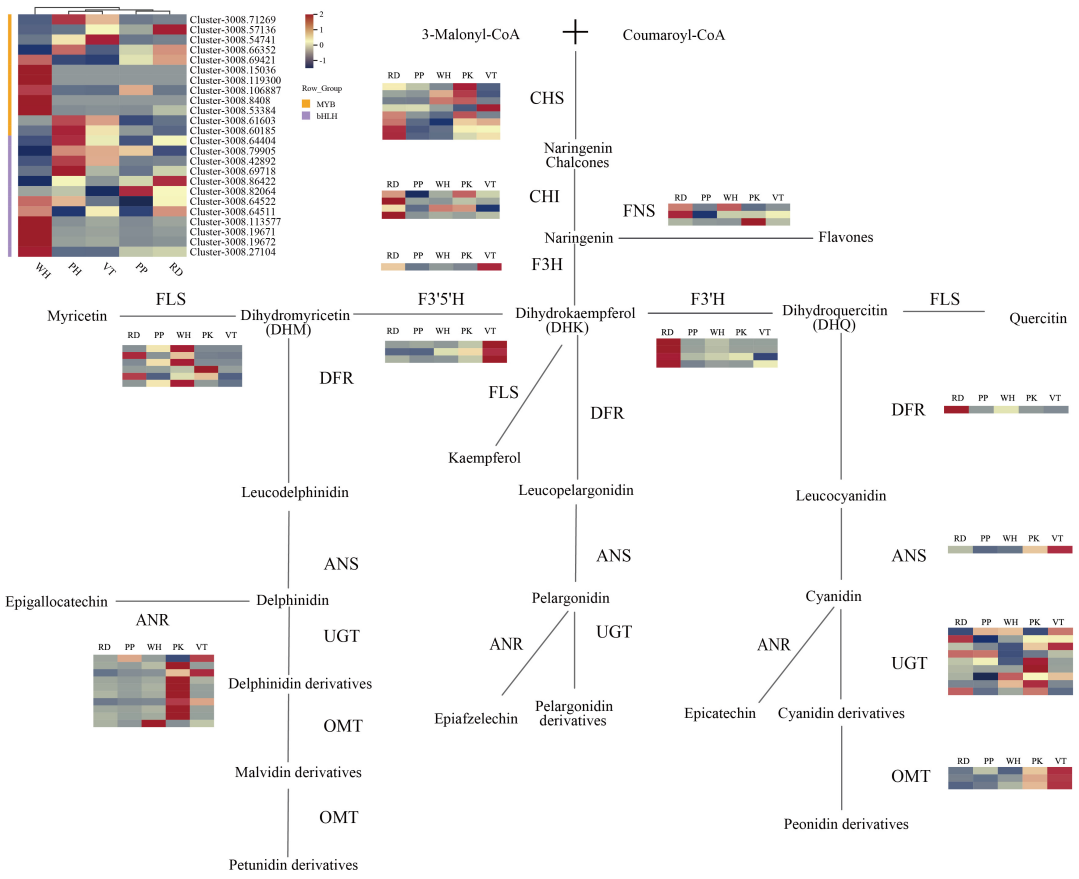


FIGURE 4
Anthocyanin biosynthesis pathway in *L. indica* flowers. Each colored cell represents the average FPKM value standardized by the Z-Score of each gene.

toward the cyanidin and peonidin derivatives. The metabolome's findings corroborated these findings. Although the expression of pathway genes was low in the PP, it indicated that pathway genes of PP may be more efficient when coding relevant protein. However, the flower color of PP was darker than PK and VT, and the total content of anthocyanin was also high. In contrast to the RD of the same dark-colored flowers, the expression level of the gene was low. It is, therefore, possible that the active period for anthocyanin accumulation in PP may not be during the full bloom period.

The expression levels of *FNS* genes and *FLS* genes were higher in white samples, while the expression levels of structural genes in the anthocyanin synthesis pathway were lower. It indicated that the final product of the catalytic substrate did not ultimately flow toward the synthesis of anthocyanins but more toward the synthesis of flavanones. *DFR* and *ANS* were downstream genes of anthocyanin synthesis. The expression level of *DFR* in PK was low, and *FLS* expressed highly in PK, which meant more DHQ flowed to the flavone synthesis. *OMT* genes and *UGT* genes were responsible for the final modification of anthocyanidins. It could be seen from the heatmap that all colored samples had high expression levels of *UGT*. The expression of *OMT* genes in VT samples was higher than in other samples, followed by PK. *OMT* genes could be involved in the methylation of anthocyanins. There were three kinds of anthocyanins in PK and VT that need to undergo multiple methylations including malvidin-3,5-di-*O*-glucoside, peonidin-3-*O*-glucoside, and petunidin-3,5-di-*O*-glucoside. The expression of *OMT* genes was lower in PP and RD than in PK and VT, but the seven anthocyanins in PP and RD were required to undergo multiple methylations.

Identification of candidate transcription factors and correlation network of key differentially expressed genes and differential accumulated anthocyanins

In addition to pathway genes, TFs also played an important role in regulating the synthesis of plant anthocyanins. There

were three main transcription factors involved in the regulation of anthocyanin synthesis including MYB, bHLH, and WD40. A total of 5317 TFs were annotated in this study, of which 184 (3.46%) were annotated for MYB, 257 (4.83%) for bHLH, and WD40 was not annotated. Differential expression analysis of the annotated TFs was performed, and finally, 11 MYBs and 5 bHLHs that might be involved in the regulation of anthocyanin synthesis were screened. To better know the relationship between anthocyanins, DEGs, and candidate TFs, a correlation network map was constructed by 14 anthocyanins, 34 structural genes, and 24 TFs. It could be seen from **Figure 5** that there were 22 genes directly related to anthocyanins, including 11 TFs and 10 structural genes. 11 TFs included nine MYBs and two bHLHs, 10 pathway genes included one *CHS* gene, two *CHI* genes, one *F3′H* gene, one *F3′5′H* gene, one *OMT* gene, and four *UGT* genes.

As a result, only 12 anthocyanins were strongly correlated with candidate structural genes and TFs, and more genes were correlated with delphinidin-3-*O*-galactoside, delphinidin-3-*O*-sambubioside, malvidin-3,5-di-*O*-glucoside, peonidin-3-*O*-rutinoside, petunidin-3,5-di-*O*-glucoside. Malvidin-3,5-di-*O*-glucoside, delphinidin-3-*O*-galactoside, delphinidin-3-*O*-sambubioside, and petunidin-3,5-di-*O*-glucoside are derivatives of malvidin, delphinidin, and petunidin, which all belonged to the final products of the lower tributaries catalyzed by *F3′5′H* enzyme. Among malvidin-3,5-di-*O*-glucoside related genes, three *MYB* genes were negatively correlated with malvidin-3,5-di-*O*-glucoside, and four pathway genes were positively correlated with malvidin-3,5-di-*O*-glucoside (one *F3′5′H* gene and three *OMT* genes). The correlation of transcription factor MYB was generally higher than that of pathway genes, and delphinidin-3-*O*-sambubioside, delphinidin-3-*O*-galactoside, and Petunidin-3,5-di-*O*-glucoside were consistent with that of malvidin-3,5-di-*O*-glucoside. In petunidin-3,5-di-*O*-glucoside and petunidin-3-*O*-glucoside-5-*O*-arabinoside, three identical MYB TFs were negatively correlated with them, and three pathway genes were positively correlated with them. The three structural genes included two *UGT* genes and one *F3′5′H* gene.

Six of the seven genes with a high correlation with peonidin-3-*O*-rutinoside were TFs and only one pathway gene. In addition, only one TF was positively correlated with the seven genes, and the value of | PCC| was higher than the other six genes. Four genes related to cyanidin-3,5-di-*O*-glucoside contained two TFs and two pathway genes, all those two TFs had a positive correlation with cyanidin-3,5-di-*O*-glucoside. One of the two pathway genes is positively correlated and the other is negatively correlated. All three genes related to peonidin-3-*O*-glucoside were positively correlated, including two TFs and one pathway gene. All genes associated with pelargonidin-3-*O*-glucoside and pelargonidin-3,5-*O*-diglucoside were positively correlated. Four genes (two transcription factors and two pathway genes) were related to pelargonidin-3-*O*-glucoside, and two structural genes were related to pelargonidin-3,5-*o*-diglucoside.

Among all TFs highly correlated with anthocyanins, most were negatively correlated, including seven MYBs and one bHLH. Only three TFs were positively correlated, including two MYBs and one bHLH. Most of the structural genes were positively correlated with anthocyanins, and only two were negatively correlated, namely, one *UGT* and one *OMT*. Structural genes are responsible for anthocyanin skeleton synthesis f transcription factors are responsible for the regulation of structural genes. Therefore, it is reasonable that most structural genes are positively correlated while most transcription factors are negatively correlated.

The correlation coefficients of the *F3′5′H* gene (Cluster-3008.85637) and the *F3′H* gene (Cluster-3008.63603) were significantly higher than upstream genes *CHS* (Cluster-3008.59967) and *CHI* (Cluster, 3008.65108-3008.45885) in anthocyanin synthesis pathway. It can be concluded that the *F3′5′H* gene and *F3′H* gene play a more important role in anthocyanin synthesis.

Discussion

Malvidin-3,5-di-*O*-glucoside may be responsible for the color diversity in *Lagerstroemia indica* flowers

Anthocyanins take the main responsibility for the coloration of petals of plants (Zhuang et al., 2019). Like the flowers of other species, anthocyanin was also enriched in the petals of *L. indica*. There were 14 different kinds of anthocyanin identified in five cultivars, and the level of accumulation of anthocyanin was higher in colored petals than in white petals. In addition, the color of the flower is mainly decided by the type and content of the pigment in the plant tissue. In this study, five types only existed in specific-colored samples, these pigments might play an important role in establishing the color. And seven DAAs were identified among WH vs. RD, WH vs. PP, WH vs. PK, and WH vs. VT, which might be the contribution to the different petal colors. Analyzing the relative content of these seven anthocyanins in different samples, the relative content ratio of malvidin-3,5-di-*O*-glucoside varied greatly in different samples. In addition, from the correlation network, malvidin-3,5-di-*O*-glucoside had a high correlation with candidate genes involved in the synthesis of flavonoids. This finding could support the results of Zhang et al. (2008), who pointed out that malvidin-3-*O*-glucoside is one of the main pigments in crape myrtle flowers. Malvidin was mainly connected with the purple flowers. Malvidin-3,5-di-*O*-glucoside is detected in the flower of *Salvia miltiorrhiza* and *Glycine soja* and was responsible for the formation of the purple flowers (Takahashi et al., 2010; Jiang et al., 2020). In this study, malvidin-3,5-di-*O*-glucoside

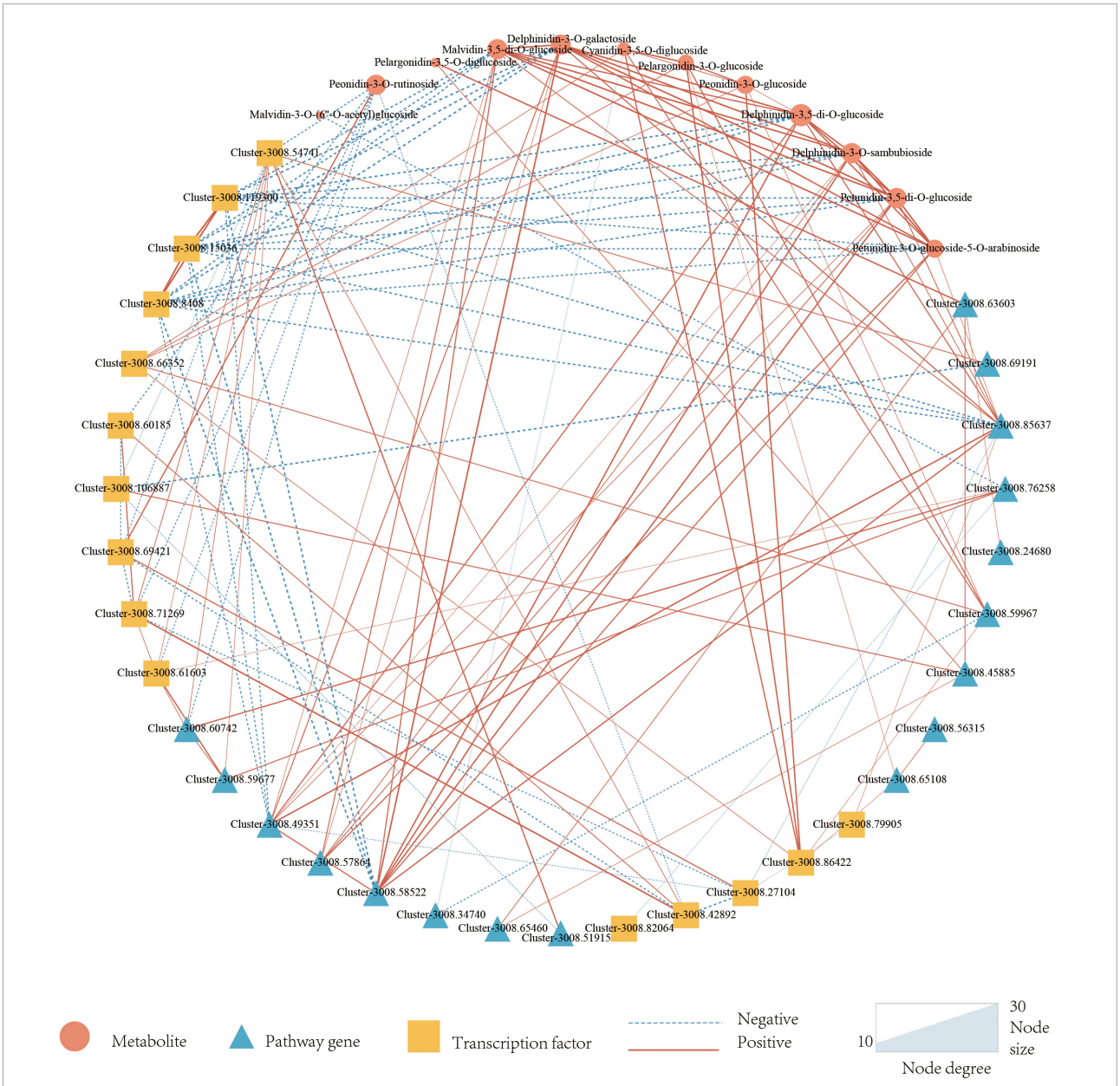


FIGURE 5
Correlation network of candidate metabolites, Pathway genes, and Transcription factors. Triangles represent metabolites, circles represent pathway genes, squares represent transcription factors, dotted lines represent negative associations, and solid lines represent positive associations. Metabolites are set as target nodes, which are determined by the number of genes associated with them.

was presented in all colored flowers and was highly accumulated in PP, and the level of accumulation in VT is higher than in PK. While RD and PK also contained malvidin-3,5-di-*O*-glucoside, it indicated that the final formation of red color and pink color in RD and PK needs the participation of other kinds of anthocyanins. In the study of Zhang et al. (2008), the content of delphinidin-3-*O*-glucoside was mentioned to be one of the key pigments for breeding the blue-colored crape myrtle flower. In previous studies, delphinidin derivatives were considered the key pigment of most blue flowers (Qi et al., 2013). In this study, four delphinidin derivatives (delphinin-3-*O*-galactoside, delphinidin-3-*O*-sambubioside, delphinidin-3,5-*O*-diglucoside, and delphinidin-3-*O*-rutinoside-7-*O*-glucoside) were found. Malvidin derivatives were the downstream production of delphinidin derivatives, delphinidin derivatives could be transformed into malvidin derivatives through the modification by the gene *OMT*. It indicated that there was a competition between the synthesis of delphinidin and malvidin that might hinder the accumulation of delphinidins for blue flowers. While the true content of every pigment in each sample

would not be known through relative content detection. Further study is needed to know the main pigment of each sample and analyze the absolute quantification analysis of each colored sample.

The expression of *F3′5′H* and *F3′H* may have a great influence on color formation

The biosynthesis pathway of anthocyanin was reported in Holton and Cornish (1995), and the function of key genes in the anthocyanin pathway was explored in recent studies. As reported, the expression of *F3′5′H* and *F3′H* is the key that can control the carbon flux to different branches in the pathway of anthocyanin synthesis (Shimada et al., 2001; Jeong et al., 2006).

The pigment synthesis of red or purple/blue flowers was controlled by *F3′H* and *F3′5′H*, respectively. Malvidin was replaced by peonidin as the primary anthocyanin in petunia transgenics overexpressing grape *VvF3′5′H* in the petunia mutant *ht1*, and the flower color turned to purple (Bogs et al., 2005). The fruit skin of *Lycium barbarum* appeared red because of the lack of expression of *F3′5′H* (Zeng et al., 2014). Suppressing the expression of *F3′5′H* led to the decrease of delphinidin derivatives and resulted in the red flower in the gentian (Nakatsuka et al., 2008). Therefore, the different transcriptional levels of *F3′H* and *F3′5′H* would decide the type and ratio of final pigments in flowers and ultimately form different flower colors. In this study, the levels of *F3′H* and *F3′5′H* expressions in five colored samples (white, pink, red, violet, and purple) were different. RD (red flowers) had a high level of *F3′H* expression, and *F3′5′H* was expressed highly in VT (violet flowers). And this would result in the color diversity of *L. indica*.

In addition, *F3′5′H* and *F3′H* as the center of the anthocyanin synthesis branch had different preferences and catalytic activity for the substrate (Olsen et al., 2010; Wang et al., 2014; Miyahara et al., 2016). And *F3′5′H* and *F3′H* would compete for common substrates (DHK), so the different preferences of *F3′H* and *F3′5′H* for substrates could ultimately affect the direction of substrate flow to a certain extent and influence the diversity of pigments in plants and lead to the different color formation. Transforming *F3′5′H* of prairie gentian to pink petunia, enzyme *F3′H* competed with the DHK with enzyme *F3′5′H*, and the color of the flowers turned to purplish red. In this experiment, we noticed *F3′H* had expression in purplish flowers and *F3′5′H* was expressed also in reddish flowers. It meant *F3′H* and *F3′5′H* would have a high possibility of competing for the same substrate in different samples, which could affect the ratio of substrates assigned to the two tributaries, thus affecting the formation of the final color of flowers. Besides, *F3′5′H* would control the synthesis of the key anthocyanin malvidin-3,5-di-*O*-glucoside. Thus, we considered *F3′H* and *F3′5′H* as the key to decoding the diverse color in *L. indica*.

For TFs shown in the correlation network diagram (**Figure 5**), 11 transcription factors were shown to be related to the synthesis of anthocyanins, among which only one of the three MYBs was negatively correlated with the content of malvidin-3,5-di-*O*-glucoside. Some studies have reported that the MYB transcription factor in *L. indica* had a higher preference for the regulation of anthocyanins (Yu et al., 2021). However, how MYB regulated the expression of structural genes and thus influenced the synthesis of anthocyanins is still unclear. In future research, we will focus on the study of the expression of the *F3′5′H* gene and *F3′H* gene in the crape myrtle, as well as the condition of interactions between MYB transcription factors and structural genes in crape myrtle.

Taken together, our study showed that malvidin-3,5-di-*O*-glucoside might be the key pigments for the color diversity of crape myrtle. *F3′H* and *F3′5H* are two key genes in the anthocyanin synthesis pathway, that might have a great influence on the synthesis of anthocyanins and the final color formation of flowers. 11 TFs which might have a great influence on the key pigment synthesis were identified, including nine MYBs and two bHLH.

Data availability statement

The original contributions presented in this study are publicly available. This data can be found here: NCBI, BioProject PRJNA818829 and BioSample SAMN26884420.

Author contributions

SH contributed to the conceptualization, performed the methodology, investigated and validated the data, wrote the original draft, and visualized the data. JW contributed to the conceptualization, performed the methodology, validated and visualized the data, and wrote the original draft. QW contributed to the conceptualization, performed the methodology, investigated and visualized the data, and wrote the original draft. YZ and GZ performed the methodology and validated the data. QM performed the methodology and investigated the data. ZW contributed to the conceptualization, performed the methodology, and wrote, reviewed, and edited the manuscript. JM contributed to the conceptualization, performed the methodology, and wrote, reviewed, and edited the manuscript. CG contributed to the conceptualization, performed the methodology, carried out the resources, supervised the data, and carried out the funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by grants from the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY21C160001) and the Zhejiang Provincial of China Agricultural New Varieties Breeding Major Science and Technology Projects (Grant No. 2021C02071-4).

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.970023/full#supplementary-material

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bairoch, A., and Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* 19(Suppl.), 2247–2249. doi: 10.1093/nar/19.suppl.2247

Bogs, J., Ebadi, A., McDavid, D., and Robinson, S. P. (2005). Identification of the flavonoid hydroxylases from grapevine and their regulation during fruit development. *Plant Physiol.* 140, 279–291. doi: 10.1104/pp.105.073262

Borevitz, J. O., Xia, Y., Blount, J., Dixon, R. A., and Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* 12, 2383–2393. doi: 10.1105/tpc.12.12.2383

Brueggemann, J., Weisshaar, B., and Sagasser, M. (2010). A WD40-repeat gene from *Malus* × *domestica* is a functional homologue of *Arabidopsis thaliana* TRANSPARENT TESTA GLABRA1. *Plant Cell Rep.* 29, 285–294. doi: 10.1007/s00299-010-0821-0

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cabrera, R. I. (2004). Evaluating and promoting the cosmopolitan and multipurpose lagertroemia. *Acta horticult.* 630, 177–184.

Cain, C. C., Saslowsky, D. E., Walker, R. A., and Shirley, B. W. (1997). Expression of chalcone synthase and chalcone isomerase proteins in Arabidopsis seedlings. *Plant Mol. Biol.* 35, 377–381. doi: 10.1023/A:1005846620791

Chagné, D., Lin Wang, K., Espley, R. V., Volz, R. K., How, N. M., Rouse, S., et al. (2012). An ancient duplication of apple MYB transcription factors is responsible for novel red fruit-flesh phenotypes. *Plant Physiol.* 161, 225–239. doi: 10.1104/pp.112.206771

Chen, Y., Zhang, R., Song, Y., He, J., Sun, J., Bai, J., et al. (2009). RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: Finding potential biomarkers for breast cancer. *Analyst* 134, 2003–2011. doi: 10.1039/b907243h

Dao, T. T. H., Linthorst, H. J. M., and Verpoorte, R. (2011). Chalcone synthase and its functions in plant resistance. *Phytochem. Rev.* 10, 397–412. doi: 10.1007/s11101-011-9211-7

Deng, C., Li, S., Feng, C., Hong, Y., Huang, H., Wang, J., et al. (2019). Metabolite and gene expression analysis reveal the molecular mechanism for petal colour variation in six *Centaurea cyanus* cultivars. *Plant Physiol. Biochem.* 142, 22–33. doi: 10.1016/j.plaphy.2019.06.018

Espley, R. V., Hellens, R. P., Putterill, J., Stevenson, D. E., Kutty-Amma, S., and Allan, A. C. (2007). Red colouration in apple fruit is due to the activity of the MYB transcription factor, *MdMYB10*. *Plant J.* 49, 414–427. doi: 10.1111/j.1365-313X.2006.02964.x

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2013). Pfam: The protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Fu, M., Yang, X., Zheng, J., Wang, L., Yang, X., Tu, Y., et al. (2021). Unraveling the regulatory mechanism of color diversity in *Camellia japonica* petals by integrative transcriptome and metabolome analysis. *Front. Plant Sci.* 12:685136. doi: 10.3389/fpls.2021.685136

Gonzalez, A., Zhao, M., Leavitt, J. M., and Lloyd, A. M. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* 53, 814–827. doi: 10.1111/j.1365-313X.2007.03373.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

He, H., Ke, H., Keting, H., Qiaoyan, X., and Silan, D. (2013). Flower colour modification of chrysanthemum by suppression of *F3′H* and overexpression of the exogenous *Senecio cruentus F3′5′H* gene. *PLoS One* 8:e74395. doi: 10.1371/journal.pone.0074395

Hichri, I., Barrieu, F., Bogs, J., Kappel, C., Delrot, S., and Lauvergeat, V. (2011). Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *J. Exp. Bot.* 62, 2465–2483. doi: 10.1093/jxb/erq442

Holton, T. A., and Cornish, E. C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* 7, 1071–1083. doi: 10.1105/tpc.7.7.1071

Irmisch, S., Ruebsam, H., Jancsik, S., Man Saint Yuen, M., Madilao, L. L., and Bohlmann, J. (2019). Flavonol biosynthesis genes and their use in engineering the plant antidiabetic metabolite montbretin A. *Plant Physiol.* 180, 1277–1290. doi: 10.1104/pp.19.00254

Jeong, S. T., Goto-Yamamoto, N., Hashizume, K., and Esaka, M. (2006). Expression of the flavonoid 3′-hydroxylase and flavonoid 3′,5′-hydroxylase genes and flavonoid composition in grape (*Vitis vinifera*). *Plant Sci.* 170, 61–69. doi: 10.1016/j.plantsci.2005.07.025

Jiang, T., Zhang, M., Wen, C., Xie, X., Tian, W., Wen, S., et al. (2020). Integrated metabolomic and transcriptomic analysis of the anthocyanin regulatory networks in *Salvia miltiorrhiza* Bge. flowers. *BMC Plant Biol.* 20:349. doi: 10.1186/s12870-020-02553-7

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2013). PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42, D1182–D1187. doi: 10.1093/nar/gkt1016

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36(Suppl._1), D480–D484. doi: 10.1093/nar/gkm882

Koes, R., Verweij, W., and Quattrocchio, F. (2005). Flavonoids: A colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci*. 10, 236–242. doi: 10.1016/j.tplants.2005.03.002

Landi, M., Tattini, M., and Gould, K. S. (2015). Multiple functional roles of anthocyanins in plant-environment interactions. *Environ. Exp. Bot*. 119, 4–17. doi: 10.1016/j.envexpbot.2015.05.012

Li, B., and Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform*. 12:323. doi: 10.1186/1471-2105-12-323

Liang, J., and He, J. (2018). Protective role of anthocyanins in plants under low nitrogen stress. *Biochem. Biophys. Res. Commun.* 498, 946–953. doi: 10.1016/j.bbrc.2018.03.087

Liu, Y., Zetter, R., Ferguson, D., and Zou, C. (2008). Lagerstroemia (Lythraceae) pollen from the Miocene of eastern China. *Grana* 47, 262–271. doi: 10.1080/00173130802457255

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lukačin, R., Urbanke, C., Gröning, I., and Matern, U. (2000). The monomeric polypeptide comprises the functional flavanone 3β-hydroxylase from *Petunia hybrida*. *FEBS Lett*. 467, 353–358. doi: 10.1016/S0014-5793(00)01116-9

Miyahara, T., Hamada, A., Okamoto, M., Hirose, Y., Sakaguchi, K., Hatano, S., et al. (2016). Identification of flavonoid 3′-hydroxylase in the yellow flower of *Delphinium zalil*. *J. Plant Physiol*. 202, 92–96. doi: 10.1016/j.jplph.2016.07.013

Moreau, C., Ambrose, M. J., Turner, L., Hill, L., Ellis, T. H. N., and Hofer, J. M. I. (2012). The *b* gene of pea encodes a defective flavonoid 3′,5′-hydroxylase, and confers pink flower color. *Plant Physiol*. 159, 759–768. doi: 10.1104/pp.112.197517

Nakatsuka, T., Mishiba, K.-I., Abe, Y., Kubota, A., Kakizaki, Y., Yamamura, S., et al. (2008). Flower color modification of gentian plants by RNAi-mediated gene silencing. *Plant Biotechnol*. 25, 61–68. doi: 10.5511/plantbiotechnology.25.61

Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (2000). The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in Arabidopsis siliques. *Plant Cell* 12, 1863–1878. doi: 10.1105/tpc.12.10.1863

Olsen, K. M., Hehn, A., Jugdé, H., Slimestad, R., Larbat, R., Bourgaud, F., et al. (2010). Identification and characterisation of CYP75A31, a new flavonoid 3′5′-hydroxylase, isolated from *Solanum lycopersicum*. *BMC Plant Biol*. 10:21. doi: 10.1186/1471-2229-10-21

Payyavula, R. S., Singh, R. K., and Navarre, D. A. (2013). Transcription factors, sucrose, and sucrose metabolic genes interact to regulate potato phenylpropanoid metabolism. *J. Exp. Bot*. 64, 5115–5131. doi: 10.1093/jxb/ert303

Pounders, C., Rinehart, T., Edwards, N., and Knight, P. (2007). An analysis of combining ability for height, leaf out, bloom date, and flower color for crapemyrtle. *Hortscience* 42, 4. doi: 10.21273/HORTSCI.42.6.1496

Qi, Y., Lou, Q., Quan, Y., Liu, Y., and Wang, Y. (2013). Flower-specific expression of the Phalaenopsis flavonoid 3′, 5′-hydoxylase modifies flower color pigmentation in *Petunia* and *Lilium*. *Plant Cell Tissue Organ Cult*. 115, 263–273. doi: 10.1007/s11240-013-0359-2

Qiao, Z., Liu, S., Zeng, H., Li, Y., Wang, X., Chen, Y., et al. (2019). Exploring the molecular mechanism underlying the stable purple-red leaf phenotype in *Lagerstroemia indica* cv. ebony embers. *Int. J. Mol. Sci*. 20:5636. doi: 10.3390/ijms20225636

Rahim, M. A., Busatto, N., and Trainotti, L. (2014). Regulation of anthocyanin biosynthesis in peach fruits. *Planta* 240, 913–929. doi: 10.1007/s00425-014-2078-2

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Shimada, Y., Ohbayashi, M., Nakano-Shimada, R., Okinaka, Y., Kiyokawa, S., and Kikuchi, Y. (2001). Genetic engineering of the anthocyanin biosynthetic pathway with flavonoid-3′,5′-hydroxylase: Specific switching of the pathway in petunia. *Plant Cell Rep*. 20, 456–462. doi: 10.1007/s002990100319

Stracke, R., Ishihara, H., Huep, G., Barsch, A., Mehrtens, F., Niehaus, K., et al. (2007). Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J*. 50, 660–677. doi: 10.1111/j.1365-313X.2007.03078.x

Szabolcs, J. (1989). *Plant Carotenoids*. Boston, MA: Springer. doi: 10.1007/978-1-4613-0849-2_3

Takahashi, R., Dubouzet, J. G., Matsumura, H., Yasuda, K., and Iwashina, T. (2010). A new allele of flower color gene *W1* encoding flavonoid 3′5′-hydroxylase is responsible for light purple flowers in wild soybean *Glycine soja*. *BMC Plant Biol*. 10:155. doi: 10.1186/1471-2229-10-155

Tanaka, Y. (2006). Flower colour and cytochromes P450. *Phytochem. Rev*. 5, 283–291. doi: 10.1007/s11101-006-9003-7

Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: Anthocyanins, betalains and carotenoids. *Plant J*. 54, 733–749. doi: 10.1111/j.1365-313X.2008.03447.x

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 28, 33–36. doi: 10.1093/nar/28.1.33

Teng, S., Keurentjes, J., Bentsink, L., Koornneef, M., and Smeekens, S. (2005). Sucrose-specific induction of anthocyanin biosynthesis in *Arabidopsis* requires the MYB75/PAP1 Gene. *Plant Physiol.* 139, 1840–1852. doi: 10.1104/pp.105.066688

Thévenot, E. A., Roux, A., Xu, Y., Ezan, E., and Junot, C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J. Proteome Res*. 14, 3322–3335. doi: 10.1021/acs.jproteome.5b00354

Timoneda, A., Feng, T., Sheehan, H., Walker-Hale, N., Pucker, B., Lopez-Nieves, S., et al. (2019). The evolution of betalain biosynthesis in Caryophyllales. *N. Phytol.* 224, 71–85. doi: 10.1111/nph.15980

Walker, A. R., Lee, E., Bogs, J., McDavid, D. A. J., Thomas, M. R., and Robinson, S. P. (2007). White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J*. 49, 772–785. doi: 10.1111/j.1365-313x.2006.02997.x

Wang, Y.-S., Xu, Y. J., Gao, L. P., Yu, O., Wang, X. Z., He, X. J., et al. (2014). Functional analysis of Flavonoid 3′,5′-hydroxylase from Tea plant (*Camellia sinensis*): Critical role in the accumulation of catechins. *BMC Plant Biol*. 14:347. doi: 10.1186/s12870-014-0347-7

Winkel Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol*. 126, 485–493. doi: 10.1104/pp.126.2.485

Yonekura Sakakibara, K., Higashi, Y., and Nakabayashi, R. (2019). The origin and evolution of plant flavonoid metabolism. *Front. Plant Sci*. 10:943. doi: 10.3389/fpls.2019.00943

Yu, C., Lian, B., Fang, W., Guo, A., Ke, Y., Jiang, Y., et al. (2021). Transcriptome-based analysis reveals that the biosynthesis of anthocyanins is more active than that of flavonols and proanthocyanins in the colorful flowers of *Lagerstroemia indica*. *Biol. Futur*. 72, 473–488. doi: 10.1007/s42977-021-00094-0

Zeng, S., Wu, M., Zou, C., Liu, X., Shen, X., Hayward, A., et al. (2014). Comparative analysis of anthocyanin biosynthesis during fruit development in two *Lycium* species. *Physiol. Plant*. 150, 505–516. doi: 10.1111/ppl.12131

Zhang, J., Wang, L., Gao, J., Shu, Q., Li, C., Yao, J., et al. (2008). Determination of anthocyanins and exploration of relationship between their composition and petal coloration in crape myrtle (*Lagerstroemia* hybrid). *J. Integr. Plant Biol*. 50, 581–588. doi: 10.1111/j.1744-7909.2008.00649.x

Zhao, D., and Tao, J. (2015). Recent advances on the development and regulation of flower color in ornamental plants. *Front. Plant Sci*. 6:261. doi: 10.3389/fpls.2015.00261

Zheng, Y., Jiao, C., Sun, H., Hernan Rosli, G., Marina Pombo, A., Zhang, P., et al. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhuang, H., Lou, Q., Liu, H., Han, H., Wang, Q., Tang, Z., et al. (2019). Differential regulation of anthocyanins in green and ourple turnips revealed by combined de novo transcriptome and metabolome analysis. *Int. J. Mol. Sci*. 20:4387. doi: 10.3390/ijms20184387

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 November 2022
DOI 10.3389/fpls.2022.1006291

OPEN ACCESS

EDITED BY
Wilhelm Boland,
Max Planck Institute for Chemical Ecology, Germany

REVIEWED BY
Shihong Luo,
Shenyang Agricultural University, China
Kean-Jin Lim,
Zhejiang Agriculture and Forestry University, China

*CORRESPONDENCE
Hui Yu
yuhui@scbg.ac.cn

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 29 July 2022
ACCEPTED 01 November 2022
PUBLISHED 15 November 2022

CITATION
Fan S, Jia Y, Wang R, Chen X, Liu W and Yu H (2022) Multi-omics analysis the differences of VOCs terpenoid synthesis pathway in maintaining obligate mutualism between *Ficus hirta* Vahl and its pollinators.
Front. Plant Sci. 13:1006291.
doi: 10.3389/fpls.2022.1006291

Multi-omics analysis the differences of VOCs terpenoid synthesis pathway in maintaining obligate mutualism between *Ficus hirta* Vahl and its pollinators

Songle Fan[1,2,3], Yongxia Jia[1], Rong Wang[4], Xiaoyong Chen[4], Wanzhen Liu[1,2] and Hui Yu[1,2]*

[1]Key Laboratory of Plant Resource Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China, [2]Guangdong Provincial Key Laboratory of Digital Botanical Garden, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China, [3]University of Chinese Academy of Sciences, Beijing, China, [4]School of Ecological and Environmental Sciences, Tiantong National Station for Forest Ecosystem Research, East China Normal University, Shanghai, China

Inroduction: Volatile organic compounds (VOCs) emitted by the receptive syconia of *Ficus* species is a key trait to attract their obligate pollinating fig wasps. *Ficus hirta* Vahl is a dioecious shrub, which is pollinated by a highly specialized symbiotic pollinator in southern China. Terpenoids are the main components of VOCs in *F. hirta* and play ecological roles in pollinator attraction, allelopathy, and plant defense. However, it remains unclear that what molecular mechanism difference in terpenoid synthesis pathways between pre-receptive stage (A-phase) and receptive stage (B-phase) of *F. hirta* syconia.

Methods: Transcriptome, proteome and Gas Chromatography-Mass Spectrometer (GC-MS) were applied here to analyze these difference.

Results and discussion:: Compared to A-phase syconia, the genes (*ACAT2*, *HMGR3*, *GGPS2*, *HDR*, *GPS2*, *TPS2*, *TPS4*, *TPS10-4*, *TPS14*) related to the terpenoid synthesis pathway had higher expression level in receptive syconia (B-phase) according to transcriptome sequencing. Seven differentially expressed transcription factors were screened, namely *bHLH7*, *MYB1R1*, *PRE6*, *AIL1*, *RF2b*, *ANT*, *VRN1*. Specifically, *bHLH7* was only specifically expressed in B-phase. 235 differentially expressed proteins (DEPs) were mainly located in the cytoplasm and chloroplasts. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showed that the DEPs were mainly enriched in the metabolic process. A total of 9 terpenoid synthesis proteins were identified in the proteome. Among them, 4 proteins in methylerythritol phosphate (MEP) pathway were all down-regulated. Results suggested the synthesis of terpenoids precursors in B-phase bracts were mainly accomplished through the mevalonic acid (MVA) pathway in cytoplasm. Correlation analysis between the transcriptome and proteome, we

detected a total of 1082 transcripts/proteins, three of which are related to stress. From the VOCs analysis, the average percent of monoterpenoids emitted by A-phase and B-phase syconia were 8.29% and 37.08%, while those of sesquiterpenes were 88.43% and 55.02% respectively. Monoterpenes (camphene, myrcene, camphor, menthol) were only detected in VOCs of B-phase syconia. To attract pollinators, B-phase syconia of *F. hirta* need more monoterpenoids and less sesquiterpenes. We speculate that transcription factor *bHLH7* may regulate the terpenoid synthesis pathway between A- and B-phase syconia. Our research provided the first global analysis of mechanism differences of terpenoid synthesis pathways between A and B phases in *F. hirta* syconia.

KEYWORDS

Ficus hirta Vahl, proteome, terpenoid, transcriptome, VOCs

Introduction

In the Angiospermae, more than 90% of flowering plants are pollinated by insects (Kearns et al., 1998). To attract pollinators, many flowering plants release volatile organic compounds (VOCs) from their flowers, inflorescences or specific tissues of inflorescences (Dudareva et al., 2006; Hu et al., 2020). Plants take advantage of VOCs to communicate and interact with their surroundings (Rosenkranz et al., 2021). Mutual adaptation of plant VOCs and insects play a vital role in adaptive evolution. The VOCs perception process of insects is a complex process that the VOCs penetrate pore tubules of the sensillum are bound and dissolved by OBPs and CSPs, transported through the sensillum lymph, and reach the sensory dendrite to activate the membrane-bound OR (Brito et al., 2016).

The *Ficus* species (Moraceae, *Ficus*) and their pollinating fig wasps known today as they form the closest mutually beneficial obligate symbiotic system. Pollinating fig wasps pollinate syconia, and the syconia provide breeding sites for them. For the mutualism, the olfactory attraction of VOCs is key link between the receptive stage (B-phase) syconia and pollinators to maintain this system (Hossaert-McKey et al., 2010). Pre-receptive stage (A-phase) begins with the appearance of the syconium buds. When the syconia are developed to be ready for pollination, B-phase begins. B-phase lasts until fig wasps are attracted by the VOCs, enter the syconia and lay eggs. VOCs attract obligate pollinating fig wasps in the B-phase syconia, while other developmental stages may repel pollinating fig wasps to maintain the specificity between pollinators and host plants (Gu et al., 2012). Therefore, the unique volatile substances were only detected in B-phase syconia may be the main signal to attract pollinators. WpumOBP2 was a major odor binding protein of unique volatile decanal, and *F. pumila* var. *pumila* attract pollinating fig wasps by the binding of decanal with WpumOPB2 in the receptive stage (Wang et al., 2021).

The main components of VOCs in plants are terpenes, fatty acid derivatives, amino acid derivatives and phenylpropane/benzene compounds (Rosenkranz and Schnitzler, 2016). Generally speaking, VOCs is species-specific and different *Ficus* species emit considerably different VOCs composition (Hossaert-McKey et al., 2016). VOCs compounds of five *Ficus* species (*F. benguetensis*, *F. septica*, *F. variegata*, *F. erecta*, and *F. virgata*) were different, which consist mainly of terpenoids and benzenoids (Okamoto and Su, 2021). VOCs components emitted by different development stages were different in the same *Ficus* species. For example, there were significant differences in the compounds of VOCs between A- and B-phase syconia of *Ficus pumila* var. pumila (Wang et al., 2021). Long-term co-evolution makes fig wasps have a preference for specific VOCs emitted by B phase syconia (Hossaert-McKey et al., 2016; Proffit et al., 2020).

The main biosynthetic pathways of VOCs in plants include mevalonic acid (MVA) pathway in the cytoplasm, methylerythritol phosphate (MEP) pathway in the chloroplasts, shikimate pathway and lipoxygenase (LOX) pathway (Dudareva et al., 2013; Rosenkranz and Schnitzler, 2016). Terpenoids are the main compounds of VOCs that are emitted by plants and attract pollinating insects (Soler et al., 2011; Hossaert-McKey et al., 2016). The terpenoid precursors are synthesized by MVA pathway and MEP pathway, including isopentenyl pyrophosphate (IPP), dimethylallyl pyrophosphate (DMAPP), geranyl diphosphate (GPP), geranyl pyrophosphate (GGPP) and farnesyl pyrophosphate (FPP). GPP is a precursor of monoterpenoids (Kumar et al., 2018), FPP is a precursor of sesquiterpenoids (Ma et al., 2019), and GGPP is a precursor of diterpenoids (Liu et al., 2017). Then, terpenoids are synthesized under the catalysis of terpenoid synthases (TPSs) (Dudareva et al., 2013). Structural properties of TPS proteins are drivers of reaction mechanisms leading to the formation of multiple products and underling the molecular evolution of terpene

diversity (Tholl, 2006). TPSs form a large family that underlies the diversity of terpenoids (Champagne and Boutry, 2016).

Ficus hirta Vahl is a dioecious shrub or small tree that grows in tropical and subtropical regions (Berg, 2003; Yu and Nason, 2013). *F. hirta* bears syconia asynchronously on individual trees (Yu et al., 2006). Roots of *F. hirta* are rich in active ingredients that can be used as medicine and plant-derived popular food (Yi et al., 2013; Wan et al., 2017; Ye et al., 2020). Like other dioecious *Ficus* species, female trees of *F. hirta* bear female syconia that contain female flowers only and produce seeds. Male trees bear male syconia functionally that contain both male and female flowers. The development of syconia of *F. hirta* were also divided into five phases (A-E phases) (Galil and Eisikowitch, 1968; Yu et al., 2006). VOCs emitted by the stomata bracts of syconia were candidate source to attract pollinators over long distances (Hossaert-McKey et al., 1994; Souza et al., 2015; Hu et al., 2020). VOCs emitted by B-phase of *F. hirta* attract obligate pollinators and maintain the obligate mutualism. Terpenoids play important ecological roles in pollinator attraction, allelopathy and plant defense (Mahmoud and Croteau, 2002; Tholl, 2006; Parvin et al., 2014). However, the molecular mechanism differences in terpenoid synthesis pathways between *F. hirta* A-phase and B-phase syconia remains unclear.

Previous studies transcriptomic data was applied to describe VOCs preliminary (Hu et al., 2020). However, transcriptome data only reflect the expression of genes at the transcriptional level. There are many modifications between genes and proteins, which not only affect the expression abundance of proteins directly, but also affect the composition of VOCs. Therefore, these frequent inconsistencies suggest that complementary proteome and VOCs analyses are needed to further be analyzed, including key genes, proteins and metabolites of terpenoid synthesis pathway in syconia bract tissues of different periods. On the basis of comparing the differences of VOCs between A-phase and B-phase syconia, combined transcriptomic and proteomic analysis provide a deeper understanding of the molecular mechanism of how the host *Ficus* species attract obligate pollinators.

In this study, transcriptome sequencing, proteomics and Gas Chromatography-Mass Spectrometer (GC-MS) experiment were used to analyze the difference of genes, proteins and metabolites between A-phase and B-phase syconia, respectively. This study aimed to screen key genes, proteins and metabolites in the terpenoids synthesis pathway to understand the biological process from A-phase to B-phase of *F. hirta* syconia. What is the cause of the differences in terpenoid synthesis pathways between A- and B-phase. The application of multi-omics for jointly analyzing the differences of the terpenoids synthesis pathways in the bracts between A-phase and B-phase. It is helpful to understand the molecular mechanism of maintaining obligate mutualism between *Ficus* species and their pollinator wasps.

Materials and methods

Studies species and sample collection

F. hirta male syconia were collected at the South China Botanical Garden (SCBG), in Guangdong Province. The area has a subtropical maritime climate, with an annual mean temperature of almost 22°C. 24~36 pre-receptive stage (A-phase) syconia and 24-36 receptive stage (B-phase) syconia were randomly collected from male trees in SCBG, and divided into 3 groups. To avoid pollinator visitation, male syconia were bagged before syconia receptivity. Ostiole bracts were dissected from A-phase and B-phase male syconia without pollinated, and put into Sample Protector for RNA (Takara). RNA sequencing (RNA-seq) was performed on Illumina Hiseq platform.

Transcriptome data analysis

Transcriptome sequencing data assembly, annotation and differential gene expression analysis had been completed in previous papers (Hu et al., 2020). Both Expectation-Maximization (RSEM) and the most common method Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced (FPKM) were applied to estimate gene expression level (Li and Dewey, 2011). DESeq R package (1.10.1) was used to analyze differential gene expression (p-adjusted <0.05 and fold change (FC)>2) (Anders and Huber, 2010). Among the 60,299 unigenes detected in both A- and B-phase of *F. hirta* syconia, 187 (0.31%) differentially expressed genes (DEGs) were found (Hu et al., 2020). Gene Ontology (GO) enrichment of DEGs was analyzed by GOseq R packages. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment of DEGs was analyzed using KOBAS (2.0) software (Xie et al., 2011). iTAK software was used to predict plant transcription factors. According to the annotation information of the database, transcription factors and the genes related to terpenoid synthesis pathways were identified, and the relevant information was extracted and sorted out. The heatmap of the target genes expression level was showed by TBtools (Chen et al., 2020).

Label-free quantitative proteome

Samples of proteome were collected as same as the transcriptome materials. Total protein of tissue samples was extracted by SDT(4%SDS, 10 mM DTT, 100 mM TEAB)-acetone method, and protein precipitate was dissolved with dissolution buffer (8 M Urea, 100 mM TEAB, pH 8.5) (Wu et al., 2014). Draw a standard curve with the absorbance of the standard protein solution and calculate the protein

concentration of the sample. Take 20 μg of protein for 12% sodium dodecyl sulfatepolyacrylamide gel electrophoresis (SDS-PAGE). The protein samples were digested by trypsin and then detected by Liquid Chromatography-Mass Spectrometry (LC/MS).

LC/MS analysis was performed using an ultra-nanoflow high-performance liquid chromatography (EASY-nLCTM 1200 nanoscale UHPLC, Thermo Fisher/LC140) system and a Q ExactiveTM series mass spectrometer. The pre-column was a home-made C18 Nano-Trap column (4.5cm×75μm×3μm), and the peptides analytical column was a home-made analytical column (15cm×150μm×1.9μm). Using a Q Exactive™ series mass spectrometer, with Nanospray Flex™ (ESI) ion source, set the ion spray voltage to 2.1 kV and the ion transfer tube temperature to 320°C. The full scan range of the mass spectrometer was m/z 350-1500. The precursor ions with the ion strength of TOP 40 in the full scan were selected for fragmentation by high-energy collisional fragmentation (HCD) method, and secondary mass spectrometry was performed for detection. The peptide fragmentation normalized collision energy was set as 27%. The threshold intensity was 2.2×10^4, and the dynamic exclusion parameter was 20 s. The raw data of MS detection was named as ".raw".

The resulting spectrum was searched using Proteome Discoverer 2.2 (PD2.2) according to the Gene Ontology (GO), KEGG and Clusters of Orthologous Groups (COG) (Madzharova and Sabino, 2019). To improve the quality of the analysis results, the PD2.2 software filtered the search results: i) Peptide Spectrum Matches (PSMs) with a reliability of more than 99% were trusted PSMs, and ii) proteins containing at least one unique peptide are trusted proteins, iii) keep only credible spectrum peptides and proteins, and iv) estimate false discovery rate (FDR). Proteomics quality control including peptide length distribution, precursor ion tolerance, unique peptide number, protein coverage, and protein molecular weight. Principal component analysis (PCA) and coefficient of variance (CV) were performed by R packages. The protein quantitation (protein abundance) results were statistically analyzed by *T*-test. The relative quantitative value of each protein in the two comparison samples was tested by *T*-test, and the corresponding *P* value was calculated as the significance index, and the default *P* value was ≤0.05 (Hussain et al., 2021). The up-regulated proteins were screened, when Fold Change ≥ 1.2 and *P* value ≤ 0.05. The down-regulated expression proteins were screened, when Fold Change ≤ 0.83 and *P* value ≤ 0.05. Cell-PLoc 2.0 was used to predict the subcellular localization of proteins in bract tissues (Chou and Shen, 2008). K-mean cluster analysis of DEPs was performed *via* the R package.

GO and interpro (IPR) functional annotation were performed using Interproscan software (Jones et al., 2014). Volcano plot analysis, K-mean cluster analysis, GO, IPR and KEGG were performed for differentially expressed proteins (DEPs) (Huang Da et al., 2009). STRING DB software (http://STRING.embl.de/) was applied for protein interactions analysis (Franceschini et al., 2013).

VOCs collection and GC-MS analysis

30~35 pre-receptive (A-phase) male syconia and 20-25 receptive (B-phase) male syconia were randomly collected from male trees in SCBG, divided into 3 groups, and put directly into polyethylene terephthalate bags. Before sampling, 120ng/ul mixture of n-nonane and dodecane was added to the sampling tube as the internal standard. Three biological replicates were performed for each stage. The steps of odour collection include, i) putting a syconia into a Teflon film bag with both ends open, ii) fastening one end of the bag with a thin wire, iii) sitting the syconia for 30 minutes, iv) setting the air flow rate at 300 ml/min, v) pumping air into the bag for 5 min and a cycle was completed (Hossaert-McKey et al., 2016). Collect the air in the current sampling environment as a blank control. Samples were stored in a -20°C freezer.

VOCs emitted by syconia were analysed using Gas Chromatography-Mass Spectrometer (GC–MS) system (GCMS-QP2010PLUS). The column used was HP-5MS quartz capillary column (30m × 250μm × 0.25μm). PTV1 injection port need the injection adopts the split mode, and the split ratio set to 10:1. The column temperature was kept at 40°C for 5 minutes in the begin, the increased to 280°C at 7.5°C per minute, and kept for 8 minutes. The inlet temperature was maintained at 40°C for 2.5 minutes. The column flow was set to 2.0 ml/min, the linear velocity was 51 cm/sec, and the purge flow was 3 ml/min. The pre-column pressure was 112 kPa and the injection volume was 2.0 μl. MS conditions were set as ionization by electron bombardment, scanning range 20-45u, electron energy 70 Ev, transmission line temperature 250°C, ion source temperature 230°C. Compounds were mainly identified by matching mass spectra with the standard spectral library (FFNSC1.3, NIST14S, NIST05s) and by comparing Kovats retention indices with that reported in the NIST chemistry Web Book (http://webbook.nist.gov) and published data. The peak area of each compound was quantified as relative quantities of each component based on the normalization method (Soler et al., 2012).

Data analysis

Each protein has its corresponding transcript, and R packages were used to analyze the correlation between transcriptome and proteome. Correlations were calculated based on Pearson's statistical method. K-mean cluster heatmap analysis *via* the R package. Key genes, proteins and metabolites involved in terpenoid synthesis pathway in transcriptome, proteomics and VOCs were analyzed and identified according to annotation databases, then using

Adobe Illustrator CS6 (AI CS6) to map terpenoid synthesis pathways. The heatmap of the target genes/proteins expression level was showed by TBtools. One-way ANOVA analyzed metabolites data at $P < 0.05$ level and multiple comparisons were performed by *Tukey* test.

Results

Transcriptome analysis of bracts in A- and B-Phase of *F. hirta* syconia

A total of 60,299 unigenes were obtained in both A- and B-Phase syconia, 79 up-regulated genes and 108 down-regulated genes were found to be differently expressed (Table S1). The KEGG pathway and GO categorization of 187 differentially expressed genes (DEGs) were analyzed. The KEGG pathways showed the main functions of up-regulated genes were phenylpropanoid biosynthesis, brassinolide biosynthesis, folate biosynthesis, ubiquinone, other terpene quinone biosynthesis, phenylalanine metabolism and RNA transport (Figure 1A). The down-regulated genes of KEGG enriched pathways were mainly cutin, suberine and wax biosynthesis, flavonoid biosynthesis, inositol phosphate metabolism, pentose and glucuronate interconversions, ribosome biogenesis in eukaryotes, plant hormone signal transduction, starch and sucrose metabolism, protein processing in endoplasmic reticulum (Figure 1B). GO categorization only showed that DEGs were mainly concentrated in molecular functions (e.g. heme binding, tetrapyrrole binding, redox process, peroxidase activity) and biological processes (e.g. peroxidase reaction and oxidative stress response) (Figure 1C).

The secondary metabolic process of VOCs were usually regulated by transcription factors, such as *MYB*, *NAC*, *WRKY*, *bHLH*, etc. Therefore, according to the Plant Transcription Factor Database, we analyzed transcription factors from DEGs. The result shown *bHLH7*, *MYB1R1*, *PRE6*(*bHLH163*), *AIL1*, *RF2b*, *ANT*, *VRN1* were significantly differentially expressed. Among them, *bHLH7* and *RF2b* were up-regulated. *MYB1R1*, *PRE6*(*bHLH163*), *AIL1*, *ANT*, *VRN1* were down-regulated (Figure 1D). Interestingly, bHLH7 was only expressed in B-phase syconia bract.

Proteomics analysis of bracts in A- and B-Phase of *F. hirta* syconia

The quantitative proteomic was performed to compare the expression differences of proteins between bracts of A- and B-phase syconia. After mass spectrometry data retrieval, the peptide and protein were checked for quality control (Figure S1). PC1 represents the score of the experimental group, accounting for 54.48% of the total variation, while PC2 represents repeatability of the experimental group, explaining 18.28% of the total variation (Figure S2A). Principal component analysis (PCA) represented a closer association of biological replicates rather than different phases (Figure S2A). Coefficient of Variance results showed that CuB samples have better repeatability (Figure S2B). A total of 668,534 spectra, with 93,239 matching those of known peptides. Among them, 11,223 specific peptides and 2,729 proteins were identified (Table S2). 1,380 proteins were identified in the GO database, 2,122 proteins were identified in the KEGG database, 1,373 proteins were identified in the COG database, and 1,943 domains were identified in the IPR database (Figure 2A). A total of 235 significantly DEPs were identified (Table S3), of which 57 were up-regulated and 178 were down-regulated (Figure 2B).

To analyze the functions of DEPs between A- and B-phase, 235 DEPs were mapped into KEGG pathways and GO categorization, respectively. The KEGG pathway enrichments showed that the major enrichment pathways of DEPs were oxidative phosphorylation, glutathione metabolism, terpenoid backbone synthesis, phenylephrine, tyrosine and tryptophan biosynthesis (Figure 3A). GO categorization showed that these proteins were mainly concentrated in biological processes such as transmembrane transport and tetrapyrrole binding (Figure 3B). The subcellular localization analysis of the DEPs showed that they were mainly located in the cytoplasm and chloroplast (Figure 3C).

In the constructed network, cluster-21669.46738_ORF1 (Methionyl-trNA synthetase) was the most important protein up-regulation hub associated with Aminoacyl-tRNA biosynthesis pathway. Cluster-21669.43468_orf1 (Ubiquilin) related to protein processing in endoplasmic reticulum pathway was an important protein down-regulation hub. The DEPs interaction networks and scores were shown in Figure 4 and Table S4. Among them, the interaction analysis of terpenoid synthesis pathway-related proteins showed that Cluster-21669.24202_orf1 (4-diphosphocytidyl methylerythritol kinase, CMK) could interact with Cluster-21669.36343_orf1 (MEP).

Transcriptome and proteome association analysis

It is well known that the relationship between mRNA and protein is complex. Therefore, we performed a correlation analysis of the transcriptome and proteome. Pearson Correlation Coefficient analysis (Pearson Correlation = 0.047) showed that transcriptome was weakly correlated with proteome (Figure S3). Transcriptome and proteome were associated with 1082 transcripts/proteins (Table S5). Among them, 3 DEGs corresponding to DEPs, namely Cluster-21669.61808 (aspartic protease), Cluster-21669.47934 (mannose-binding lectin), Cluster-21669.43117 (non-specific lipid transfer protein)

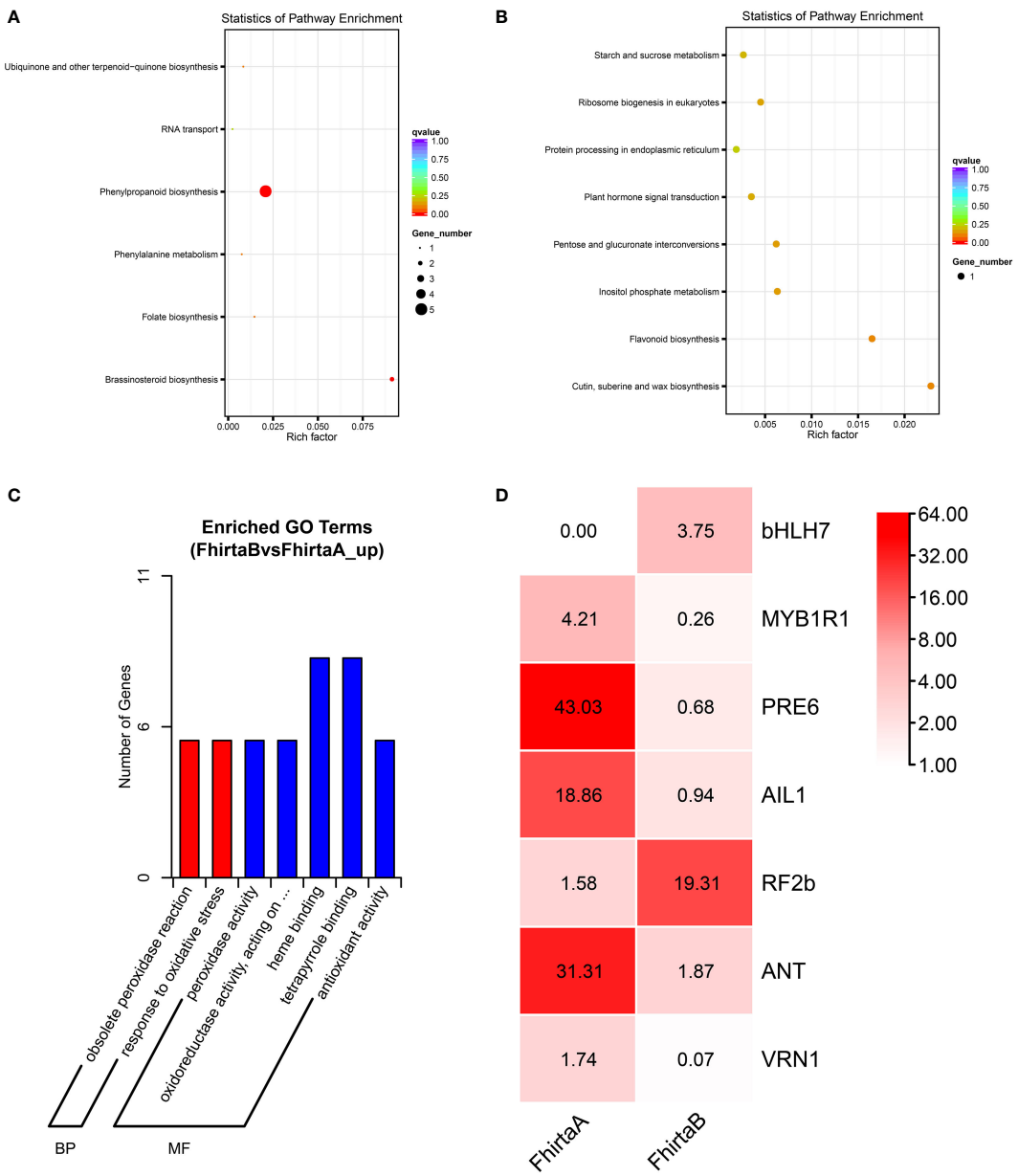


FIGURE 1
Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analysis of differentially expressed genes (DEGs) between A- and B-phase of *F hirta* bracts. **(A)** KEGG classification pathway of up-regulated genes. **(B)** KEGG pathway of down-regulated genes. **(C)** GO classification of up-regulated genes. **(D)** Heatmaps of expression profiles of differentially expressed transcription factors genes according to FPKM. The data inside the box represents FPKM. FhirtaA and FhirtaB represent A phase and B phase of *F hirta* bracts, respectively.

(Figure 5A). Three genes all play an essential role in coping with adversity stress. DEPs of 4 clusters were enriched in 23 KEGG pathways, mainly enriched in the metabolic process. Among them, transcripts/proteins of cluster1 and cluster 3 were both down-regulated (Figure 5B). 4 DEPs were enriched in the terpenoid synthesis pathway. DEPs of 4 clusters were enriched in 43 GO processes (Figure 5C). These results suggested that there are both complex post-transcriptomic and post-translational modifications in the bracts of syconia from A-phase to B-phase.

VOCs emitted by male syconia of *F. hirta*

The VOCs of *F. hirta* in A- and B-phase syconia contained more than 60 compounds including 2 fatty acid derivatives, 20

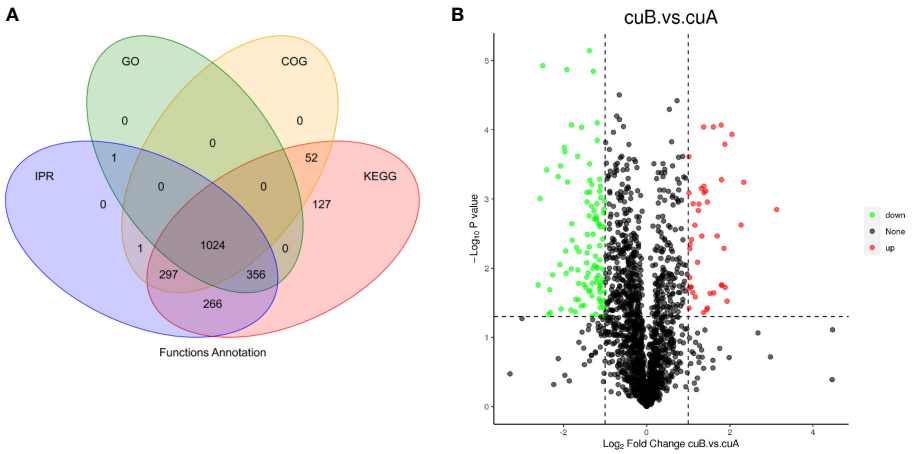


FIGURE 2
Annotation results and expression profiles of differentially expressed (DEPs) proteins. **(A)** Venn diagram of GO, KEGG, Clusters of Orthologous Groups (COG), Interpro (IPR) database annotation results. **(B)** Volcano maps representing expression pattern of DEGs. Red spots represent upregulated DEGs. Green spots indicate downregulated DEGs. Those shown in black are proteins that did not show obvious changes in A phase and B phase of *F hirta* bracts. CuA VS CuB represent multiples of protein expression in both A phase and B phase of *F hirta* bracts.

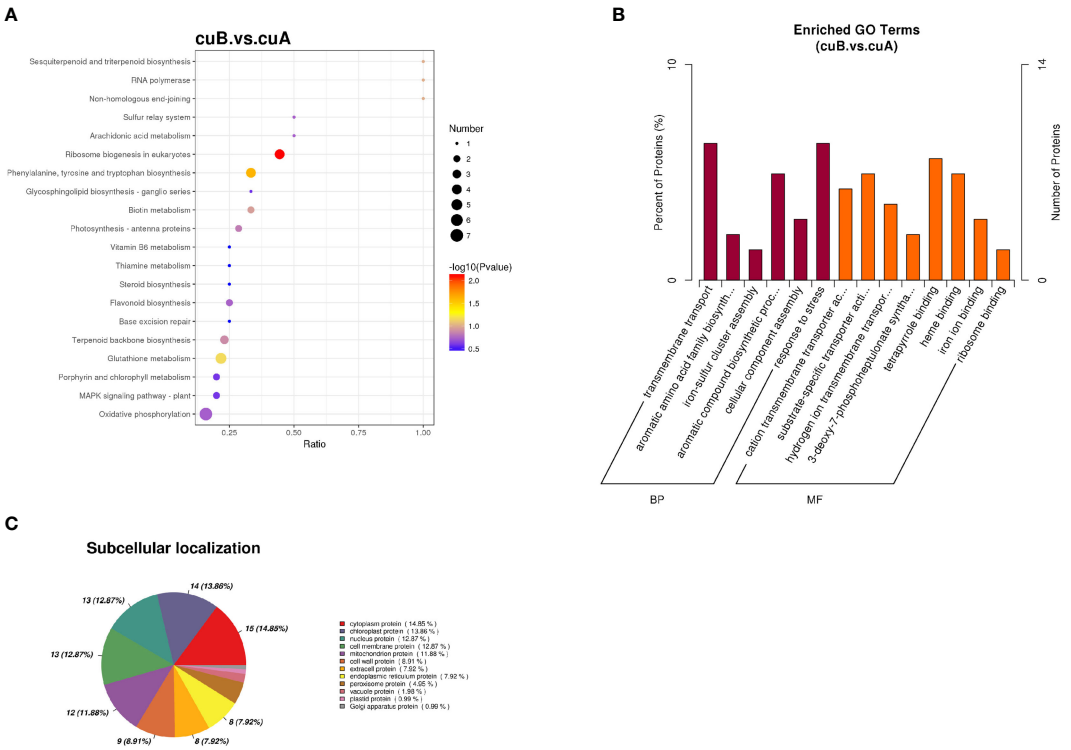


FIGURE 3
Differential expression proteins (DEPs) enrichment and localization analysis. **(A)** KEGG pathway of DEPs. **(B)** GO enrichment of DEPs. **(C)** Pie chart of subcellular localization of DEPs.

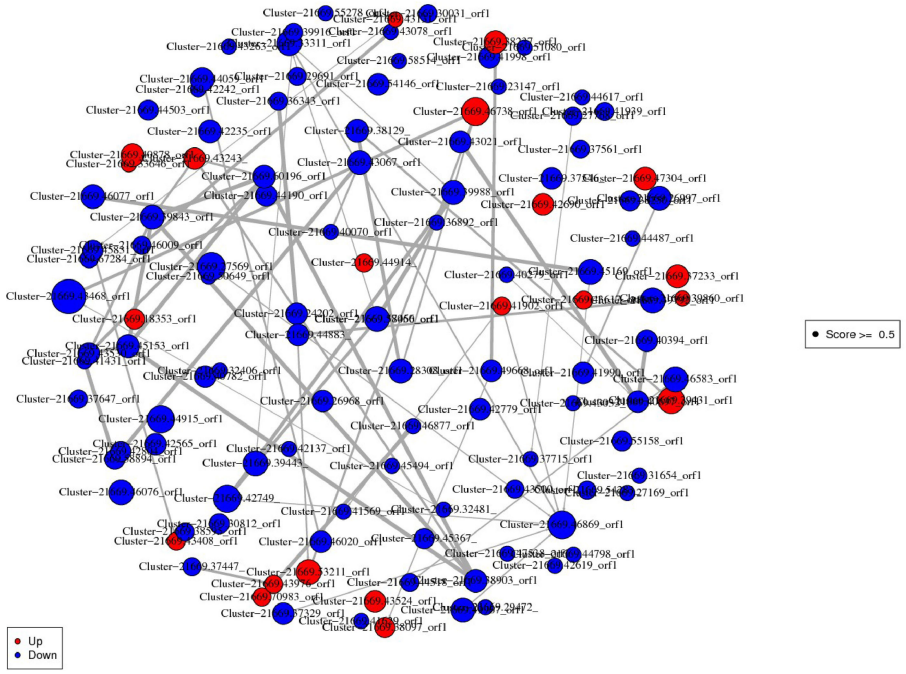


FIGURE 4
Interaction network diagram of DEPs. Red spots represent upregulated DEGs. Green spots indicate downregulated DEGs.

monoterpenes, 44 sesquiterpenes 1 benzenoids, and 1 Nitrogen (Table S6). Total of average percent of monoterpenoids emitted by A-phase and B-phase syconia was 8.29% and 37.08%, respectively. Average percent of sesquiterpenes emitted by A-phase and B-phase syconia was 88.43% and 55.02%, respectively. This variation was mainly due to the significant difference in dispersion between A-phase males and B-phase males syconia (P = 0.05). Caryophyllene was the most abundant compound emitted by A phase syconia and was the second by B phase syconia. Ocimene was the most abundant compound emitted by B phase syconia and fifth by A-phase syconia. The compounds were detected in B phase syconia but not in A-phase syconia monoterpenes were pinene, camphene, myrcene, camphor, menthol; and the sesquiterpenes were 7-Epi-sesquithujene, farnesene (Table 1).

Key genes, proteins and metabolites involved in terpenoid synthesis pathway in *F. hirta*

29 candidate enzyme genes were involved in the pathway of terpenoid synthesis precursors and 20 terpenoid synthases (TPSs) were identified in the transcriptome (Table S7). For the terpenoid synthesis pathway, except for 4-diphosphocytidyl methylerythritol synthase (CMS) and mevalonate kinase (MVK), at least one transcript of candidate enzyme genes was up-regulated in B-phase compared with *F. hirta* A-phase. Acetyl-CoA C-acetyltransferase protein (*ACAT2*), 3-hydroxy-3-methylglutaryl-coenzyme A reductase (*HMGR3*), geranylgeranyl diphosphate synthase (*GGPS2*), 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (*HDR*), geranyl diphosphate synthase (*GPS2*), *TPS2*, *TPS4-2*, *TPS4-3*, *TPS4-7*, *TPS14*, and *TPS10-4* were highly expressed in *F. hirta* B-phase Syconia (Figure 6A).

A total of 9 terpenoid synthesis proteins were identified in the proteome. Among them ACAT and hydroxymethylglutaryl-CoA synthase (HMGS) proteins in the MVA pathway were up-regulated, while 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR), 4-(cytidine 5'-diphospho)-2-c-methyl-d-erythritol kinase (CMK), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS) and MEP in the MEP pathway were down-regulated (Figure 6B).

20 monoterpenes and 44 sesquiterpenes were detected from *F. hirta* A-phase and B-phase Syconia (Table S6). We found pinene, myrcene and linalool contents were higher in the VOCs of B-phase syconia than in those of A-phase (Figure 6A and Table S6). The change of these metabolites content was consistent with the expression pattern of the corresponding *TPSs* (*TPS4*, *TPS10*, *TPS14*) gene.

10.3389/fpls.2022.1006291

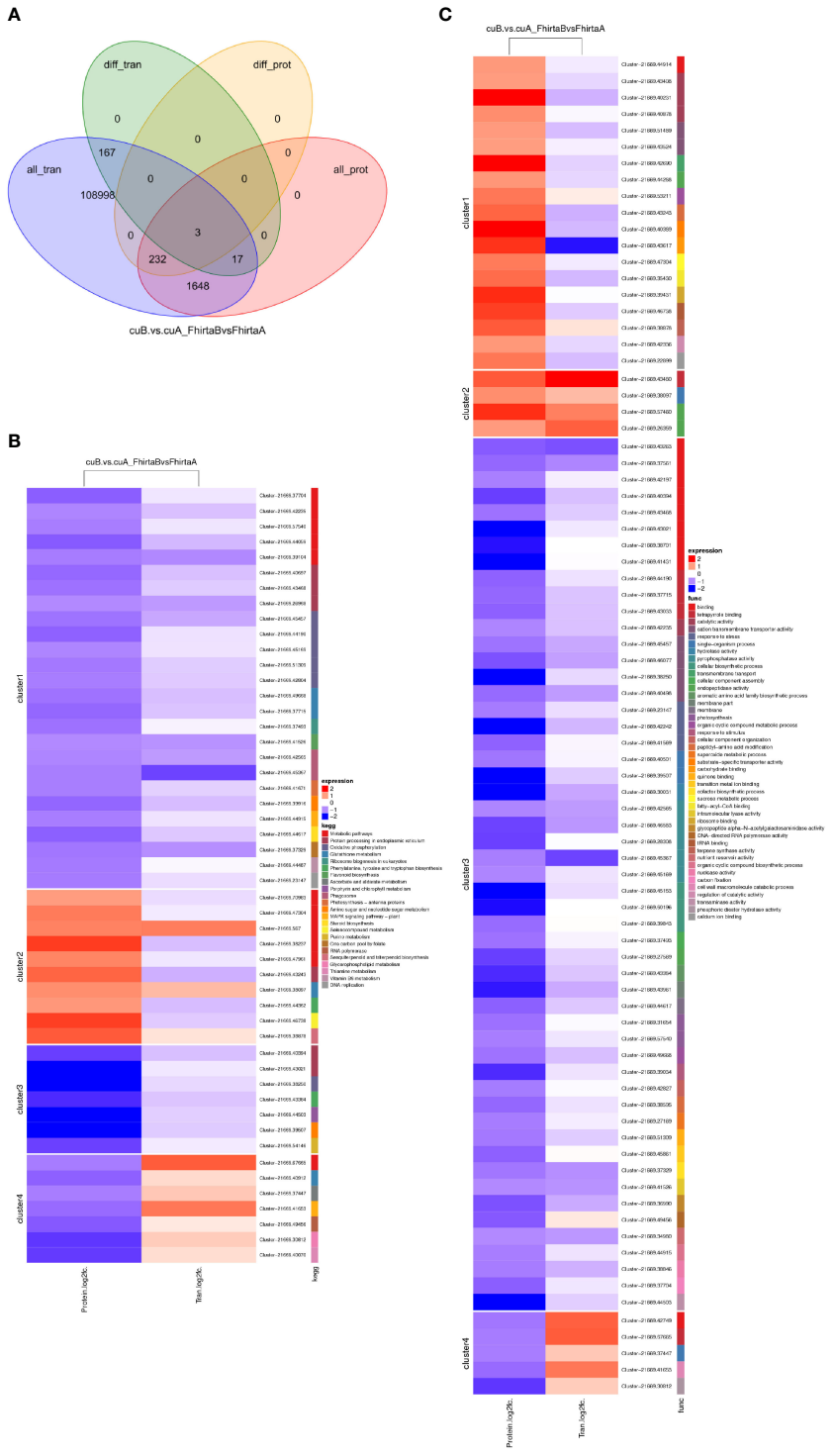


FIGURE 5
Correlation analysis of transcriptome and proteome. **(A)** Venn diagram of proteins expression and genes expression. **(B)** KEGG pathway of DEGs corresponding to DEPs. Red represents up-regulated and green represents down-regulated. FhirtaA VS FhirtaB represent multiples of transcript expression in both A phase and B phase of *F hirta* bracts. CuA VS CuB represents multiples of protein expression in both A phase and B phase of *F hirta* bracts. Fold Change (FC) represents multiples of genes/proteins expression in two groups. **(C)** GO categorization of DEGs corresponding to DEPs.

TABLE 1 Volatile compounds monoterpenes and sesquiterpenes emitted by *F. hirta* syconia.

Metabolite name	Normalized amount of volatile compound (%)	
Monoterpenes	A Phase	B Phase
Tricyclene	0.279701 ± 0.093901	0
Thujene <alpha->	0.106975 ± 0.076027	0.607962 ± 0.190082
.alpha.-Pinene	0	0.380554 ± 0.620119
Camphene	0	0.061219 ± 0.031016
Sabinene	0.420924 ± 0.134895	1.339409 ± 0.804251
.beta.-Myrcene	0	0.331641 ± 0.142777
Terpinene alpha	0.111256 ± 0.048199	0.196848 ± 0.086431
Cymene <para->	0.089512 ± 0.074377	0.159837 ± 0.018086
D-Limonene	0.142464 ± 0.109211	5.614226 ± 2.811749
Ocimene <(Z)-, beta->	0.148691 ± 0.031176	0.630069 ± 0.145269
Ocimene <(E)-, beta->	3.298357 ± 2.106637	20.33461 ± 2.174713
gamma-Terpinene	2.855996 ± 0.55994	3.084461 ± 0.1595
unknown 1084	0.035741 ± 0.040984	0.304717 ± 0.152347
Terpinolene	0.063345 ± 0.014799	0.160307 ± 0.099771
Linalool	0.191161 ± 0.0889	0.287481 ± 0.021421
Perillene	0.276044 ± 0.187177	0.062331 ± 0.027489
1,3,8-p-Menthatriene	0.026553 ± 0.028402	0.500437 ± 0.236772
Cosmene	0.246179 ± 0.138258	2.6029 ± 0.673882
Camphor	0	0.246274 ± 0.108866
Menthol	0	0.172707 ± 0.202493
Sesquiterpenes		
Elemene <delta->	0.495017 ± 0.239381	0.389853 ± 0.02331
Cubebene <alpha->	1.461843 ± 0.78094	0.370632 ± 0.038567
Cyclosativene	1.359822 ± 1.427221	0.659959 ± 0.214766
Copaene <alpha->	11.18039 ± 2.558977	1.925423 ± 0.530602
Daucene	0.517397 ± 0.096735	0.716682 ± 0.056523
Bourbonene <beta->	0.114903 ± 0.057929	0.02933 ± 0.003621
Cubebene <beta->	0.143267 ± 0.011609	0.244524 ± 0.066354
Elemene <beta->	3.690693 ± 1.060517	0.592203 ± 0.077555
Sesquithujene <7-epi->	0	0.741083 ± 0.152043
alpha.-Funebrene	3.460877 ± 2.168216	9.467905 ± 0.882513
Cedrene <alpha->	1.39091 ± 0.557673	1.520184 ± 0.181372
Caryophyllene <(E)->	25.1436 ± 12.23496	12.01404 ± 1.182516
Maaliene <beta->	0.134936 ± 0.040833	0
unknown 1428	0.952317 ± 0.693508	0.456136 ± 0.063767
.gamma.-Elemene	0.131153 ± 0.084364	0.128221 ± 0.069312
Calarene	0.053222 ± 0.026647	0.186355 ± 0.095928
Isogermacrene D	0.334866 ± 0.288424	0.169613 ± 0.031951
Guaiene <alpha->	0.673739 ± 0.593171	0.328577 ± 0.082898
Farnesene <(Z)-, beta->	0	0.292418 ± 0.138151
Humulene <alpha->	6.136025 ± 0.973794	3.524929 ± 0.271481
Caryophyllene <9-epi-(E)->	0.659508 ± 0.27048	0.073844 ± 0.127902
Muurola-4(14),5-diene <cis->	0.633604 ± 0.298112	0.355672 ± 0.100485
unknown 1469	0.082155 ± 0.030938	0.213291 ± 0.181688
Muurolene <gamma->	0.948525 ± 0.602066	0.46667 ± 0.198895
Germacrene D	4.588433 ± 2.555332	3.284334 ± 0.566488
Acoradiene <beta->	0.069279 ± 0.066496	0.328642 ± 0.091315

(Continued)

TABLE 1 Continued

Metabolite name	Normalized amount of volatile compound (%)	
Bicyclogermacrene	0.920439 ± 0.771135	1.678372 ± 1.106348
Selinene <beta->	0.450857 ± 0.271269	0.235989 ± 0.047297
Muurolene <alpha->	1.634127 ± 0.821371	2.416825 ± 2.211517
Amorphene <epsylon->	0.346487 ± 0.019319	0.320938 ± 0.093198
Cadinene <gamma->	0.917877 ± 0.578696	1.671503 ± 0.317758
Bulnesene	0.923225 ± 0.507382	0.792212 ± 0.2399
Bisabolene <(Z)-, alpha->	0.068381 ± 0.09625	0.080953 ± 0.055912
Cadinene <delta->	1.852318 ± 0.221276	1.057476 ± 0.186988
Cadinene <alpha->	0.171423 ± 0.038027	0.137721 ± 0.023957
Bisabolene <(E)-, gamma->	0.178962 ± 0.119737	0.321178 ± 0.049371
Calacorene <alpha->	0.2378 ± 0.050368	0.173671 ± 0.031566
Germacrene B	0.173741 ± 0.105013	0.188113 ± 0.069762
Calacorene <beta->	0.047576 ± 0.008562	0.063438 ± 0.025872
Cedrene <alpha-, epoxy->	8.765918 ± 12.64395	3.92387 ± 0.84527
Cedranone	3.272034 ± 2.72078	1.367145 ± 0.280374
Isolongifolen-5-one	0.577825 ± 0.418587	0.117931 ± 0.059512
.tau.-Cadinol	2.86869 ± 2.108536	1.706591 ± 0.526422
9-Cedranone	0.661865 ± 0.63051	0.287667 ± 0.096888

Normalized amount of volatile compound = (peak area of volatile compound)/(total peak area of all volatile compounds). Values represent means of three independent determinations.

Discussion

In previous research, it was speculated that phenylpropanoid and terpenoids were the main enriched VOCs components of *F. hirta* B-phase syconia by analyzing the genes related to VOCs (Hu et al., 2020). Terpenoids play ecological roles in pollinator attraction, allelopathy, and plant defense (Mahmoud and Croteau, 2002; Tholl, 2006). For the first time, we used a comprehensive analysis of transcriptome, proteome, and metabolism to understand the differences of the terpenoids synthesis pathways in the bracts of both A-phase and B-phase in *F. hirta*. It is important for us to understand molecular mechanism of maintaining obligate mutualism between the *Ficus* species and their pollinator wasps.

The VOCs emitted by syconia were mainly composed of terpenoids, which play ecological roles in pollinator attraction, allelopathy, and plant defense (Mahmoud and Croteau, 2002; Tholl, 2006; Parvin et al., 2014). Terpenoids were predominant in VOCs emitted by both A-phase and B-phase in *F. hirta* (Table S6). Most of the monoterpenoids were emitted by B-phase syconia higher than the A-phase, while most of sesquiterpenes were lower (Table 1). Monoterpenes Ocimene, D-Limonene, gamma-Terpinene were important component of VOCs emitted by B phase syconia (Table 1). β-myrcene, E-β-ocimene and D-limonene were elicited by *Mimulus lewisii* flowers evoke significant neural responses in bumblebee. Besides, a synthetic blend of the three chemical compound evoke the same responses as natural scents (Byers et al., 2014). The monoterpenes that were only detected in the VOCs of B-phase syconia were pinene, camphene, myrcene, camphor, menthol (Table 1). Compared with A-phase, the unique volatile monoterpenes emitted by B-phase syconia may play an important role in attracting pollinating fig wasps. Monoterpenoids, such as linalool, camphene, and cineole, were involved in pollinator attraction and allelopathy (Mahmoud and Croteau, 2002). The olfactory response experiment of weevil showed that camphene had an attracting effect on them (Sharaby and Al-Dosary, 2014). The olfactory receptor neuron of the pine weevil exhibited a strong response to α-pinene (Wibe et al., 1998). The olfactory responses of *Mahanarva spectabilis* to forage VOCs suggested that menthone, eucalyptol and camphor were all compounds likely to cause loss of attractiveness or repellence (Silva et al., 2019). Menthol exhibited moderate repellent effects on *Drosophila suzukii* (Corda et al., 2020). Compared with A-phase, the unique decanal in *Ficus pumila* var. *pumila* B-phase attracts pollinating fig wasps (Wang et al., 2021). We speculated that pinene, camphene, myrcene, camphor, and menthol play important roles to attract obligate pollinating fig wasps. Caryophyllene, funebrene, cedrene were main component of volatile sesquiterpenes emitted by B phase syconia (Table 1). (E)-caryophyllene was significantly increased in the damaged plants (Riffel et al., 2021). Red-rot disease infected plants release greater amounts of (+)-β-funerene than herbivore-infected plants (Peñaflor and Bento, 2018). Cedrene accumulated after whitefly infects tobacco,

10.3389/fpls.2022.1006291

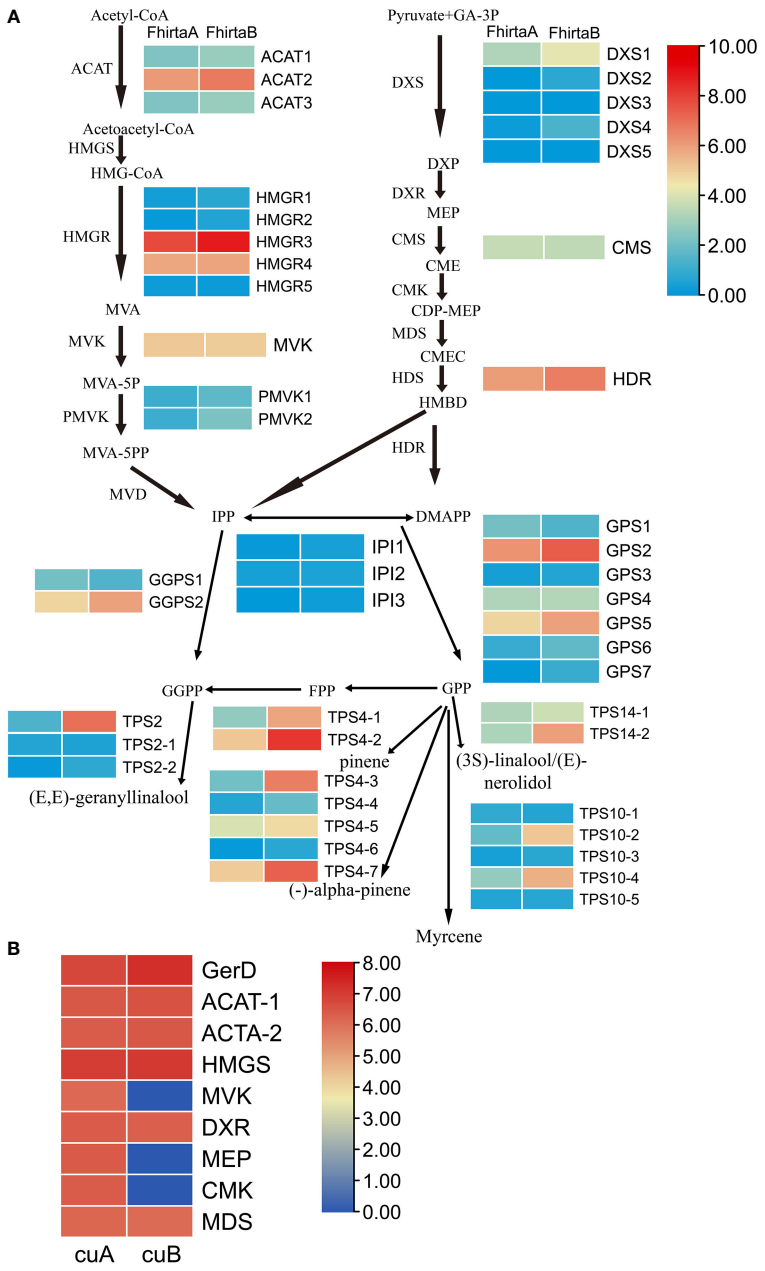


FIGURE 6
Expression profiles of genes and proteins involved in the terpenoid synthesis pathway of *F hirta*. **(A)** Expression levels for genes in A- and B-phase of *F hirta* bracts were shown by heatmap according to $\log_2$ (FPKM). **(B)** Expression levels for proteins in A- and B-phase of *F hirta* bracts were shown by heatmap according to $\log_{10}$ (Protein Abundance). CuA and CuB represent A phase and B phase of *F hirta* bracts, respectively.

conferring resistance to the whitefly (Luan et al., 2013). We suspect that the emitting of caryophyllene, funebrene and cedrene by B phase syconia are related to plant defense. In conclusion, monoterpenoids mainly attract pollinators, while sesquiterpenoids were related to plant defense response in *F. hirta*. If we want to know what attracts pollinating wasps, we need to perform experiments of electroantennographic detection coupled with gas chromatography (GC-EAD) and Y-tube olfactometer tests for identification. Combine these results and our results suggested that B-phase syconia need to produce more monoterpenoids and reduce the accumulation of sesquiterpenes to attract pollinators in *F. hirta*.

For terpenoids synthesis, transcription factors activate/repress their activities by binding to homeopathic elements in the promoter region of target gene, then regulate metabolic pathways. At the transcriptional level, the expression levels of *ACAT2*, *HMGR3*, *GGPS2*, *HDR*, *GPS2*, *TPS2*, *TPS4-2*, *TPS4-3*, *TPS4-7*, *TPS14*, *TPS10-4* increased in the bracts in B-phase (Figure 6A). Interestingly, pinene, myrcene and linalool were also higher in B-phase syconia than in those by A-phase (Table S6). This result indicates that up-regulation of *TPSs* gene expression can directly increase the content of related metabolites. Seven differentially expressed genes were screened from the transcriptome, *bHLH7*, *MYB1R1*, *PRE6* (*bHLH163*), *AIL1*, *RF2b*, *ANT*, and *VRN1* (Figure 1D). Transient expression of *AabHLH1* in *Artemisia annua* leaves increased the transcript level of *HMGR* (Ji et al., 2014). Transcription factors *MYB*, *NAC*, *ARF*, *WRKY*, *MYC*, *ERF* and *GRAS* were co-expressed with terpenoid biosynthesis genes, which may regulate terpenoid biosynthesis (Yu et al., 2021). *HY5*, a member of the *Arabidopsis* bZIP family of transcription factors, mediated the regulation of the terpene synthase AtTPS03 (Michael et al., 2020). MYB21 and MYC2 complex regulated *FhTPS1* expression in *Freesia hybrida* and *Arabidopsis* (Yang et al., 2020). Transcription factor *CitERF71* activated the terpene synthase gene *CitTPS16* by binding to the ACCCGCC and GGCGGG motifs of promoter, and was involved in the synthesis of E-geraniol in sweet orange fruit (Li et al., 2017). EREB58 can bind to the GCC-box in the promoter region of *TPS10* to promote its expression (Li et al., 2015). According to research progress in other plants and Plant Transcription Factor Database (http://planttfdb.gao-lab.org/index.php), we speculated *RF2b*, *VRN1* (Loukoianov et al., 2005), *ANT*, *PRE6* (Ferrero et al., 2019) and *MYB1R1* (Wang et al., 2016) were involved in regulating growth and development, and *AIL1* were involved in regulating ethylene signaling pathway. *bHLH7* may regulate metabolic processes (Chen, 2018). The differences in terpenoid synthesis pathways between A- and B-phase may be mainly caused by transcription factors regulating the expression of key enzyme genes, thereby regulating the synthesis of terpenoids. Interestingly, bHLH7 was not expressed in A-phase syconia bract but in the B phase. Based on this, we speculate that a key transcription factor *bHLH7* may regulate the expression of key enzymes involved in the terpenoid synthesis pathway in B-phase syconia.

In proteome, ACAT and HMGS in the MVA pathway were up-regulated, while DXR, CMK, MDS and MEP in the MEP pathway were down-regulated (Figure 6B). When one pathway is blocked with specific inhibitors, compensation can be observed with precursors produced by the other pathway (Henry et al., 2015; Mendoza-Poudereux et al., 2015). These results suggested that, at the protein level, the synthesis of terpenoids precursors in the B-phase of syconia were mainly synthesized through the MVA pathway.

In addition, there were three differentially expressed transcript/proteins that were related to adversity stress, namely Cluster-21669.61808 (aspartic protease), Cluster-21669.47934 (mannose-binding lectin), and Cluster-21669.43117 (non-specific lipid transfer protein) in transcriptome and proteome (Figure 5A, Tables S3, S5). The different developmental phases of syconia bracts involve many physiological and biochemical processes regulated by genes (Zhang et al., 2020; Wang et al., 2021). Aspartic proteinase was involved in protein processing and degradation, male and female gamete development, and played a vital role in plant coping with adversity stress (Huang et al., 2013; Scandola and Samuel, 2019; Soares et al., 2019). Mannose-binding lectin was important in controlling pests and diseases, resisting pathogenic microorganisms, and resisting higher herbivorous animals (Chen et al., 2021). The pepper mannose-binding lectin gene *CaMBL1* played a vital role in regulating cell death and defense responses to microbial pathogens (*Pseudomonas syringae* pv tomato, *Alternaria brassicicola*) (Hwang and Hwang, 2011). Overexpression of *OsJRL* (jacalin-related mannose-binding lectin) enhanced the salt tolerance and increased the expression levels of many stress-related genes in rice (He et al., 2017). Non-specific lipid-transfer proteins were involved in biotic stress, abiotic stress, and various metabolic processes (Tomassen et al., 2007; Liu et al., 2015; Gangadhar et al., 2016). These results indicate that many metabolic processes occurred in the bracts of A- and B-phase. And these processes produce metabolites in response to adversity stress, which provide a suitable environment for *F. hirta* B-phase syconia to attract pollinator fig wasps.

These findings contribute to understand the mechanism differences of terpenoids synthetic pathways in A- and B-phase syconia. Langenheim (1994) argued that terpenoids may be factors determining some properties of terrestrial plant communities and ecosystems. Higher plant terpenoids were closely related to many ecologically relevant characteristics (plant pollination, direct plant defense, allelopathy, formation of reactive gases in troposphere) (Langenheim, 1994; Mahmoud and Croteau, 2002; Tholl, 2006; Rosenkranz et al., 2021). Studying the mechanism differences of terpenoids synthetic pathways in both A- and B-phase syconia is important for further understanding the obligate mutualism and ecological implications.

Data availability statement

The transcriptome data presented in the study are deposited in the NCBI's Short Read Archive (SRA) repository, accession number PRJNA491590; The proteome data presented in the study are deposited in the integrated proteome resources

repository, accession number IPX0003971000 (https://www.iprox.cn/page/project.html?id=IPX0003971000).

Author contributions

SF, YJ and WL performed the experiments. SF and RW analyzed the data. FS, XC and HY wrote the paper. All authors read and approved the final paper.

Funding

This work was supported by the National Natural Science Foundation of China (Grants Nos. 31971568).

Acknowledgments

We thank Liao Yao-lin and Cheng Yu-fen for guidance on the GC-MS experiment. We also thank Novogene for guidance on proteome of *Ficus hirta* Vahl bracts.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1006291/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Protein qualitative data quality control. **(A)** Diagram of protein coverage distribution. **(B)** Diagram of precursor ions tolerance distribution. **(C)** Statistical diagram of protein molecular weight. **(D)** Diagram of unique peptide distribution.

SUPPLEMENTARY FIGURE 2
All samples were analyzed by principal component analysis (PCA) and coefficient of variance (CV). **(A)** PCA of the proteome data in a 2D graph of PC1 and PC2. The plot shows the effect for A-phase (CuA), B-phase (CuB). CuA and CuB represent A phase and B phase of *F. hirta* bracts, respectively. **(B)** Display of reproducibility CV of all samples.

SUPPLEMENTARY FIGURE 3
Pearson Correlation Coefficient analysis of the transcriptome and proteome. FhirtaA VS FhirtaB represent multiples of transcript expression in both A phase and B phase of F. hirta bracts. CuA VS CuB represent multiples of protein expression in both A phase and B phase of *F. hirta* bracts.

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Nat. Precedings* 1-1. doi: 10.1038/npre.2010.4282.1

Berg, C. C. (2003). Flora malesiana precursor for the treatment of moraceae 1: The main subdivision of ficus: The subgenera. *Blumea - Biodiver. Evol. Biogeogr. Plants* 48, 166–177. doi: 10.3767/000651903x686132

Brito, N. F., Moreira, M. F., and Melo, A. C. (2016). A look inside odorant-binding proteins in insect chemoreception. *J. Insect Physiol.* 95, 51–65. doi: 10.1016/j.jinsphys.2016.09.008

Byers, K. J., Bradshaw, H. D.Jr., and Riffell, J. A. (2014). Three floral volatiles contribute to differential pollinator attraction in monkeyflowers (Mimulus). *J. Exp. Biol.* 217, 614–623. doi: 10.1242/jeb.092213

Champagne, A., and Boutry, M. (2016). Proteomics of terpenoid biosynthesis and secretion in trichomes of higher plant species. *Biochim. Biophys. Acta* 1864, 1039–1049. doi: 10.1016/j.bbapap.2016.02.010

Chen, T.-Y. (2018). Analysis of bioinformatics and expression level of bHLH transcription factors in scutellaria baicalensis. *Chin. Traditional Herbal Drugs* 24, 671–677. Available at: https://pesquisa.bvsalud.org/portal/resource/pt/wpr-852222.

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, P., De Schutter, K., Pauwels, J., Gevaert, K., Van Damme, E. J. M., and Smagghe, G. (2021). Binding of orysata lectin induces an immune response in insect cells. *Insect Sci.* 29, 717–729. doi: 10.1111/1744-7917.12968

Chou, K.-C., and Shen, H.-B. (2008). Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. https://www.nature.com/articles/nprot.2007.494#article-info

Corda, G., Solari, P., Dettori, M. A., Fabbri, D., Delogu, G., Crnjar, R., et al. (2020). Association between olfactory sensitivity and behavioral responses of drosophila suzukii to naturally occurring volatile compounds. *Arch. Insect Biochem. Physiol.* 104, e21669. doi: 10.1002/arch.21669

Dudareva, N., Klempien, A., Muhlemann, J. K., and Kaplan, I. (2013). Biosynthesis, function and metabolic engineering of plant volatile organic compounds. *New Phytol.* 198, 16–32. doi: 10.1111/nph.12145

Dudareva, N., Negre, F., Nagegowda, D. A., and Orlova, I. (2006). Plant volatiles: Recent advances and future perspectives. *Crit. Rev. Plant Sci.* 25, 417–440. doi: 10.1080/07352680600899973

Ferrero, V., Viola, I. L., Ariel, F. D., and Gonzalez, D. H. (2019). Class I TCP transcription factors target the gibberellin biosynthesis gene GA20ox1 and the growth-promoting genes HBI1 and PRE6 during thermomorphogenic growth in arabidopsis. *Plant Cell Physiol.* 60, 1633–1645. doi: 10.1093/pcp/pcz137

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

Galil, J., and Eisikowitch, D. (1968). Flowering cycles and fruit types of ficus sycomorus in Israel. *New Phytol.* 67, 745–758. doi: 10.1111/j.1469-8137.1968.tb05497.x

Gangadhar, B. H., Sajeesh, K., Venkatesh, J., Baskar, V., Abhinandan, K., Yu, J. W., et al. (2016). Enhanced tolerance of transgenic potato plants over-expressing non-specific lipid transfer protein-1 (StnsLTP1) against multiple abiotic stresses. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01228

Gu, D., Compton, S. G., Peng, Y., and Yang, D. (2012). 'Push' and 'pull' responses by fig wasps to volatiles released by their host figs. *Chemoecology* 22, 217–227. doi: 10.1007/s00049-012-0108-8

He, X., Li, L., Xu, H., Xi, J., Cao, X., Xu, H., et al. (2017). A rice jacalin-related mannose-binding lectin gene, OsJRL, enhances escherichia coli viability under high salinity stress and improves salinity tolerance of rice. *Plant Biol. (Stuttg)* 19, 257–267. doi: 10.1111/plb.12514

Henry, L. K., Gutensohn, M., Thomas, S. T., Noel, J. P., and Dudareva, N. (2015). Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10050–10055. doi: 10.1073/pnas.1504798112

Hossaert-McKey, M., Gibernau, M., and Frey, J. (1994). Chemosensory attraction of fig wasps to substances produced by receptive figs. *Entomologia experimentalis applicata* 70, 185–191. doi: 10.1111/j.1570-7458.1994.tb00746.x

Hossaert-McKey, M., Proffit, M., Soler, C. C., Chen, C., Bessiere, J. M., Schatz, B., et al. (2016). How to be a dioecious fig: Chemical mimicry between sexes matters only when both sexes flower synchronously. *Sci. Rep.* 6, 21236. doi: 10.1038/srep21236

Hossaert-McKey, M., Soler, C., Schatz, B., and Proffit, M. (2010). Floral scents: their roles in nursery pollination mutualisms. *Chemoecology* 20, 75–88. doi: 10.1007/s00049-010-0043-5

Huang Da, W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, J., Zhao, X., Cheng, K., Jiang, Y., Ouyang, Y., Xu, C., et al. (2013). OsAP65, a rice aspartic protease, is essential for male fertility and plays a role in pollen germination and pollen tube growth. *J. Exp. Bot.* 64, 3351–3360. doi: 10.1093/jxb/ert173

Hussain, T., Asrar, H., Zhang, W., Gul, B., and Liu, X. (2021). Combined transcriptome and proteome analysis to elucidate salt tolerance strategies of the halophyte panicum antidotale retz. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.760589

Hu, R., Sun, P., Yu, H., Cheng, Y., Wang, R., Chen, X., et al. (2020). Similitudes and differences between two closely related ficus species in the synthesis by the ostiole of odors attracting their host-specific pollinators: A transcriptomic based investigation. *Acta Oecologica* 105, 103554. doi: 10.1016/j.actao.2020.103554

Hwang, I. S., and Hwang, B. K. (2011). The pepper mannose-binding lectin gene CaMBL1 is required to regulate cell death and defense responses to microbial pathogens. *Plant Physiol.* 155, 447–463. doi: 10.1104/pp.110.164848

Ji, Y., Xiao, J., Shen, Y., Ma, D., Li, Z., Pu, G., et al. (2014). Cloning and characterization of AabHLH1, a bHLH transcription factor that positively regulates artemisinin biosynthesis in artemisia annua. *Plant Cell Physiol.* 55, 1592–1604. doi: 10.1093/pcp/pcu090

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., Mcanulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kearns, C. A., Inouye, D. W., and Waser, N. M. (1998). Endangered mutualisms: the conservation of plant-pollinator interactions. *Annu. Rev. Ecol. systematics* 29, 83–112. https://www.jstor.org/stable/221703

Kumar, S. R., Shilpashree, H. B., and Nagegowda, D. A. (2018). Terpene moiety enhancement by overexpression of geranyl(geranyl) diphosphate synthase and geraniol synthase elevates monomeric and dimeric monoterpene indole alkaloids in transgenic catharanthus roseus. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00942

Langenheim, J. H. (1994). Higher plant terpenoids: a phytocentric overview of their ecological roles. *J. Chem. Ecol.* 20, 1223–1280. doi: 10.1007/BF02059809

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.* 12, 1–16. doi: 10.1186/1471-2105-12-323

Liu, M., Wang, W. G., Sun, H. D., and Pu, J. X. (2017). Diterpenoids from isodon species: an update. *Nat. Prod Rep.* 34, 1090–1140. doi: 10.1039/c7np00027h

Liu, F., Zhang, X., Lu, C., Zeng, X., Li, Y., Fu, D., et al. (2015). Non-specific lipid transfer proteins in plants: presenting new advances and an integrated functional analysis. *J. Exp. Bot.* 66, 5663–5681. doi: 10.1093/jxb/erv313

Li, S., Wang, H., Li, F., Chen, Z., Li, X., Zhu, L., et al. (2015). The maize transcription factor EREB58 mediates the jasmonate-induced production of sesquiterpene volatiles. *Plant J.* 84, 296–308. doi: 10.1111/tpj.12994

Li, X., Xu, Y., Shen, S., Yin, X., Klee, H., Zhang, B., et al. (2017). Transcription factor CitERF71 activates the terpene synthase gene CitTPS16 involved in the synthesis of e-geraniol in sweet orange fruit. *J. Exp. Bot.* 68, 4929–4938. doi: 10.1093/jxb/erx316

Loukoianov, A., Yan, L., Blechl, A., Sanchez, A., and Dubcovsky, J. (2005). Regulation of VRN-1 vernalization genes in normal and transgenic polyploid wheat. *Plant Physiol.* 138, 2364–2373. doi: 10.1104/pp.105.064287

Luan, J. B., Yao, D. M., Zhang, T., Walling, L. L., Yang, M., Wang, Y. J., et al. (2013). Suppression of terpenoid synthesis in plants by a virus promotes its mutualism with vectors. *Ecol. Lett.* 16, 390–398. doi: 10.1111/ele.12055

Madzharova, E., and Sabino, F. (2019). "Exploring extracellular matrix degradomes by TMT-TAILS n-terminomics," in *Collagen* (New York, NY, Springer), 115–126.

Mahmoud, S. S., and Croteau, R. B. (2002). Strategies for transgenic manipulation of monoterpene biosynthesis in plants. *Trends Plant Sci.* 7, 366–373. doi: 10.1016/s1360-1385(02)02303-8

Ma, L. T., Lee, Y. R., Liu, P. L., Cheng, Y. T., Shiu, T. F., Tsao, N. W., et al. (2019). Phylogenetically distant group of terpene synthases participates in cadinene and cedrane-type sesquiterpenes accumulation in taiwania cryptomerioides. *Plant Sci.* 289, 110277. doi: 10.1016/j.plantsci.2019.110277

Mendoza-Poudereux, I., Kutzner, E., Huber, C., Segura, J., Eisenreich, W., and Arrillaga, I. (2015). Metabolic cross-talk between pathways of terpenoid backbone biosynthesis in spike lavender. *Plant Physiol. Biochem.* 95, 113–120. doi: 10.1016/j.plaphy.2015.07.029

Michael, R., Ranjan, A., Kumar, R. S., Pathak, P. K., and Trivedi, P. K. (2020). Light-regulated expression of terpene synthase gene, AtTPS03, is controlled by the bZIP transcription factor, HY5, in arabidopsis thaliana. *Biochem. Biophys. Res. Commun.* 529, 437–443. doi: 10.1016/j.bbrc.2020.05.222

Okamoto, T., and Su, Z.-H. (2021). Chemical analysis of floral scents in sympatric ficus species: highlighting different compositions of floral scents in morphologically and phylogenetically close species. *Plant Systematics Evol.* 307, 1–12. doi: 10.1007/s00606-021-01767-y

Parvin, R., Shahrokh, K. O., Mozafar, S., Hassan, E., and Mehrdad, B. (2014). Biosynthesis, regulation and properties of plant monoterpenoids. *J. Medicinal Plants Res.* 8, 983–991. doi: 10.5897/jmpr2012.387

Peñaflor, M. F. G. V., and Bento, J. M. S. (2018). Red-rot infection in sugarcane attenuates the attractiveness of sugarcane borer-induced plant volatiles to parasitoid. *Arthropod-Plant Interact.* 13, 117–125. doi: 10.1007/s11829-018-9629-6

Proffit, M., Lapeyre, B., Buatois, B., Deng, X., Arnal, P., Gouzerh, F., et al. (2020). Chemical signal is in the blend: bases of plant-pollinator encounter in a highly specialized interaction. *Sci. Rep.* 10, 10071. doi: 10.1038/s41598-020-66655-w

Riffel, A., Silva Filho, B. F., Santos, S. P. A., Silva, W. L., Ribeiro, T. F. L., Oliveira, D. J. A., et al. (2021). Exposure to sugarcane borer-induced plant volatile (E)-caryophyllene enhances parasitoid recruitment. *Entomologia Experimentalis Applicata* 169, 937–946. doi: 10.1111/eea.13081

Rosenkranz, M., Chen, Y., Zhu, P., and Vlot, A. C. (2021). Volatile terpenes - mediators of plant-to-plant communication. *Plant J* 108, 617-631. doi: 10.1111/tpj.15453

Rosenkranz, M., and Schnitzler, J. P. (2016). "Plant volatiles," in *eLS*, 1–9. doi: 10.1002/9780470015902.a0000910.pub3

Scandola, S., and Samuel, M. A. (2019). A flower-specific phospholipase d is a stigmatic compatibility factor targeted by the self-incompatibility response in brassica napus. *Curr. Biol.* 29, 506–512.e504. doi: 10.1016/j.cub.2018.12.037

Sharaby, A., and Al-Dosary, M. (2014). An electric air flow olfactometer and the olfactory response of rhynchophorous ferrugineus weevil to some volatile compounds. *J. Agric. Ecol. Res. Int.* 1, 40–50. https://www.researchgate.net/publication/343671410_Article_no_JAERI2014004_SCIENCEDOMAIN_international_Original_Research_Article_Sharaby_and_Al-Dosary

Silva, S. E. B., Auad, A. M., Moraes, J. C., Alvarenga, R., Fonseca, M. G., Marques, F. A., et al. (2019). Olfactory response of mahanarva spectabilis (Hemiptera: Cercopidae) to volatile organic compounds from forage grasses. *Sci. Rep.* 9, 10284. doi: 10.1038/s41598-019-46693-9

Soares, A., Ribeiro Carlton, S. M., and Simoes, I. (2019). Atypical and nucellin-like aspartic proteases: emerging players in plant developmental processes and stress responses. *J. Exp. Bot.* 70, 2059–2076. doi: 10.1093/jxb/erz034

Soler, C., Hossaert-Mckey, M., Buatois, B., Bessiere, J. M., Schatz, B., and Proffit, M. (2011). Geographic variation of floral scent in a highly specialized pollination mutualism. *Phytochemistry* 72, 74–81. doi: 10.1016/j.phytochem.2010.10.012

Soler, C. C., Proffit, M., Bessiere, J. M., Hossaert-Mckey, M., and Schatz, B. (2012). Evidence for intersexual chemical mimicry in a dioecious plant. *Ecol. Lett.* 15, 978–985. doi: 10.1111/j.1461-0248.2012.01818.x

Souza, C. D., Pereira, R. A., Marinho, C. R., Kjellberg, F., and Teixeira, S. P. (2015). Diversity of fig glands is associated with nursery mutualism in fig trees. *Am. J. Bot.* 102, 1564–1577. doi: 10.3732/ajb.1500279

Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr. Opin. Plant Biol.* 9, 297–304. doi: 10.1016/j.pbi.2006.03.014

Tomassen, M. M., Barrett, D. M., van der Valk, H. C., and Woltering, E. J. (2007). Isolation and characterization of a tomato non-specific lipid transfer protein involved in polygalacturonase-mediated pectin degradation. *J. Exp. Bot.* 58, 1151–1160. doi: 10.1093/jxb/erl288

Wan, C., Chen, C., Li, M., Yang, Y., Chen, M., and Chen, J. (2017). Chemical constituents and antifungal activity of ficus hirta vahl. fruits. *Plants (Basel)* 6, 44. doi: 10.3390/plants6040044

Wang, J., Wu, F., Zhu, S., Xu, Y., Cheng, Z., Wang, J., et al. (2016). Overexpression of OsMYB1R1-VP64 fusion protein increases grain yield in rice by delaying flowering time. *FEBS Lett.* 590, 3385–3396. doi: 10.1002/1873-3468.12374

Wang, R., Yang, Y., Jing, Y., Segar, S. T., Zhang, Y., Wang, G., et al. (2021). Molecular mechanisms of mutualistic and antagonistic interactions in a plant-pollinator association. *Nat. Ecol. Evol.* 5, 974–986. doi: 10.1038/s41559-021-01469-1

Wibe, A., Borg-Karlson, A.-K., Persson, M., Norin, T., and Mustaparta, H. (1998). Enantiomeric composition of monoterpene hydrocarbons in some conifers and receptor neuron discrimination of α-pinene and limonene enantiomers in the pine weevil, hylobius abietis. *J. Chem. Ecol.* 24, 273–287. doi: 10.1023/A:1022580308414

Wu, X., Xiong, E., Wang, W., Scali, M., and Cresti, M. (2014). Universal sample preparation method integrating trichloroacetic acid/acetone precipitation with phenol extraction for crop proteomic analysis. *Nat. Protoc.* 9, 362–374. https://www.nature.com/articles/nprot.2014.022

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Yang, Z., Li, Y., Gao, F., Jin, W., Li, S., Kimani, S., et al. (2020). MYB21 interacts with MYC2 to control the expression of terpene synthase genes in flowers of freesia hybrida and arabidopsis thaliana. *J. Exp. Bot.* 71, 4140–4158. doi: 10.1093/jxb/eraa184

Ye, X., Tian, W., Wang, G., Zhang, X., Zhou, M., Zeng, D., et al. (2020). Phenolic glycosides from the roots of ficus hirta vahl. and their antineuroinflammatory activities. *J. Agric. Food Chem.* 68, 4196–4204. doi: 10.1021/acs.jafc.9b07876

Yi, T., Chen, Q., He, X., So, S., Lo, Y., Fan, L., et al. (2013). Chemical quantification and antioxidant assay of four active components in ficus hirta root using UPLC-PAD-MS fingerprinting combined with cluster analysis. *Chem. Cent J.* 7, 115. doi: 10.1186/1752-153X-7-115

Yu, N., Chen, Z., Yang, J., Li, R., and Zou, W. (2021). Integrated transcriptomic and metabolomic analyses reveal regulation of terpene biosynthesis in the stems of sindora glabra. *Tree Physiol.* 41, 1087–1102. doi: 10.1093/treephys/tpaa168

Yu, H., and Nason, J. D. (2013). Nuclear and chloroplast DNA phylogeography of ficus hirta: obligate pollination mutualism and constraints on range expansion in response to climate change. *New Phytol.* 197, 276–289. doi: 10.1111/j.1469-8137.2012.04383.x

Yu, H., Zhao, N.-X., Chen, Y.-Z., Deng, Y., Yao, J.-Y., and Ye, H.-G. (2006). Phenology and reproductive strategy of a common fig in guangzhou. *Bot. Stud.* 47, 435–441. https://ejournal.sinica.edu.tw/bbas/content/2006/4/Bot474-10.pdf

Zhang, X., Wang, G., Zhang, S., Chen, S., Wang, Y., Wen, P., et al. (2020). Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell* 183, 875–889.e817. doi: 10.1016/j.cell.2020.09.043

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 December 2022
DOI 10.3389/fpls.2022.1029540

OPEN ACCESS

EDITED BY
Rong Wang,
East China Normal University, China

REVIEWED BY
Huan Fan,
Xishuangbanna Tropical Botanical Garden (CAS), China
Chao Hu,
Shanghai Chenshan Plant Science Research Center (CAS), China

*CORRESPONDENCE
Nagarjun Vijay
nagarjun@iiserb.ac.in

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 27 August 2022
ACCEPTED 17 November 2022
PUBLISHED 12 December 2022

CITATION
Patil AB, Vajja SS, Raghavendra S, Satish BN, Kushalappa CG and Vijay N (2022) Jack of all trades: Genome assembly of Wild Jack and comparative genomics of *Artocarpus*. *Front. Plant Sci.* 13:1029540. doi: 10.3389/fpls.2022.1029540

Jack of all trades: Genome assembly of Wild Jack and comparative genomics of *Artocarpus*

Ajinkya Bharatraj Patil[1], Sai Samhitha Vajja[1], S. Raghavendra[2], B. N. Satish[3], C. G. Kushalappa[3] and Nagarjun Vijay[1]*

[1]Computational Evolutionary Genomics Lab, Department of Biological Sciences, Indian Institute of Science Education and Research (IISER), Bhopal, Madhya Pradesh, India, [2]College of Agriculture Hassan, University of Agricultural Sciences (UAS), Bangalore, Karnataka, India, [3]College of Forestry, Ponnampet, Karnataka, India

Artocarpus (Moraceae), known as breadfruits for their diverse nutritious fruits, is prized for its high-quality timber, medicinal value, and economic importance. Breadfruits are native to Southeast Asia but have been introduced to other continents. The most commonly cultivated species are *Artocarpus heterophyllus* (Jackfruit) and *Artocarpus altilis* (Breadfruit). With numerous smaller but nutritionally comparable fruits on a larger tree, *Artocarpus hirsutus*, also called "Wild Jack" or "Ayani", is an elusive forest species endemic to Indian Western Ghats. In this study, we sequenced and assembled the whole genome of *Artocarpus hirsutus* sampled from the sacred groves of Coorg, India. To decipher demographic and evolutionary history, we compared our Wild Jack genome with previously published Jackfruit and Breadfruit genomes. Demographic history reconstruction indicates a stronger effect of habitat rather than phylogeny on the population histories of these plants. Repetitive genomic regions, especially LTR Copia, strongly affected the demographic trajectory of *A. heterophyllus*. Upon further investigation, we found a recent lineage-specific accumulation of LTR Copia in *A. heterophyllus*, which had a major contribution to its larger genome size. Several genes from starch, sucrose metabolism, and plant hormone signal transduction pathways, in *Artocarpus* species had signatures of selection and gene family evolution. Our comparative genomic framework provides important insights by incorporating endemic species such as the Wild Jack.

KEYWORDS

Wild Jack, *Artocarpus*, Breadfruit, Jackfruit, Western Ghats, gene family evolution, positive selection, lineage-specific selection

Introduction

Genus *Artocarpus* (Moraceae), or "Breadfruits," are tropical plants famous for their nectary and fleshy fruits (Jarrett, 1977). This genus comprises ~70 species with considerable variability in size, height, flower/fruit morphology, developmental processes, and functional properties (Zerega et al., 2010; Gardner et al., 2021). Most of the members of the genus provide a rich resource of food, timber, and other valuable products, popularising them in their native regions (Jagtap and Bapat, 2010; Xavier et al., 2014; Ragone, 2018). As a consequence of such properties, some species have been introduced to various parts of the world. The two most widely distributed domesticated species, *Artocarpus heterophyllus* (Jackfruit) and *Artocarpus altilis* (Breadfruit), currently have oriental distribution in the tropical and subtropical regions (Zerega et al., 2010; Williams et al., 2017). However, *Artocarpus* trees are native to the region extending from the Western Ghats, South-East Asia, to the Oceanic Islands. Although a recent study suggested the diversification of *Artocarpus* from Borneo followed by subsequent dispersal and divergence during the Miocene (Williams et al., 2017), multiple fossils from India dated to the Palaeocene suggest an earlier presence (Mehrotra et al., 1984; Srivastava, 1998). Despite being unlikely, the Bornean origin of *Artocarpus* suggests overwater or overland dispersal across large distances as the only possibility for Indian *Artocarpus* species to exist (Williams et al., 2017). Hence, the biogeographical history of these plants is yet to be established and is a matter of further research. Differences in the bioclimatic properties of their habitats and the fauna involved in their pollination/dispersal might have played an instrumental role in adapting these species by developing divergent characteristics from their ancestral counterparts.

Artocarpus trees are well known for their diversity of unique unisexual inflorescences and composite syncarpous fruits (Jarrett, 1977). The phenotypic diversity among the syncarps is such that the taxonomy of this genus is entirely dependent upon inflorescence morphology and structure (Zerega et al., 2010). Even though the focus has been on the floral diversity for delineating these species, these plants have evolved several other species-specific characteristics. The trees of *A. heterophyllus* reach a height of 15-20 meters and have reticulate branching close to the soil, whereas the trees of *A. altilis* reach up to 30 meters and are moderately branched at a medium height from the ground (Figure 1). As opposed to these two, Wild Jack (*Artocarpus hirsutus*) are large forest trees that reach above 50 meters, some extending to 70 m with no branching until the apices. The male inflorescences differ in all three species. *A. heterophyllus* has smaller cylindrical inflorescence than *A. altilis*, which has longer and thicker apices. In contrast, *A. hirsutus* has a thin, long, filamentous stalk and the male inflorescence differs entirely from the other two species. Female inflorescences also differ in these three species, so their fruit morphology is quite diverse. Jackfruit (*A. heterophyllus*) bears multiple, low-hanging, larger, ellipsoidal, fleshy, nectary, and green-sheathed fruits of sizes up to 100 cm. The Breadfruit (*A. altilis*) bears numerous medium-sized, oval, starchy, and green-sheathed fruits of 12-20 cm, hanging at apices of branches of medium heights. In comparison, Wild Jack (*A. hirsutus*) bears multiple oval/ellipsoidal, fleshy, smaller, orange/yellow sheathed fruits of size 6-10 cm at the apices of branches of higher heights. Therefore, such diverse phenotypic characteristics suggest differentiated pollinator/disperser networks and mechanisms (Jarrett, 1977; Matthew et al., 2006; Jagtap and Bapat, 2010; Ragone, 2018; Buddhisuharto et al., 2021).

The Wild Jack (*A. hirsutus*) is unique in its phenotype compared to the other two popular *Artocarpus* species. Due to its endemic distribution in the Western Ghats and its forests, it has received minimal attention and is still understudied (Matthew et al., 2006). However, in its native range, it is a multipurpose plant of economic and ecological importance. It provides long, high-quality, pathogen-resistant timber and has been widely used for building houses, boats, and large, long-lasting structures (Matthew et al., 2006; Xavier et al., 2014; Meenu et al., 2021). The other parts, like fruits and seeds, are used as a rich source of energy, and the constituents of this fruit are comparable or superior to the other two species, which demands further research into this plant (Solanki et al., 2020). Lastly, leaves, seeds, and bark are used in traditional medicinal treatments (Matthew et al., 2006; Jagtap and Bapat, 2010; Solanki et al., 2020; Buddhisuharto et al., 2021; Meenu et al., 2021). Wild Jack reaches maturity for timber harvesting in around 20 years, and the present distribution of the species cannot fulfill the increasing demand of the timber market, which makes it vulnerable to population decline. Hence the re-assessment of conservation status and efforts to effectively conserve this plant is warranted (Matthew et al., 2006; Liu et al., 2020). The sacred groves of the Western Ghats are fragmented forests protected by locals due to religious importance. But recently, these forests have been threatened due to deforestation and developmental projects (Matthew et al., 2006; Osuri et al., 2014; Wang et al., 2020). The biodiversity hotspot of the Western Ghats is home to multiple endemic flora and fauna (Myers et al., 2000), but there is a paucity of genomic data from these species. One such species is the Wild Jack, a ubiquitous constituent of these ancient forests conserving this species (Chandrashekara and Sankar, 1998; Tambat et al., 2005). By generating genomic resources, our study aims to incorporate Wild Jack in a comparative framework with other *Artocarpus* species. Hence, our goal is to identify genomic changes related to differential phenotypes and acclimatization to their habitats.

The phenotypic characteristics of *Artocarpus* species are quite distinct including their inflorescence structures, which tend to be associated with adaptive changes (Harder and Johnson, 2009; Harder and Prusinkiewicz, 2013). To uncover the genomic basis of phenotypic diversity between *Artocarpus*

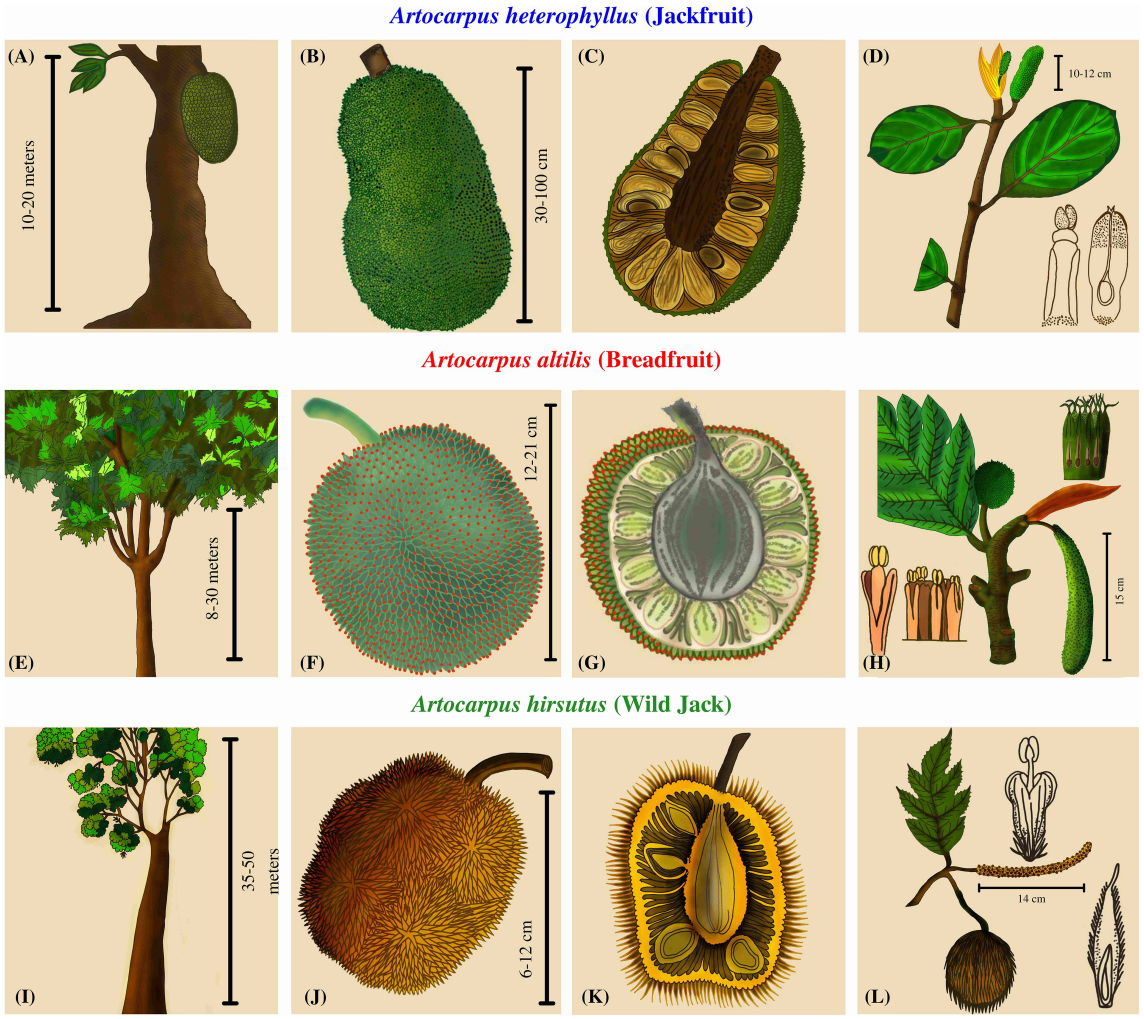


FIGURE 1
Plant morphology and comparative phenotypic characters of three *Artocarpus* species. **(A, E, I)** The first column depicts the overall structure of the trees. **(B, F, J)** The second column contains drawings of the fruit describing the color, shape, and size. **(C, G, K)** The third column has a cross-sectional view of the fruits depicting the number and arrangement of seeds. **(D, H, L)** The fourth column has drawings of the inflorescences.

species, we require genomic information of Wild Jack to compare it with other genomes. Lineage-specific gene family changes and/or signatures of selection are potent drivers of phenotypic evolution and adaptation (Lespinet et al., 2002; Van Der Lee et al., 2017). Similarly, repeat accumulation can also lead to lineage-specific phenotypes (Negi et al., 2016; Li et al., 2018; Wang et al., 2018; Ramakrishnan et al., 2021). Therefore, we

a) Sequence and assemble the Wild Jack genome (and plastome), i.e., *Artocarpus hirsutus*.
b) Perform gene annotation to identify orthologous gene sequences across order Rosales members and construct a species tree.
c) Analyse gene family evolution in these species, especially in *Artocarpus* members.
d) Employ multiple methods to detect the signatures of selection in the genes of all *Artocarpus* species.
e) Detailed repeat sequence annotation to understand repeat accumulation dynamics.

We wanted to evaluate the role of differential bioclimatic history on the demographic trends of species from differing habitats and whether it will be strongly affected by the ecological differences. We chose phylogenetically related species from distinct habitats to address this question. Hence, we

a) Use demographic reconstructions to analyse population size history and

b) Species distribution modelling to assess species range dynamics for *Artocarpus* species.

Materials and methods

Sample collection, genome sequencing, and assembly

A. hirsutus is endemic to the Western Ghats and its forests. We located a fruit-bearing tree near the College of Forestry, Ponnampet (GPS coordinates 12°08′56.5″N 75°54′32.5″E; Altitude: 829–850 m Above Sea level) and sampled some leaves for the sequencing. The samples collected for sequencing are preserved and cataloged (ART_HIR_WG_IISERB). A photograph of the leaf specimen is provided in Supplementary Figure 1. A leaf was cleaned, sanitised, and then cut into pieces for further processing. The whole genomic DNA was extracted from the leaves using DNeasy plant mini kit from QIAGEN. The quality of the extracted DNA was evaluated by observing the DNA band on 1% Agarose gel for shearing. The concentration and purity of extracted DNA were assessed using QUBIT 3.0 and Nanodrop. The purified DNA was then used to prepare Illumina short-read (150 bp) libraries with TruSeq DNA Nano Library Prep Kit with an insert size of 450 ± 50 bp. We sequenced ~ 88X coverage of the genome with Illumina short-read paired-end data using the Illumina Novaseq 6000 sequencer.

The quality of whole-genome sequencing (WGS) paired-end reads was assessed using FASTQC. Barcode sequences were trimmed if present using Cutadapt (Martin, 2011). These sequencing reads were used for the estimation of genome size. We used Jellyfish (Marçais and Kingsford, 2011) to perform Kmer analysis using a kmer-size (k) of 21 and hash size (-m) of 100M on the sequencing reads. GenomeScope (Vurture et al., 2017) was then used to estimate genome size and heterozygosity. We used the Celera assembler implemented in MaSuRCA version 4.0.6 (Maryland Super Read Cabog Assembler) for assembling the sequencing data (Zimin et al., 2013). The published assemblies of *A. heterophyllus* and *A. altilis* were used as a reference for the synteny-assisted assembly step of MaSuRCA. We used Quast (Gurevich et al., 2013) to calculate genome assembly metrics such as N50 and L50 (Supplementary Table 1). BUSCO version 3 (Simão et al., 2015) was used to assess the genome completeness with the eudicotyledons_odb10 dataset (Supplementary Table 2).

Repeat annotation and analyses

For *de-novo* identification and annotation of the repetitive genomic regions and/or transposable elements, we used RepeatModeler version 2 (Flynn et al., 2020), with the LTR_struct option to include LTR models identified by programs such as LTR-FINDER (Xu and Wang, 2007) and LTR-Harvest (Gremme et al., 2013). The consensus fasta library obtained by RepeatModeler 2 was then used as input to RepeatMasker (Smit, AFA, Hubley, R & Green) to annotate, mask and tabulate the repeat content and their types. The resultant output file was then used to soft mask the genome for further analyses. The RepeatMasker.align output was used to calculate Kimura two-parameter divergence estimates (TE age) between the repeat families for all species using accessory scripts provided with RepeatMasker suite like buildSummary.pl, calcDivergenceFromAlign.pl, and createRepeatLandscape.pl. The obtained output was summarised to plot histograms of Kimura divergence values to visualise the distribution of repeat families across the time scale. The genome size of the species from order Rosales was correlated with the percent repeat content in their assemblies. To further nullify the effect of phylogenetic relatedness on the correlation, a correction was done using the PIC (Phylogenetic Independent Contrast) method implemented in the R package phytools.

Genome annotation

We used MAKER version 2 (Campbell et al., 2014) to annotate the genome's coding regions. Three rounds of the maker pipeline were executed to obtain the final annotated genesets. In the first round of homology-based annotation, we used protein fasta sequences from all the species of order Rosales available on NCBI, including *A. altilis*, and *A. heterophyllus* (Supplementary Table 3). The mRNA evidence from *A. altilis* was provided as alternative transcript sequences. The obtained genesets from this round were then used to generate training gene models for *de-novo* gene annotation algorithms like SNAP (Johnson et al., 2008) and AUGUSTUS (Stanke and Morgenstern, 2005). New gene models were identified during both rounds, and existing gene models were refined. Genesets after the third round were considered final and used to get coding sequences and translated protein sequences. We performed BUSCO on the resultant protein dataset to assess the quality of the annotations. We further used EggNOG functional annotation algorithm to get the gene names and GO annotations. We also used blastp with ARAPORT 11 database to validate the gene models.

Chloroplast assembly, annotation, and analysis

The chloroplast sequence was independently assembled using WGS reads with NOVOPlasty version 4.3.1

(Dierckxsens et al., 2017). The chloroplast genome sequence of *A. altilis* (NCBI accession: NC_059002.1) was used as a reference for the algorithm, and the Maturase K gene sequence of *A. hirsutus* (NCBI accession: KU856362.1) was used as a seed, which is used as assembly generation point. The resultant assembly produced two contigs with only one arrangement possibility leading to a complete circular genome sequence spanning ~162Kbp. The chloroplast assembly was then annotated using GeSeq (Tillich et al., 2017), and the circular genome was depicted and visualised using OGDRAW (Greiner et al., 2019) implemented in CHLOROBOX. Currently available chloroplast genomes from the *Artocarpus* genus and outgroup species *Ficus religiosa* and *Morus indica* were downloaded from NCBI. To investigate rearrangements between these chloroplast genomes, they were aligned with ProgressiveMauve aligner (Darling et al., 2010) and visualised in Mauve alignment viewer (Darling et al., 2004). To identify the phylogenetic positions of these genomes, we aligned the genomes using the MAFFT aligner (Katoh et al., 2002). The appropriate substitution model was estimated using Modeltest-ng (Darriba et al., 2020), and the phylogenetic tree was constructed using Raxml-ng (Kozlov et al., 2019) with 1000 bootstraps. The chloroplast genomes of *A. heterophyllus* and *A. integer* show an inversion for the SSC (Small Single Copy) region compared to other *Artocarpus* sp. plastomes (Supplementary Figure 2).

Identification of orthologous sequences and construction of species tree

The translated coding sequences of *A. hirsutus* and 13 other species (*A. altilis, A. heterophyllus, Morus notabilis, Parasponia andersonii, Trema orientale, Cannabis sativa, Rhamnella rubrinervis, Ziziphus jujuba, Malus baccata, Malus domestica, Prunus persica, Fragaria vesca,* and *Rosa chinensis*) were concatenated and used to find orthologs. We used Orthofinder (Emms and Kelly, 2019) to find orthologous genic sequences across 14 species with parameters to use MSA alignments to obtain the orthogroups using diamond blast (Buchfink et al., 2021), MAFFT (Katoh et al., 2002) and fasttree (Price et al., 2009). The orthologous gene sequences in which *A. hirsutus* is present were tabulated. These gene IDs were used to get corresponding CDS sequences from each species. These CDS sequences for all the genes were then aligned using GUIDANCE version 2 (Sela et al., 2015) with the MAFFT aligner (Katoh et al., 2002). All the resultant CDS alignments were concatenated and used to find partitions and models using IQTREE version 2 (Minh et al., 2020). After that, the loci and concatenated trees were obtained to get bootstrap support with additional metrics such as the Gene concordance factor (gCF) and Site concordance factor (sCF). Following these evaluations, the tree was exported and rooted at a branch leading to *F. vesca* and *R. chinensis*.

Comparative genomics and gene family analyses

The translated coding sequences of 4 species, *A. hirsutus, A. altilis, A. heterophyllus*, and *M. notabilis* were used to identify overlapping and non-overlapping gene clusters using Orthovenn version 2 (Xu et al., 2019). Orthovenn identified and constructed the unique and common gene clusters for all four species. Unique gene clusters of *Artocarpus* species were subjected to GO enrichment analysis. We used CAFÉ version 5 software (Mendes et al., 2021) for gene family analyses of contractions and expansions. We first concatenated protein sequences of 14 species used for species tree construction and made a blast database. This 14-species protein database was used as a subject to perform all vs. all protein blast (blastp) (Camacho et al., 2009). The blast results were then used as input for mcxload to create network and sequence dictionary files. The clustering was performed using mcl clustering software (Li et al., 2003) with an inflation parameter (-I) of 3. The cluster files were then reformatted, and the gene families with large gene copy numbers were removed from the analyses. The constructed species tree was converted to an ultrametric tree using r8s software (Sanderson, 2003) using a divergence estimate of 87 MYA (Million Years Ago) between *P. persica* and *Z. jujube* obtained from TimeTree (Kumar et al., 2017). The filtered clustering file of MCL and the ultrametric tree were then used as input for the CAFÉ 5. The clade-based gene family expansion/contraction results were then summarised and represented on the phylogeny. Out of all significant gene family contractions, we selected only those gene families with a difference of five gene copies at the least between the species. By enforcing these stringent criteria, we got seventeen, three, and seven gene families significantly expanded in *A. hirsutus*, *A. heterophyllus*, and *A. altilis*, respectively.

Lineage-specific selection tests in *Artocarpus* genes

Tests of selection intensity among species for the same orthologous genes in a phylogenetic framework provide opportunities to identify loci under relaxed or intensified pressures in a focal species of interest. This selection pressure analysis helps us identify evolutionary changes and signs of adaptations to their bioclimatic niche. We used multiple approaches to identify selection pressures in *Artocarpus* to understand the evolutionary mechanisms and processes these species have undergone. We used branch-site models implemented in PAML version 4.9 (Yang, 2007) and aBSREL (Adaptive Branch-site Random Effects Likelihood) (Smith et al., 2015) implemented in HYPHY to test for positively selected branches. We also used RELAX (Wertheim et al., 2015)

(intensification parameter; K > 1) implemented in HYPHY to identify the genes under intensified selection. For detecting strong purifying or relaxed selection, we implemented the branch site model of PAML version 4.9 and RELAX (relaxed parameter; K < 1) of HYPHY. To reduce the false positive results, we compared the list of genes identified as positively selected by all three methods and considered only those genes that were overlapping/common between them. The functional roles of positively selected genes were cross-referenced using KEGG (Kanehisa and Goto, 2000), FLOR-ID (Bouché et al., 2016), ARAPORT11 (Cheng et al., 2017), and TAIR (Rhee et al., 2003) databases.

Demographic history reconstruction

The genomic sequencing reads of one individual each of *A. hirsutus*, *A. altilis*, and *A. heterophyllus* were mapped to the respective genome assemblies using the BWA MEM aligner (Li, 2013). The alignments were converted to binary, sorted, and indexed using samtools (Li et al., 2009). These binary alignments were then used to call consensus sequence using bcftools (Li and Barrett, 2011). To assess the effect of genomic regions such as exonic, intronic, intergenic, and repetitive elements on the demographic estimation, we masked each part to evaluate the impact of the respective fraction. We masked the respective genomic region using BEDTOOLS maskfasta and followed similar steps mentioned above to get the demographic estimation. To assess the effect of each individual repeat family/type, we followed the published protocol/scripts (Patil and Vijay, 2021). We quantified the concordance between the trajectories estimated from different repeat types within each species using a non-parametric measure of intraclass correlation implemented in the "nopaco: Non-Parametric Concordance Coefficient" R package (Rothery, 1979). We calculated the difference in N_e estimates between the Unmasked and Masked trajectories and the differences between the Unmasked and each repeat type. Using the difference between unmasked and masked trajectories as the maximum deviation in trajectories, we evaluated which repeat types had a similarly large deviation from the unmasked estimates. For this, we performed Wilcoxon tests between the (Unmasked-masked) and (Unmasked-each repeat type) (see Supplementary Table 4). The repeat types with significant differences are not major contributors to the masked estimates. Therefore, the comparisons with non-significant p-values between (Unmasked-masked) and (Unmasked-each repeat type) are the repeats that have contributed the most to the deviation from unmasked estimates.

We used filters like -C50, -Q30, -q20 for bcftools mpileup to ensure quality bases and mapped reads to be considered in the variant calling. The consensus calls were converted to the required (fastq) format using vcfutils vcf2fq using a quality filter of 25, whereas calls with less than one-third and more than twice the mean coverage were excluded during this step to exclude false calls. These consensus calls were then converted to input format (.psmcfa) for psmc using fq2psmcfa. The input psmcfa file was then used to run the psmc program (Li and Durbin, 2011) with options -N 25 -t 5 -r 5 -p 4 + 25*2+4+6. The output of psmc was inspected for a sufficient number of recombination events. At first, we used a mutation rate (μ) of *Populus trichocarpa*, i.e., 2.5e-09 per site per year (Tuskan et al., 2006) with a generation time of 15 years to execute the psmc_plot.pl script to get scaled demographic trajectories for each species.

Estimation of mutation rate

Since we were trying to study demographic effects on these three species comparatively, we needed to understand the bottleneck events for these species from their native ranges. *A. hirsutus* and *A. heterophyllus* are native to the Western Ghats and should have experienced similar demographic events. Our initial scaled PSMC plots for both species with the same mutation rates did not align with the starting point of the trajectory. These trajectories created a possibility that there might be mutation rate differences between these three species. To obtain a reliable mutation rate estimate, we sampled orthologous alignments in which only four species (*A. hirsutus, A. altilis, A. heterophyllus*, and *M. notabilis*) are present. We fixed an input un-rooted tree structure to allow branch-specific comparisons possible. We used codon alignments of ~1500 genes to estimate a (d_4) 4-fold degenerate site substitution rate (parameters used, model = 0, NSsites=0, seqtype=1, CodonFreq=2, runmode=0) using PAML version 4.9. The obtained d_4 rates for all alignments were summarised, and the mean value for these estimates was considered d_4 for individual species. These mean d_4 estimates were then divided by the divergence time between compared branches or species. The estimates obtained were then considered a proxy of the respective species' mutation rates (Nadachowska-Brzyska et al., 2015).

Species distribution modelling

We downloaded species occurrence data corresponding to the native range of each species as identified earlier (Williams et al., 2017) from the GBIF (Global Biodiversity Information Facility) database for all three *Artocarpus* species (Gbif.Org, 2022). We used the method of Ecological niche modelling (ENM) to predict the species distribution during three paleoclimatic eras: Last Glacial Maximum (LGM, approx. 20,000 years ago), Last interglacial (LIG, approx. 110,000-130,000 years ago), and Marine Isotope Stage 19 (MIS19, approx. 750,000-790,000). The environmental variables for

these periods were extracted from PaleoClim (Brown et al., 2018) at a resolution of 2.5 min arc. Environmental layers were resized to the species' native range using the software DIVA-GIS (version 7.5) (Hijmans et al., 2012). We considered the annual and excluded the seasonal parameters for highly correlated bioclimatic variables. The set of variables used was chosen based on species-specific considerations for the compared periods (see Supplementary Tables 5B-C).

The ENM was performed using the software MaxEnt (version 3.4.4). The settings for MaxEnt were species and paleoclimatic era-specific. We used the R package ENMeval, which identifies settings that balances model fit and increases the predictive ability (see Supplementary Table 5E) (Muscarella et al., 2014). The following settings were set by default: 10000 background points, 500 maximum iterations, ten runs of cross-validations, and the regularisation multiplier were explicitly based on ENMeval results. We saved the output in cloglog form, which is the simplest to understand and the default output format. It gives the probability of occurrence estimate between 0 to 1. We selected the mean of all ten replicate runs to represent each species across each time period. The average of the population count across these 10 runs was calculated as the Population Count (PC). The number of grid cells with a habitat suitability index > 0.9 was calculated as the Grid cell Count (GC). The accuracy in the prediction of species distribution was analysed through the use of a receiver operating characteristics (ROC) plot. In the ROC plots, all the values fell between 0 and 1 (AUC: Area Under the Curve). All the values were above 0.5 and are considered better than random when the curve lies above the diagonal, indicated by the AUC (see Supplementary Table 5A) (Merow et al., 2013). A Jackknife test was performed to find the different contributions of variables and to identify the ones with a maximum contribution (see Supplementary Table 5D). The habitat suitability maps of species distribution were generated using R.

Results

Genome sequencing, assembly, and annotation

The whole genome sequencing of Wild Jack (*A. hirsutus*) yielded ~475 million Illumina short reads (71.65 Gigabases). The genome assemblies of previously published congeneric species vary from ~800 Mbp (*A. altilis*) to ~980 Mbp (*A. heterophyllus*); their Kmer-based genome size estimates are 812 Mbp and 1005 Mbp, respectively (Sahu et al., 2019). The 21 Kmer value-based genome size estimate for *A. hirsutus* is 635.16 Mbp with 1.16% heterozygosity, a smaller genome size estimate than other two congenerics. The resultant genome assembly was 791.16 Mbp in length. Our resultant assembly captures nearly complete genomic information for *A. hirsutus* as it is substantially greater than the Kmer-based estimate but closer to the nearest congeneric, *A. altilis*. The generated genome assembly has coverage of ~ 90X. The assembly has a contig N50 of 50.25 Kbp with L50 of 4630. Our assembled genome has 96.7% of complete BUSCO's, indicative of a nearly complete assembly. More than 98% of the sequencing reads were mapped to the genome assembly. The LTR-Retriever's LAI score for the assembly was 6.28, and indicates it is a draft genome assembly. The MAKER pipeline annotated 46,957 gene models with a mean length of 2387.62 bp. BUSCO identified 94.6% of protein gene sets from the MAKER as complete, which indicates good annotation and an almost complete gene set.

Lineage-specific gene family dynamics

Gene family expansion and diversification are prominent drivers of phenotypic evolution. Comparative analysis of *Artocarpus* genomes and outgroups from Rosales' order identified changes in gene family composition (see Figures 2A, B and Supplementary Tables 6, 7). Notably, lineage-specific genes identified by orthovenn2 in *A. hirsutus* are enriched for pollen recognition genes (GO: 0048544), which are Receptor Kinases from the Lectin domain-containing gene family (Supplementary Table 7). We also found evidence of lectin gene family expansion based on CAFÉ analysis in *A. hirsutus* (41 copies) compared to *A. altilis* (33 copies) and *A. heterophyllus* (22 copies) (see Figure 1C, RK3 panel, Supplementary Table 8). Among the lectin genes with orthologs across all three *Artocarpus* species, we detect signatures of intensified selection using RELAX and positive selection using the PAML branch site and aBSREL (Supplementary Figure 3). Lectin domain-containing proteins have diverse functions in biotic and abiotic stress response, plant growth, and development (Sun et al., 2020; Saidou and Zhang, 2022). Therefore, our results suggest diversification of lectin domain-containing proteins in *A. hirsutus*.

Apart from lectins, *A. hirsutus* also showed lineage-specific gene family expansions in at least 15 other gene families with functions varying from pollen/flower development (AGL18/15, ARA1F, PTR3, EDA17), secondary metabolite biosynthesis (F6'H, IIL1, HSD7, DTX28 (Upadhyay et al., 2020), MDL3), stress tolerance and defence, i.e., biotic (IRE1, IIL1, DTX28, MDL3) and abiotic (F6'H, CYP81D8, GLR3.6, SCPL34, PTR1/5, HSP17-6B, OPR2), growth and development (mTERF7, HSD, AGL18/15, ARA1F) and plant-pathogen interactions (IRE1, F6'H, IIL1, OPR2, MDL3) (see Figures 2C). All these numerous gene family expansions may reflect the concerted evolution of this plant to acclimatise to biotic and abiotic conditions and adapt to its habitat.

In *A. heterophyllus*, the lineage-specific genes identified by orthovenn2 are enriched for Toll-Interleukin-Resistance (TIR) domain proteins, Receptor Like Protein 33 (RLP33), and the

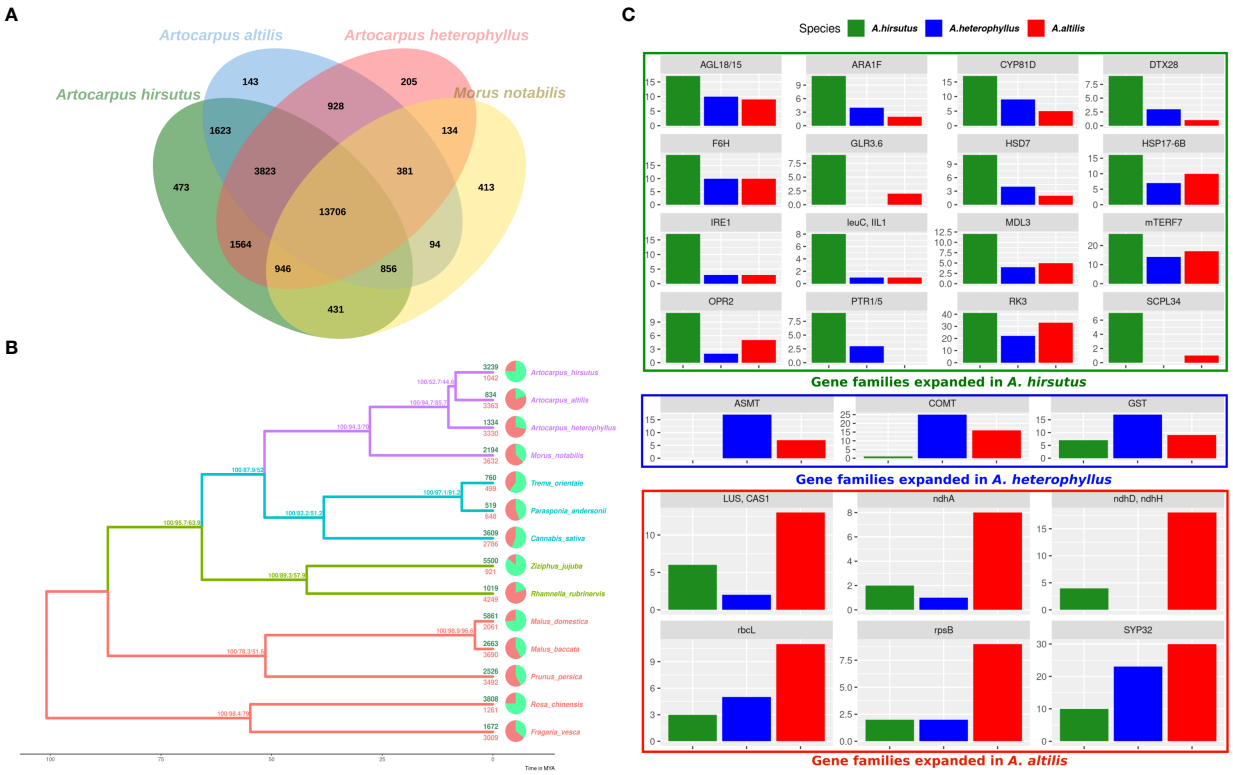


FIGURE 2
Comparative genomics of orthologous gene content evolution in *Artocarpus* and their relatives. **(A)** Orthovenn diagram of three Artocarpus species (*A. hirsutus, A. altilis,* and *A. heterophyllus*) with *M. notabilis*. Unique and shared gene clusters between these species are denoted in respective intersections. *A. hirsutus* has 473 unique gene clusters not shared with other congenerics. **(B)** Representative species tree with gene family evolution or Gene gain-losses: 14 species from Order Rosales of family Moraceae (purple), Rhamnaceae (sea green), Cannabaceae (olive green), and Rosaceae (red) are represented in the species tree based on ~4500 orthologous genes. Branch values are bootstraps values, gCF (gene concordance factors) and sCF (site concordance factors), respectively. Pie charts and associated numbers at the tips show gene families' expansions (green) and contractions (red). **(C)** Significantly expanded gene families in *A. hirsutus* (green), *A. heterophyllus* (blue), and *A. altilis* (red).

flavonoid biosynthesis pathway. TIR domain proteins and RLP33 are well known for foreign pattern recognition and providing immunity to plants from microbes (Burch-Smith and Dinesh-Kumar, 2007; Jamieson et al., 2018). In addition, the two most essential genes of the Flavanoid Biosynthesis Pathway, Chalcone Synthase (CHS) and Flavanone 3-Hydroxylase (F3H) have lineage-specific gene copies and may explain the high flavonoid content of *A. heterophyllus* (Meera et al., 2018). The copy number of both ASMT (N-Acetylserotonin Methyltransferase) and COMT (Caffeic Acid O-methyltransferase) genes is higher in *A. heterophyllus* (17 and 25 copies) compared to *A. hirsutus* (0 and 1 copies) and *A. altilis* (7 and 16 copies) (Supplementary Figure 4). ASMT and COMT genes act in the penultimate step of the melatonin pathway (Back et al., 2016; Zhao et al., 2019). Furthermore, COMT also plays an important role in the lignin biosynthesis pathway (Wang et al., 2013). Lastly, the gene family of Glutathione S-Transferases (GST), which have a role in stress tolerance, has also expanded in *A. heterophyllus*.

The lineage-specific genes of *A. altilis* are enriched for Hexokinase-3, ABCB27 (ATP-Binding Cassette B27) or ALS1 (Aluminium Sensitive 1), and mTERF15. Hexokinase-3 is involved in sugar processing, primarily glucose and plant development (Paulina Aguilera-Alvarado and Sanchez-Nieto, 2017). ABCB27 or ALS1 are transporters involved in stress response to Aluminium-rich or Acidic soils (Kar et al., 2021). The transcription factor mTERF15 modulates the expression of mitochondrial assembly factor I genes, specifically NAD2/3 (NADH ubiquinone oxidoreductases), and regulates energy generation (Hsu et al., 2014). Interestingly, we found gene family expansions in multiple genes of mitochondria and chloroplast (ndhH, ndhD, ndhA, rbcL, and rpsB) in *A. altilis*. These expansions of organellar genes and their regulators might be due to a higher energy requirement caused by oxidative stress or other stressors. The triterpenoid biosynthesis synthase genes like Cycloartenol Synthase (CAS1) and Lupeol synthase 2/5 (LUP2/5) (Thimmappa et al., 2014; Cárdenas et al., 2019) and essential pollen development proteins, syntaxin of plants

(SYP31/32) (Rui et al., 2021) were also increased in copy number.

Habitat rather than phylogeny determines the population histories

We found that *A. altilis* underwent demographic contraction ~ 2 to 1 million years ago (MYA), followed by extensive population expansion from ~ 1 MYA to 150 thousand years ago (KYA), which marks the start of the Holocene or Last Glacial Period (Figures 3C, 4A). The demographic expansion of *A. altilis* from MIS-19 to the LIG is accompanied by an increase in habitat suitability (i.e., GC increases from 1517 to 1586 and PC increases from 28267.39 to 30603.35). In contrast to *A. altilis*, the population sizes of *A. heterophyllus* and *A. hirsutus* experienced extensive expansion from ~ 2 to 1 MYA, followed by a contraction in population size from ~ 1 MYA to 200 KYA (Figures 3A, B, 4A). The demographic contraction in both species from MIS-19 to the LIG is accompanied by corresponding reductions in habitat suitability for *A. heterophyllus* (i.e., GC decreases from 1695 to 1598 and PC decreases from 52174.614 to 47000.784) and

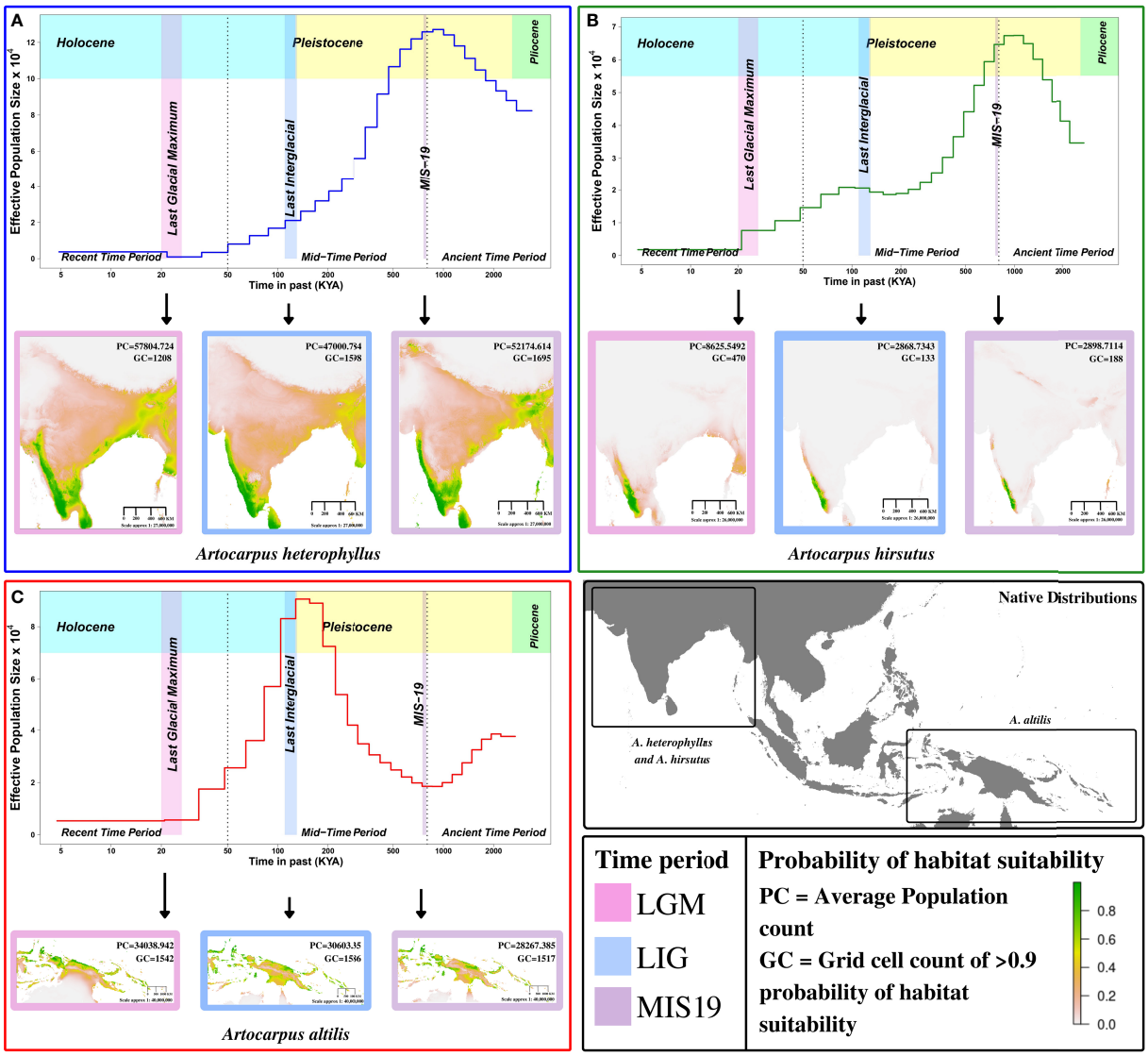


FIGURE 3
Demographic history reconstruction and species distribution modelling of *Artocarpus* species. Demographic history reconstruction using PSMC with Species distribution modelling for LGM (Last Glacial Maximum), LIG (Last Interglacial), and MIS19 (Marine Isotope Stage 19) for **(A)** *A. heterophyllus* (blue), **(B)** *A. hirsutus* (green), **(C)** *A. altilis* (red). The two dotted vertical lines are used to demarcate the recent, mid and ancient time periods following the timeline shown. The native distribution of all three species are highlighted in the bottom right map of the continent.

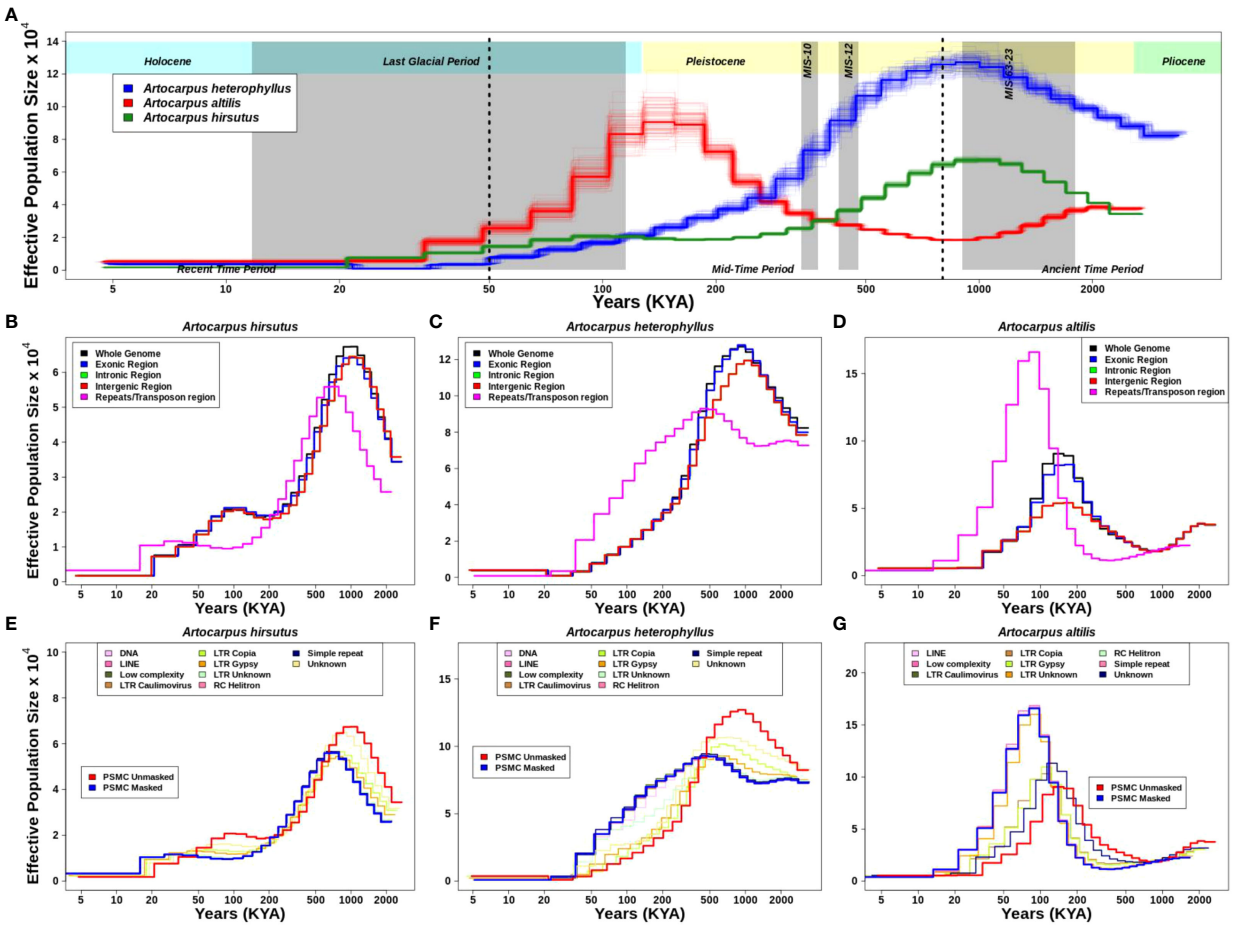


FIGURE 4
Demographic history reconstruction of three *Artocarpus* species. **(A)**. Demographic history of *A. heterophyllus* (blue), *A. altilis* (red), and *A. hirsutus* (green) are represented by the respective trajectories with additional bootstrapped lines. The two dotted vertical lines are used to demarcate the recent, mid and ancient time periods following the timeline shown. The grey vertical shading represents the corresponding glacial event. Demographic trajectories and the effect of genome fraction used for PSMC inference in **(B)**. *A. hirsutus*, **(C)**. *A. heterophyllus*, and **(D)**. *A. altilis*. The effect of different repeat types on PSMC inference is represented in **(E)**. *A. hirsutus*, **(F)**. *A. heterophyllus*, and **(G)**. *A. altilis*. The amount of change due to LTR-Copia is most prominent in *A. heterophyllus*.

A. hirsutus (i.e., GC decreases from 188 to 133 and PC decreases from 2898.7 to 2868.7343). After the onset of the Holocene, the effective population size declined in *A. altilis* and *A. heterophyllus*. However, in *A. hirsutus* the population size recovered and stabilised before undergoing another round of population decline in the Holocene. The discrepancy between SDM and PSMC in the recent time period might be due to inability of PSMC to reliable estimates in this time period. Comparing demographic histories in the mid and ancient time periods among the three *Artocarpus* species and the species distribution models suggests that bioclimatic changes and the habitat have been instrumental in shaping the population histories. In conclusion, the demographic histories of the *Artocarpus* species reflect the effects of habitat more than their phylogenetic relatedness (Figure 4A).

The estimates of historical effective population size (N_e) reflect evolutionary processes such as actual changes in population size, population structure, gene flow (Mazet et al., 2015, 2016), and linked selection (Schrider et al., 2016) and/or regions of the genome used (Patil and Vijay, 2021). Hence, we evaluated the effect of different genomic regions in estimating demographic histories. Exon, intron, and intergenic region-masked trajectories matched with the whole-genome-based trajectory, which explains that these individual regions of the genome are not drastically changing the estimates (Figures 4B-D). However, masking repetitive genomic regions resulted in two types of changes in the inferred trajectory. The less noticeable trajectory change results in a diagonal shift towards recent time intervals in *A. hirsutus* and *A. altilis*. The more drastic change in trajectory occurs in *A. heterophyllus*, where the

repeat masked and whole-genome-based N_e estimates have a lower concordance. The measures of concordance between the trajectories estimated from different repeat types were higher for *A. altilis* (0.915) and *A. hirsutus* (0.942) compared to *A. heterophyllus* (0.872). The pairwise differences in the concordance between species found that *A. altilis* and *A. hirsutus* did not differ significantly (p-value: 0.07). However, the comparison of both (*A. altilis* vs. *A. heterophyllus*: delta is 0.0465 and p-value is 0.000553 and *A. hirsutus* vs. *A. heterophyllus*: delta is 0.0723 and p-value is 1.86e-05) these species with *A. heterophyllus* has significant differences. Overall, our results indicate that repeats have influenced the demographic inferences of *A. heterophyllus* (concordance between masked and unmasked genomes psi: 0.783 and p-value: 1.83e-05) more than the other two species (*A. altilis*; psi:0.94 and p-value:1.18e-30, *A. hirsutus*; psi:0.892 and p-value:3.4e-18).

To understand which type of repeat regions affect the inference of demographic history in these species, we investigated the effect of each repeat type. In all three species, the shift in trajectories among LTRs (i.e., LTR-Unknown, LTR-Gypsy, and LTR-Copia) was highest (Figures 4E-G). Other repeat families, like simple repeats, DNA transposons, low complexity regions, etc., mirrored the masked trajectory and had no effect of masking on demography. Unknown repeat types had the most noticeable impact on the trajectories of *A. hirsutus* (Wilcoxson test p-value: 0.03513) and *A. altilis* (Wilcoxson test p-value: 0.07581), whereas LTR-Copia (Wilcoxson test p-value: 0.9934) greatly impacted the N_e estimates of *A. heterophyllus*. The most surprising result of masking repetitive regions occurs in the N_e estimates of *A. heterophyllus*, which drastically changes the trajectory in magnitude and shape mainly due to LTR-Copia.

Differential abundance/accumulation of repeat families in *A. heterophyllus*

We found that the repetitive genomic regions strongly affected the demographic analyses, which demands further detailed characterisation of repeat families and their contents. We compared the types of repeat families assembled in those 14 Rosales genomes and their abundances (see Figure 5A). Of the three *Artocarpus* genomes, *A. hirsutus* (481 Mbp; 60.51% of the genome) and *A. altilis* (505 Mbp; 60.68% of the genome) have a comparable composition of repeat types. In contrast, *A. heterophyllus* (614 Mbp; 62.56% of the genome) has a higher overall repeat content than the other two species. Specifically, the abundance of LTR-Copia in the *A. heterophyllus* genome (246.5 Mbp; 25.1% of the genome) was highly elevated compared to *A. altilis* (131 Mbp; 15.7% of the genome) and *A. hirsutus* (128 Mbp; 16.1% of the genome). Other than LTR-Copia, most other

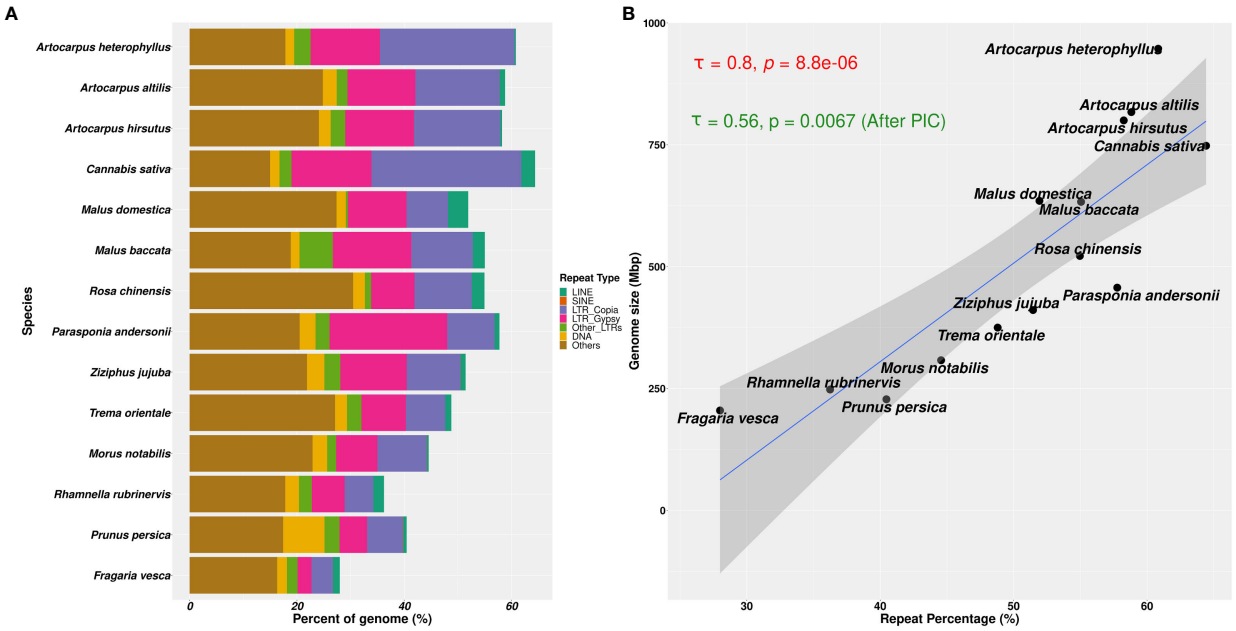


FIGURE 5
Repetitive genomic regions and their abundance in Rosales genomes. **(A)** Repeat content percentages in genomes of 14 species of Rosales with the abundance of types of repeats across these species. **(B)** Correlation of genome sizes with percentage of repeat content assembled in these 14 genomes. There is a strong positive correlation (τ = 0.8 and p-value = 8.8e-06, after Phylogenetically Independent Contrasts (PIC): τ = 0.56 and p-value = 0.0067) between the percentage of repeat sequences assembled in the genome and the assembly sizes.

families, except for some unknown/unannotated LTRs, were similarly abundant across the three *Artocarpus* species. These differences in repeat composition suggest a species-specific expansion or excessive accumulation of LTR-Copia family repeats in *A. heterophyllus*. Among the *Artocarpus* outgroup genomes, *C. sativa* has LTR-Copia and overall repeat content expansion similar to *A. heterophyllus*.

Genome size evolution is a product of various factors, including the repetitive profile of the species. Repeat sequence accumulation can inflate the genome size of a species and shape genome evolution. To address if repeat expansions and assemblages in the genomes of Order Rosales significantly impacted their genome sizes, we correlated their total assembly sizes (i.e., a proxy for genome size) and percent of repeat content. We observed a strong positive correlation between the percent repeat content in these genomes with their genome sizes (Figure 5B, Kendall's correlation coefficient = 0.8, p-value = 8.8e-06; after PIC correction, Kendall's correlation coefficient = 0.56, p-value =0.0067). The strong correlation suggests that Order Rosales underwent genome size evolution strongly influenced by repeat expansions and accumulations.

To understand the differential species-specific repeat accumulation in the order Rosales, we used Kimura two-parameter divergence estimates to reconstruct the timeline of repeat expansion (see Figure 6). All three *Artocarpus* species have a comparable repeat abundance of ~2% genomic content at a Kimura distance of ~0.1, and this likely represents their shared history of repeat accumulation. However, *A. heterophyllus* has recently accumulated species-specific repeats corresponding to ~3.5% genomic content, primarily LTR-Copia sequences at a Kimura distance of ~0.05. Hence, the recent accumulation of the LTR-Copia is most likely the reason for genome size expansion in *A. heterophyllus* after divergence from *A. altilis* and *A. hirsutus*. Like *A. heterophyllus*, *C. sativa* also has a similar pattern of recent LTR-Copia repeat accumulation compared to other members of Cannabaceae. Rosaceae family has a rich diversity of plants with flowers and fruits with economic and commercial value. Plants of this family prove to be diverse in terms of species-specific repeat-type accumulation. For instance, *Malus* species have a recent expansion of LINE sequences.

Similarly, *F. vesca* and *P. persica* have an unusual abundance of DNA CMC repeats but have accumulated at different Kimura

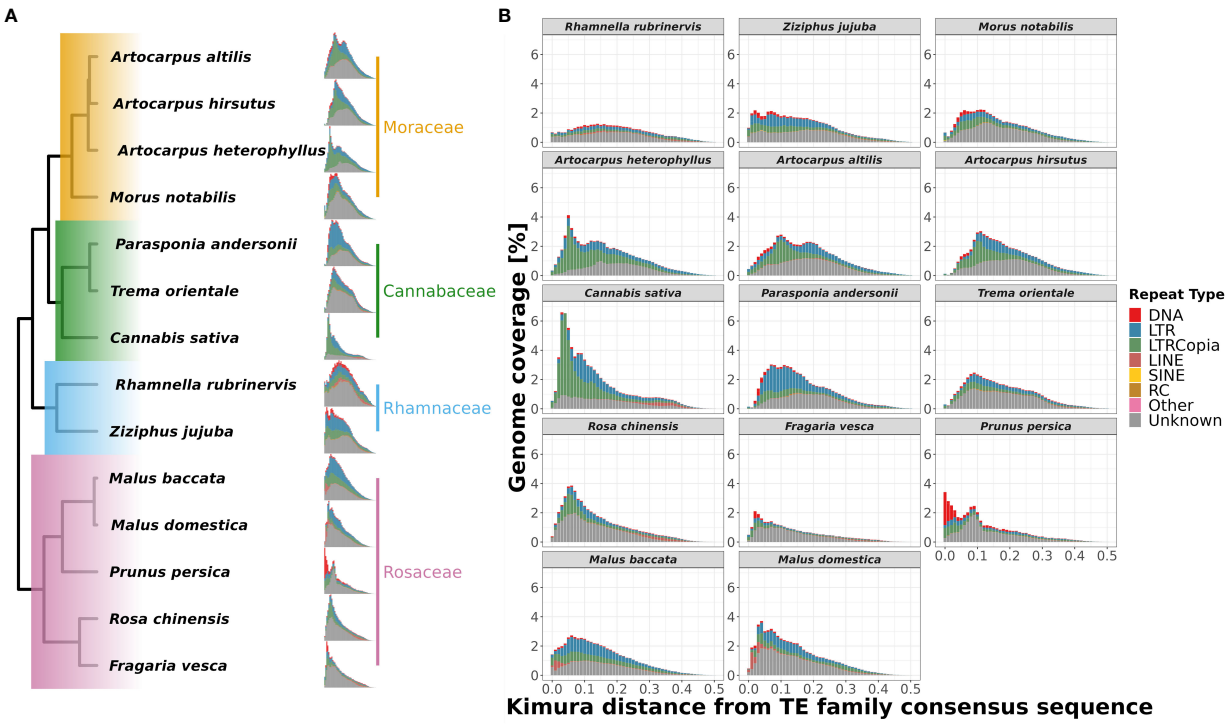


FIGURE 6
Repeat landscape evolution in 14 species of order Rosales. Family-wise changes in repeat landscapes: **(A)** The differences in abundance and insertion ages of repeat families are depicted phylogenetically. The recent spike in the fraction of the genome covered by repeats of *A. heterophyllus* suggests a recent accumulation not shared by the other two species of *Artocarpus*. **(B)** Repeat landscapes for 14 species of order Rosales showing the distribution of different repeat types and their ages of insertions depicted by Kimura distance (along the x-axis) and Genome coverage (along y-axis) percentages. *A. heterophyllus* shows a peak (3.5% of genome coverage) of LTR-Copia abundance during the recent period (~0.05 Kimura distance), which is not shared by the other two species as they have a peak (~2.5% of genome coverage) during the comparatively older period (~0.1 Kimura distance).

distances. While the repeat content in *F. vesca* has peaked at a Kimura distance of ~0.05, the accumulation of repeats in *P. persica* appears to be more recent. Further research is required to understand if this represents an ongoing insurgence of DNA CMC by comparing high-quality genomes and transcriptomes of closely related species/varieties.

Species-specific signatures of selection

To reduce false positives, we used genes identified as positively selected by the three approaches (i.e., PAML, aBSREL, and RELAX). While this approach identifies a smaller set of genes, these results are more reliable and robust to the approaches employed (see Supplementary Figure 5). We discuss the pathways with several positively selected genes in a comparative framework to understand putative species-specific adaptations (for the complete list of genes, see Supplementary Tables 9-11).

Starch and sucrose pathway

Starch and sucrose metabolism is at the heart of plant growth and development. All three *Artocarpus* species shared signatures of positive selection in genes producing (1) GBE1 (1,4 alpha-glucan branching enzyme) involved in the Starch synthesis step and (2) Cellulase/endoglucanase involved in the breakdown of cellulose (see Figure 7A). The genes coding for BGLU (Beta-Glucosidase) and EGLC (Glucan endo-1,3-beta-D-glucosidase) were positively selected in both *A. hirsutus* and *A. altilis*. These genes are involved in synthesizing D-glucose by producing multiple intermediate metabolites. However, there are multiple species-specific shifts in selection strength among the three *Artocarpus* species. For instance, *A. hirsutus* has multiple positively selected genes in different subprocesses of the starch and sucrose metabolism pathway and includes all the genes involved in the conversion of UDP-glucose to D-glucose through the production of Trehalose-6-P and alpha-Trehalose, i.e., otsA (Trehalose 6-phosphate synthase), otsB (Trehalose 6-phosphate phosphatase) and TREH (Trehalase). The BAM (Beta-amylase) gene involved in the breakdown of starch into Dextrin and Maltose through Maltodextrin was also positively selected in *A. hirsutus*. Interestingly, none of these genes had any signatures of selection in the other two *Artocarpus* species. Similarly, *A. altilis* has a species-specific positive selection in the PGM (Phosphoglucomutase) gene involved in converting D-glucose-1-phosphate to D-glucose-6-P. In conclusion, the comparative analysis of positively selected genes in this pathway highlights the differential regulation of plant developmental processes,

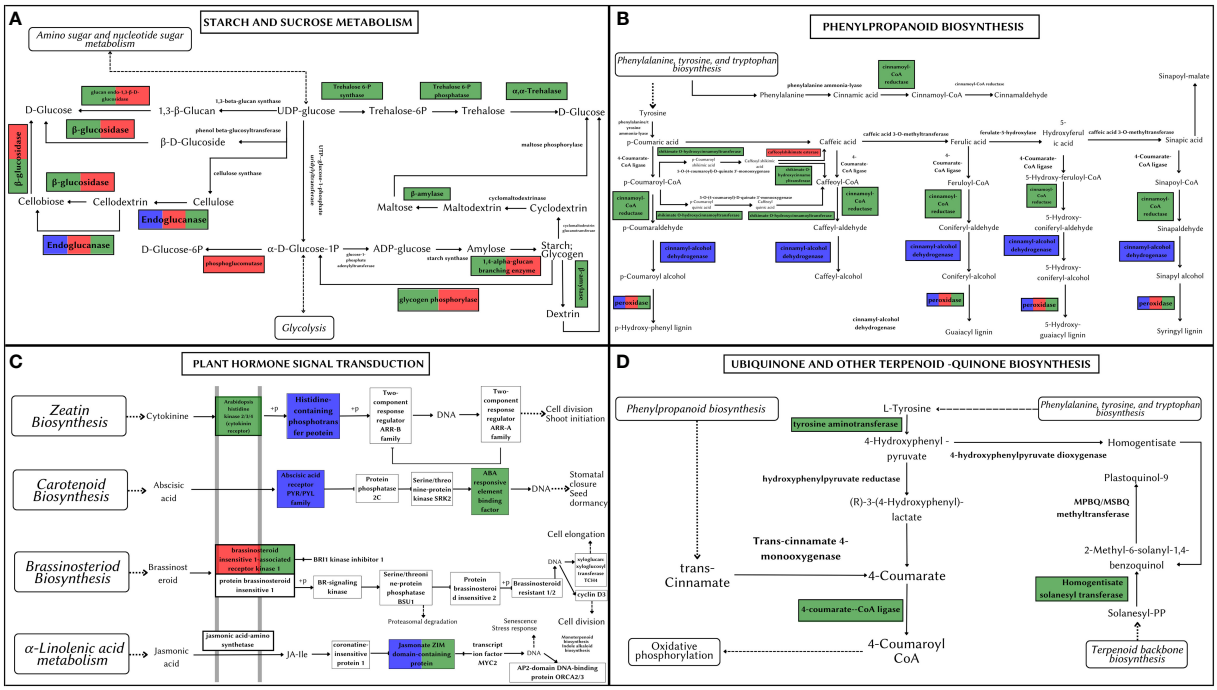


FIGURE 7
Metabolic pathways depicting positively selected genes in *Artocarpus* species. The positively selected genes in *A. hirsutus* (green), *A. heterophyllus* (blue), and *A. altilis* (red). **(A)** Starch and Sucrose Metabolism. **(B)** Phenylpropanoid biosynthesis. **(C)** Plant hormone signal transduction. **(D)** Ubiquinone and other-terpenoid-quinone biosynthesis.

especially in *A. hirsutus*, which has several positively selected genes in Trehalose synthesis and metabolism.

Phenylpropanoid biosynthesis/lignin pathway

The phenylpropanoid pathway is involved in the biosynthesis of secondary metabolites such as lignins and flavonoids using Phenylalanine, tyrosine, and tryptophan-derived compounds. All three species of *Artocarpus* have signatures of positive selection in the genes producing peroxidase (PRX/PRDX) enzyme, which catalyses the last step of lignin biosynthesis by converting lignin alcohols to lignins (see Figure 7B). Species-specific positive selection is detected in *A. hirustus* among the genes involved in the pathway's initial stages, such as 4CL (4-coumarate-coA ligase) and HCT (shikimate O-hydroxycinnamoyltransferase). Similarly, positive selection is detected in CAD (Cinnamyl-alcohol dehydrogenase) and CSE (Caffeoyl shikimate esterase) for *A. heterophyllus* and *A. altilis*, respectively. 4CL is common to both lignin and flavonoid biosynthesis pathways, while CAD, HCT, and CSE are considered lignin pathway-specific genes (Falcone Ferreyra et al., 2012; Yao et al., 2021). However, HCT is also thought to have a role in flavonoid biosynthesis (Ren et al., 2020).

Plant hormone signal transduction

Hormone signal transduction involves numerous crucial players in plant development. A transmembrane protein, BAK1 (Brassinosteroid insensitive 1-associated receptor kinase 1), is a co-receptor of BRI1 (Brassinosteroid insensitive 1) and plays a vital role in development, stress tolerance, and plant-pathogen interactions. BAK1 is positively selected in *A. hirsutus* and *A. altilis* but not in *A. heterophyllus*, suggesting differential developmental regulation in these species (see Figure 7C). Apart from BAK1, *A. hirsutus* has elevated selection pressure on another transmembrane protein HK2/3 (Histidine Kinase 2/3), the receptor for cytokinin, which is instrumental in shoot initiation and vascular bundle formation. ABF (ABA-responsive element binding factor) protein involved in regulating plant abiotic stress responses is also positively selected in *A. hirsutus*. Another important factor involved in JA (Jasmonic Acid) pathway, JAZ (Jasmonate ZIM domain-containing protein) (Pauwels and Goossens, 2011), also experiences higher selective pressure in *A. hirsutus* and *A. heterophyllus*. JA pathway is involved in almost every developmental process, including flower and root development and protection or response to multiple biotic or abiotic stress (Yang et al., 2019). Furthermore, *A. heterophyllus* also has two more genes that have elevated selection pressure, AHP (Histidine-containing phosphotransfer protein) and PYL (abscisic acid receptor PYR/PYL family) regulators of cytokinin and ABA (Abscisic acid), respectively.

Ubiquinone and terpenoid-quinone biosynthesis

Ubiquinone and other quinone-related compounds participate in multiple growth and developmental processes and act as antioxidants to provide stress tolerance (Liu and Lu, 2016). Genes involved in this pathway such as 4CL, TAT (Tyrosine aminotransferase), HST (Homogentisate solanesyltransferase), COQ6 (Ubiquinone biosynthesis monooxygenase), and NDC1 (Demethylphylloquinone reductase) all were positively selected in *A. hirsutus*, whereas neither of the two other *Artocarpus* species had any positive selection in this pathway (see Figure 7D).

Carotenoid biosynthesis

Carotenoid pathway aids in the development, stress response, and synthesis of carotenoid products (Shumskaya and Wurtzel, 2013). *A. hirsutus* has a positively selected gene PDS (15-cis-phytotene desaturase) involved in the production of carotenoids, specifically zeta-carotenoids and their derivatives. These carotenoids are yellowish and are involved in fruit ripening (Naing et al., 2019). The CYP707A gene, which is involved in catabolising ABA and its regulation, is positively selected in *A. altilis* (see Figure 8A). ABA is involved in germination and other stress responses, which suggests CYP707A may be regulating seed development processes (Kim et al., 2020).

Plant-pathogen interaction

Plant-pathogen interactions impact the survival and development of the plant. The plant can elicit pathogen-specific immune responses by assessing the nature of the pathogen. *A. hirsutus* has three positively selected genes involved in this pathway, CPK (Calcium-dependent protein kinase), CALM (Calmodulin), and BAK1 (see Figure 8B). CPK and CALM are part of fungal PAMP-triggered immunity (Pathogen-Associated Molecular Pattern) and provide fungi-specific responses. BAK1 is involved in the bacterial pathogen response of the plant. RAR1 protein (Disease resistant protein) involved in effector-triggered immunity against bacterial pathogens is positively selected in *A. heterophyllus*. The EIX receptor 1/2, which is a Pattern recognition receptor (PRR), was positively selected in both *A. heterophyllus* and *A. altilis*. Other than this PRR, *A. altilis* showed positive selection in multiple different genes involved in plant-pathogen interactions, like CNGC (Cyclic nucleotide-gated channel), a transmembrane protein providing fungal response PTI1 (pto-interacting protein 1), BAK1 and KCS (3-ketoacyl-coA synthase) involved in bacterial defence responses.

Circadian rhythm in the plant

Circadian rhythm of plants controls the molecular and cellular expression patterns to regulate better and mediate the light and dark periods, which in turn gives a fitness benefit to the plant (Venkat and Muneer, 2022). *A. heterophyllus* has two positively selected genes, PRR7 (pseudo-response regulator 7) and CSNK2A (Casein kinase II subunit alpha), which regulate

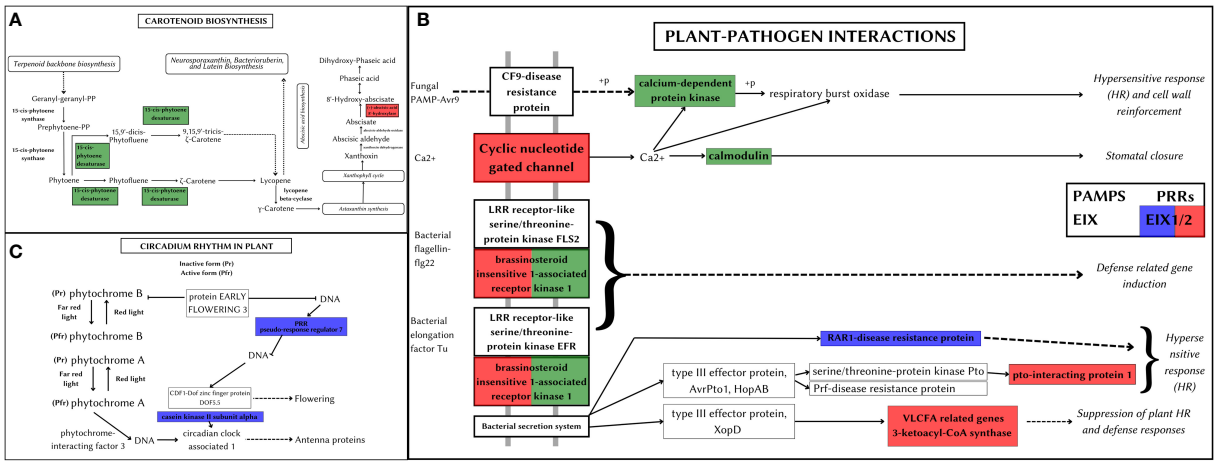


FIGURE 8
Metabolic pathways depicting positively selected genes in *Artocarpus* species. The positively selected genes in *A. hirsutus* (green), *A. heterophyllus* (blue), and *A. altilis* (red). **(A)** Carotenoid biosynthesis. **(B)** Plant-pathogen interactions. **(C)** Circadian rhythm in plants.

plant circadian rhythms (see Figure 8C). The importance of efficient control of light and dark periods could also be an evolutionary adaptation that could have led *A. heterophyllus* to distribute to a larger geographical area.

Floral gene evolution

Floral genes AS1 (Asymmetric Leaves 1) and HUA2 (Enhancer of AG-4 2) experienced strong selection pressure in both *A. hirsutus* and *A. heterophyllus*. At the same time, TIL1 (Tilted 1), TPS1 (Trehalose-6-phosphate synthase), GA2 (GA requiring 2), and MEE27 (Maternal effect embryo arrest 27) were positively selected only in *A. hirsutus*. VIP5 (Vernalisation Independence 5) and PRR7 (Psuedo-response regulator 7) experienced elevated signatures of selection only in *A. heterophyllus*. PFT1 (Phytochrome and flowering time 1), HDA9 (Histone deacetylase 9), and LNK2 (Night light-inducible and clock-regulated 2) have lineage-specific signatures of selection in *A. altilis*. Similarly, PHYE (Phytochrome E) was positively selected in both *A. altilis* and *A. heterophyllus*.

All the genes that had functions in light signaling and/or circadian rhythm regulation were under strong signatures of selection in *A. altilis*, and few of them were in *A. heterophyllus*. This might be the case because of the dependence on the tight regulation of flowering time for these plants' fitness. However, genes under strong selection pressure in *A. hirsutus* were involved in meristematic growth, fungal defense, pollen/flower development, and identity. Such processes explain the dependence on maintaining the phenotype rather than adapting to changes in light periodicity. But, *A. heterophyllus* experiences selection pressure on both circadian and other phenotype-related genes and hence might have allowed the evolution of a distinctive phenotype with a wider distribution.

Discussion

Our newly generated Wild Jack genome allowed comparative genomics of phenotypically diverse, phylogenetically distant Breadfruit species from contrasting habitats. Changes in gene content between these species reflect putative modifications in the distinct phenotypes of these plants. Notably, several genes from the same biological processes, such as flower development and response to biotic/abiotic stress, have experienced changes in copy number and may have allowed the rewiring of these pathways. Signatures of selection also occur in some genes of these same pathways leading to fine-tuning the altered phenotypes. Whole-genome sequence data allows the reconstruction of the demographic history and comparison between closely related species. Hence, such reconstructions provide helpful insight into habitat-specific responses. We complemented the information from these genomics-based reconstructions with species distribution modeling. Broadscale patterns identified by species distribution models were concordant with the trajectory inferred by genomics-based reconstructions and indicated that Jackfruit and Wild Jack species, which share the same habitat, have a comparable population size history. In contrast, the Breadfruit, which has a native range with a history of volcanic eruptions, has a unique history of population bottlenecks. While the effect of habitat on the demographic history is expected, the response of each *Artocarpus* species to the habitat is idiosyncratic.

How has the habitat shaped the genomes of *Artocarpus*

The demographic reconstruction and species distribution modelling revealed the effects of differentiated bioclimatic forces acting on their bottleneck history and distribution patterns in two diversified habitats, i.e., the Western Ghats (Jackfruit and Wild Jack) and East of Sulewasi (Breadfruit). The oceanic region of East Sulewasi islands has a history of volcanic eruptions, with accumulated ash resulting in acidic soils. These acidic soils have unique properties such as low phosphate, high iron, high Aluminium content, and other minerals or microelements. The climate and precipitation cycles are also different compared to the Western Ghats. In contrast, the Western Ghats have nutrient-rich, alkaline soils with abundant biotic meta-compositions of various taxa in the soil (Myers et al., 2000). This nutrient-rich soil contains numerous bacterial and fungal pathogens, and plants must adapt to achieve fitness in interacting or responding to these species. All these differences have impacted flora and fauna of these regions, and hence these plants must adapt to differential plant-biotic interactions. Breadfruit (*A. altilis*) has multiple gene family expansions in the OXPHOS assembly complex and chloroplast genes. We also identified lineage-specific copies of the organellar expression regulator transcription factor mTERF15 and acidic soil response transporter ABCB27. Therefore, due to harsh soil and bioclimatic properties, the mitochondrial assembly genes and their regulators in Breadfruit have experienced gene family expansion. These multiple expansions in energy-producing pathways can be explained by the higher energy demand of the plant to sustain oxidative stress response and acquire resistance to Aluminium-rich acidic soils.

Jackfruit and Wild Jack have multiple gene family changes related to plant-biotic interactions and secondary metabolite productions, which are important determinants for biotic and abiotic adaptations. The Wild Jack shows gene family expansions for IRE1, GRIP, SCPL-II, PTR1, DTX28, HSP20, MDL3, and receptor kinases, all involved in either biotic, abiotic stress tolerance/response, or plant immunity. Similarly, the Jackfruit has gene family expansion in the stress-related GST gene family and unique gene clusters of genes like the TIR domain gene involved in plant immunity and the RLP gene family involved in stress responses. In addition, both Western Ghat species have signatures of selection in anti-fungal genes such as AS1, DMS11 (Defective in meristem silencing 11), RD20 (Responsive to desiccation 20), LECRK-IX.1 (L-TYPE LECTIN RECEPTOR KINASE IX.1) conferring fungal-resistant properties to its timber.

Why such divergent phenotypes among *Artocarpus* trees

The faunal consumers prefer the fruits of Jackfruit and Wild Jack as they are sweet, fleshy, and nectary. However, the Breadfruit is a starchy fruit that is not as sweet and nectary as the other two; hence it is eaten as a vegetable rather than fruit. As discussed above, due to higher energy expenditure to sustain oxidative stress response, many other pathways with relatively more minor functions might have been impacted, reduced, or relaxed and could explain the loss of the ancestral sweet and nectary fruit phenotype in Breadfruit. Although the distribution of these plants overlaps, the two Western Ghats species, Jackfruit and Wild Jack differ in their plant height, branching, fruit size, colour, etc, and their responses to similar biotic environments may have been different due to their contrasting growth patterns. Wild Jack is a typical forest-adapted species with large trees having unidirectional growth, maintaining the apical branching to compete for sunlight efficiently, and a strong tap root system to utilize water and nutrients in dense forests. Due to this phenotype of Wild Jack, the lineage-specific gene family expansions, unique gene clusters, and genes showing selection signatures are primarily attributed to plant-pathogen interactions, stress responses, and floral evolution.

The fruits of Wild Jack are at a greater height, reducing its niche of vertebrate land dispersers such as elephants, boars, and other ruminants. The increased height ensures a different mechanism for both pollination and dispersal. The long and stalky male inflorescences of Wild Jack, in contrast to the short cylindrical inflorescences of Jackfruit, might be a switch from faunal dependence for pollination to a wind-pollinated mechanism and can explain multiple gene family expansions, unique gene clusters, and positive selection in pollen recognition genes from the lectin gene family, the Receptor kinases. Due to the switch to wind pollination, the plant must have devised some mechanisms to maintain Self Incompatibility (SI). The receptor kinases are well known to function in maintaining SI to avoid self-pollination and allow cross-pollination as much as possible (Sherman-Broyles et al., 2007). The number of fruits is more and has distinctively bright yellowish-orange colour and smaller sizes as compared to others which is an adaptation for attracting birds, bats, and primates as their dispersers (Primack, 2003; Flörchinger et al., 2010). The Lion-tailed Macaque (*Macaca silenus*), endemic to the Western Ghats, is one of the most important consumers of these fruits and can be considered their dispersers (Kumara and Santhosh, 2013). Some hornbills have also been observed eating these fruits. These pollination/disperser-specific changes in Wild Jack might be due to gene family changes and stronger positive selection in floral genes like AGL15/18, ARA1F, PTR3, EDA17, RK3, TIL1, TPS1, GA2, MEE27, AS1, HUA2, etc. In comparison, the colour of the fruit could be due to strong selection pressure on carotenoid biosynthesis genes. For example, the positively selected gene PDS is crucial for synthesizing zeta-carotene, which has a yellowish pigment. All these genomic changes have translated into the phenotype of the Wild Jack to adapt to the forest habitat and fine-tune its pollination and disperser network.

Trehalose metabolism contributes to processes involving embryogenesis and various other processes (Lunn et al., 2014).

Additionally, TPS1 regulates axillary bud outgrowth and modulation of axial shoot branching (Fichtner et al., 2021). In the Wild Jack genome, all the trehalose metabolism genes are positively selected, suggesting its importance in maintaining the phenotype of apical branching and changes in inflorescence structure. Moreover, Wild Jack has a gene family expansion in F6'H (Feruloyl-CoA 6'-hydroxylase) which catalyses the penultimate step in scopoletin synthesis, a simple coumarin. A recent study (Hoengenaert et al., 2022) demonstrated that elevated expression of Scopoletin in lignifying cells leads to higher production of monosaccharides. Due to the higher lignocellulosic mass of Wild Jack, the Trehalose pathway's involvement in generating sugars and their conduction seems likely in this plant.

In contrast to Wild Jack, Jackfruit has a short, branched tree structure with low-hanging fruits that are not suited for dense forests. The large fruits of Jackfruit are nectary and sweet with inflorescences that also impart volatile compounds, which attract a species of Gall Midge that may facilitate pollination (Gardner et al., 2018). The low-hanging Jackfruit is consumed by large mammals like elephants, wild boar, and other ruminants, facilitating its dispersal. Specifically, effective long-range dispersal is possible due to the long-distance migration of these dispersers. Therefore, the unique phenotypes of Jackfruit allow efficient fauna-based pollination/dispersal mechanisms. Gene family expansions and lineage-specific selection among genes of the flavonoid biosynthesis pathway, like Chalcone-synthase, could have facilitated the evolution of nectary fruits and inflorescences with volatile compounds. The widespread distribution of Jackfruit spans regions with differing light periodicity. Hence, the need to adapt to these changes. The strong signatures of selection in the genes involved in light signaling or circadian rhythm suggest a tight regulation of light periodicity-related pathways. Consequently, an efficient plant-pollinator/disperser network and tight regulation of circadian rhythm might have played an instrumental role in maintaining Jackfruit's wider distribution range and larger population size. Similar pollinator/disperser-influenced evolution of inflorescence has been established in closely related *Ficus* species (Zhang et al., 2020; Wang et al., 2021).

Did LTR-Copia accumulation shape Jackfruit evolution

We observed recurrent genome size changes due to repeat content dynamics in Rosales' order. We also see that the genomes of order Rosales show a strong positive correlation between their genome sizes and repeat content (Figure 5B). The increase in genome size with repeat content suggests that genome size evolution is influenced by repeat expansion. The size of the assembled *A. heterophyllus* (Jack fruit) genome (~980 Mb) is ~200Mb larger than that of *A. altilis* (~800 Mb) and *A. hirsutus* (~790 Mb). Assembled genome sizes concord with the K-mer-based estimates and is largely unaffected by assembly quality. Moreover, the genomes of *A. heterophyllus* and *A. altilis* are from a single study (Sahu et al., 2019) that uses the same methodology for sequencing and assembling both genomes, ensuring comparable genome quality. The difference in genome size among the *Artocarpus* species is primarily due to the increased prevalence (~150 Mb) of LTR-Copia in *A. heterophyllus* (*see* Figure 5A). The genome of *C. sativa* from the sister family Cannabaceae also has a larger genome, potentially due to a lineage-specific accumulation of LTR-Copia. Investigation of repeat accumulation dynamics suggests recent lineage-specific repeat expansions in these two phylogenetically distant species in a similar time frame, which suggests a role of habitat or stress-mediated induction of repeats. Repeat content change in plants has been linked to functional diversification through cis-regulatory changes or other epigenetic mechanisms (Negi et al., 2016; Hirsch and Springer, 2017). The conflict between transposable elements and the host defense mechanisms is elevated in stress conditions resulting in improved regulatory machinery (Wang et al., 2018). Hence, the accumulation of LTR-Copia in *A. heterophyllus* has played a pertinent role in the evolution of the Jackfruit genome. In future studies, gene expression data will allow the identification of ongoing transposon activity and its effect on gene regulation.

Limitations and broader implications

Gene content tends to be underestimated in fragmented genomes, and genome quality heterogeneity can confound comparative genomics. All three *Artocarpus* genomes compared in this study have similar BUSCO scores and are fairly comparable in gene content. Additionally, we put forth multiple stringent criteria to avoid false positives. We identify several candidate pathways that have experienced changes in gene content and positive selection. Detailed functional characterisation of these candidates by evaluating changes in gene expression and the consequent phenotypic changes will require further studies. The occurrence data for *Artocarpus* is limited and influenced by human-mediated dispersal, which could confound the species distribution modelling. The single genome-based demographic history reconstructions performed using the PSMC method are known to be unreliable in the recent past (<20KYA). Future studies can provide better resolution by incorporating population-level sampling.

Of the ~70 species of *Artocarpus*, our study includes only three whole genomes. Although our study highlights the potential of such comparative genomic studies, the inclusion of multiple other species would be able to provide definitive answers to questions regarding the origin, phenotypic diversity, and diversification. For instance, the genetic basis of syncarp evolution in this genus can be explored to exploit the molecular

mechanisms involved in achieving desired phenotypes. Such species-rich genera with heterogeneous phenotypes are especially well suited for agroforestry genomics (Feng and Du, 2022). Hence, *Artocarpus* can serve as a model to understand inflorescence/syncarp biology.

Conclusion

Our study has generated genomic resources for a forest tree, the Wild Jack, which is endemic to the Western Ghats. This dataset will help understand the evolution of forests and fill a gap in sampling forest tree genomes. Comparative genomic analysis with other *Artocarpus* species and members of the order Rosales has provided interesting insights into their genomic evolution. For example, habitat-driven evolution through phenotypic diversification has resulted in genomic signatures of selection and gene-family changes. Similarly, the demographic history reconstructions from genomic data and species distribution modelling strongly support the prominent role of habitat. And lastly, the adaptive changes in plant growth and development, floral morphology, and biotic interactions have shaped the Wild Jack to thrive in the forests and may explain its endemism and current fragmented distribution. In contrast, Jackfruit and Breadfruit appear tightly regulated by light signalling and circadian rhythm leading to more widespread distribution. Additionally, the fruit morphology/sizes might be due to genic evolution in floral development and may be due to the habitat-specific rewiring of the pollinator/dispersal network. Our comparative genomic analysis of *Artocarpus* con-generics exemplifies genomic changes associated with phenotypic diversity and habitat-mediated demographic changes.

Data availability statement

The data presented in the study are deposited in the ENA repository (https://www.ebi.ac.uk/ena), and the accession numbers are PRJEB55580 and ERZ12974505. Scripts and data are available at: https://github.com/Ajinkya-IISERB/Wild_Jack and https://doi.org/10.17632/vc6vwbrzs4.1.

Author contributions

AP and NV wrote the manuscript with inputs from SV and CK, BS, and SR collected the samples required for primary data generation. AP analyzed the genomic data and generated all the results. Species distribution modeling analysis was done by SV, who also prepared the illustrations used in this manuscript. All authors contributed to the article and approved the submitted version.

Funding

The Department of Biotechnology, Ministry of Science and Technology, India (Grant no. BT/11/IYBA/2018/03) and Science and Engineering Research Board (Grant no. ECR/2017/001430) provided funds used togenerate primary sequencing data published in this article and computational resources (i.e., Har Gobind Khorana Computational Biology cluster) used.

Acknowledgments

We thank the Ministry of Human Resource Development fellowship to AP. We want to thank the lab members of the PCDB lab, IISER Bhopal, for their valuable discussions. We thank Hume Centre for Ecology and Wildlife Biology for logistical support with the project.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1029540/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Leaf specimen of *A. hirsutus* (Wild Jack) sampled from the sacred groves, Kodagu, Coorg, Karnataka, India.

SUPPLEMENTARY FIGURE 2
Chloroplast genome of *A. hirsutus* and rearrangements in Genus Artocarpus. **(A)** The assembled chloroplast genome of *Artocarpus hirsutus* with a length of ~161 Kbp. **(B)** Whole plastome-based bootstrapped phylogeny of *Artocarpus* and outgroup species *Ficus religiosa* and *Morus indica*. The second panel shows Mauve alignment of variable regions of these chloroplast genomes. All *Artocarpus* genomes show a similar arrangement of inverted repeat and small single copy (SSC) regions except *A. heterophyllus* and *A. integer*. These two genomes show an inverted arrangement of sequences and genes of the SSC region.

SUPPLEMENTARY FIGURE 3
The phylogenetic relationship between gene copies of the Lectin gene family Receptor Kinase 3 (RK3). Gene copies of *Artocarpus hirsutus* are coloured in green, *A. heterophyllus* in blue, and *A. altilis* in red.

SUPPLEMENTARY FIGURE 4
Phylogenetic relationship between gene copies of ASMT and COMT gene families. The sky blue coloured cluster of genes is the COMT gene family, whereas the red coloured is ASMT. Gene copies of *Artocarpus hirsutus* are coloured in green, *A. heterophyllus* in blue, and *A. altilis* in red.

SUPPLEMENTARY FIGURE 5
Approaches used to detect positive and relaxed selection in *Artocarpus* species. Genes identified as positively selected by all the approaches are shortlisted as positively selected genes.

References

Aguilera-Alvarado, G.P., and Sanchez-Nieto, S. (2017). Plant hexokinases are multifaceted proteins. *Plant Cell Physiol.* 58, 1151–1160. doi: 10.1093/PCP/PCX062

Back, K., Tan, D. X., and Reiter, R. J. (2016). Melatonin biosynthesis in plants: multiple pathways catalyze tryptophan to melatonin in the cytoplasm or chloroplasts. *J. Pineal Res.* 61, 426–437. doi: 10.1111/JPI.12364

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in arabidopsis thaliana. *Nucleic Acids Res.* 44, D1167–D1171. doi: 10.1093/NAR/GKV1054

Brown, J. L., Hill, D. J., Dolan, A. M., Carnaval, A. C., and Haywood, A. M. (2018). PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. *Sci. Data* 5 (1), 1–9. doi: 10.1038/sdata.2018.254

Buchfink, B., Reuter, K., and Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18 (4), 366–368. doi: 10.1038/s41592-021-01101-x

Buddhisuharto, A. K., Pramastya, H., and Fidrianny, I. (2021). An updated review of phytochemical compounds and pharmacology activities of artocarpus genus. *Biointerface Res. Appl. Chem.* 11, 14898–14905. doi: 10.33263/BRIAC116.1489814905

Burch-Smith, T. M., and Dinesh-Kumar, S. P. (2007). The functions of plant TIR domains. *Sci. STKE* 2007, pe46. doi: 10.1126/STKE.4012007PE46

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-p. *Curr. Protoc. Bioinforma* 48, 4.11.1–4.11.39. doi: 10.1002/0471250953.BI0411S48

Cárdenas, P. D., Almeida, A., and Bak, S. (2019). Evolution of structural diversity of triterpenoids. *Front. Plant Sci.* 10. doi: 10.3389/FPLS.2019.01523/BIBTEX

Chandrashekara, U. M., and Sankar, S. (1998). Ecology and management of sacred groves in kerala, India. *For. Ecol. Manage.* 112, 165–177. doi: 10.1016/S0378-1127(98)00326-0

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the arabidopsis thaliana reference genome. *Plant J.* 89, 789–804. doi: 10.1111/TPJ.13415

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394. doi: 10.1101/GR.2289704

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PloS One* 5, e11147. doi: 10.1371/JOURNAL.PONE.0011147

Darriba, Di., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294. doi: 10.1093/MOLBEV/MSZ189

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18. doi: 10.1093/nar/gkw955

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/S13059-019-1832-Y/FIGURES/5

Falcone Ferreyra, M. L., Rius, S. P., and Casati, P. (2012). Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Front. Plant Sci.* 3. doi: 10.3389/FPLS.2012.00222

Feng, L., and Du, F. K. (2022). Landscape genomics in tree conservation under a changing environment. *Front. Plant Sci.* 13. doi: 10.3389/FPLS.2022.822217/BIBTEX

Fichtner, F., Barbier, F. F., Annunziata, M. G., Feil, R., Olas, J. J., Mueller-Roeber, B., et al. (2021). Regulation of shoot branching in arabidopsis by trehalose 6-phosphate. *New Phytol.* 229, 2135–2151. doi: 10.1111/NPH.17006

Flörchinger, M., Braun, J., Böhning-Gaese, K., and Schaefer, H. M. (2010). Fruit size, crop mass, and plant height explain differential fruit choice of primates and birds. *Oecologia 2010* 164 (1), 151–161. doi: 10.1007/S00442-010-1655-8

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U S A* 117, 9451–9457. doi: 10.1073/PNAS.1921046117/SUPPL_FILE/PNAS.1921046117.SAPP.PDF

Gardner, E. M., Gagné, R. J., Kendra, P. E., Montgomery, W. S., Raguso, R. A., McNeil, T. T., et al. (2018). A flower in fruit's clothing: Pollination of jackfruit (artocarpus heterophyllus, moraceae) by a new species of gall midge, clinodiplosis ultracrepidata sp. nov. (Diptera: Cecidomyiidae). *Int. J. Plant Sci.* 179, 350–367. doi: 10.1086/697115/ASSET/IMAGES/LARGE/FGA2.JPEG

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Sahromi, A. D., et al. (2021). Paralogs and off-target sequences improve phylogenetic resolution in a densely sampled study of the breadfruit genus (Artocarpus, moraceae). *Syst. Biol.* 70, 558–575. doi: 10.1093/SYSBIO/SYAA073

Gbif.Org (2022). doi: 10.15468/DL.FV8PMY. Occurrence Download.

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Gremme, G., Steinbiss, S., and Kurtz, S. (2013). Genome tools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma* 10, 645–656. doi: 10.1109/TCBB.2013.68

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/BIOINFORMATICS/BTT086

Harder, L. D., and Johnson, S. D. (2009). Darwin's beautiful contrivances: evolutionary and functional evidence for floral adaptation. *New Phytol.* 183, 530–545. doi: 10.1111/J.1469-8137.2009.02914.X

Harder, L. D., and Prusinkiewicz, P. (2013). The interplay between inflorescence development and function as the crucible of architectural diversity. *Ann. Bot.* 112, 1477–1493. doi: 10.1093/AOB/MCS252

Hijmans, R. J. (2015). DIVA-GIS, a geographic information system for the analysis of biodiversity data. Availabe at: https://diva-gis.org.

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1860, 157–165. doi: 10.1016/J.BBAGRM.2016.05.010

Hoengenaert, L., Wouters, M., Kim, H., Meester, B., Morreel, K., Vandersyppe, S., et al. (2022). Overexpression of the scopoletin biosynthetic pathway enhances lignocellulosic biomass processing. *Sci. Adv.* 8, eabo5738. doi: 10.1126/SCIADV.ABO5738

Hsu, Y. W., Wang, H. J., Hsieh, M. H., Hsieh, H. L., and Jauh, G. Y. (2014). Arabidopsis mTERF15 is required for mitochondrial nad2 intron 3 splicing and functional complex I activity. *PloS One* 9, e112360. doi: 10.1371/JOURNAL.PONE.0112360

Jagtap, U. B., and Bapat, V. A. (2010). Artocarpus: A review of its traditional uses, phytochemistry and pharmacology. *J. Ethnopharmacol.* 129, 142–166. doi: 10.1016/J.JEP.2010.03.031

Jamieson, P. A., Shan, L., and He, P. (2018). Plant cell surface molecular cypher: Receptor-like proteins and their roles in immunity and development. *Plant Sci.* 274, 242. doi: 10.1016/J.PLANTSCI.2018.05.030

Jarrett, F. (1977). The syncarp of artocarpus. a unique biological phenomenon. *Gard* 29, 35–39.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and De Bakker, P. I. W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938–2939. doi: 10.1093/BIOINFORMATICS/BTN564

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27. doi: 10.1093/NAR/28.1.27

Kar, D., Pradhan, A. A., and Datta, S. (2021). The role of solute transporters in aluminum toxicity and tolerance. *Physiol. Plant* 171, 638–652. doi: 10.1111/PPL.13214

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/NAR/GKF436

Kim, H. M., Park, S. H., Ma, S. H., Park, S. Y., Yun, C. H., Jang, G., et al. (2020). Promoted ABA hydroxylation by capsicum annuum CYP707As overexpression suppresses pollen maturation in nicotiana tabacum. *Front. Plant Sci.* 11. doi: 10.3389/FPLS.2020.583767/BIBTEX

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/BIOINFORMATICS/BTZ305

Kumara, H. N., and Santhosh, K. (2013). Development of conservation strategy for a newly discovered lion-tailed macaque macaca silenus population in sirsi-honnavara, Western ghats: II. understanding the impact of NTFP collection on lion-tailed macaque. *Sacon Tec Rep.* 116, 1–48.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/MOLBEV/MSX116

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048. doi: 10.1101/GR.174302

Li, H. (2013) *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. Available at: http://arxiv.org/abs/1303.3997 (Accessed May 1, 2019).

Li, H., and Barrett, J. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/BIOINFORMATICS/BTR509

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, Z. W., Hou, X. H., Chen, J. F., Xu, Y. C., Wu, Q., Gonzalez, J., et al. (2018). Transposable elements contribute to the adaptation of arabidopsis thaliana. *Genome Biol. Evol.* 10, 2140–2150. doi: 10.1093/GBE/EVY171

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/GR.1224503

Liu, C. N., Li, Y. Y., Wang, R., and Chen, X. Y. (2020). Genetic factors are less considered than demographic characters in delisting species. *Biol. Conserv.* 251, 108791. doi: 10.1016/J.BIOCON.2020.108791

Liu, M., and Lu, S. (2016). Plastoquinone and ubiquinone in plants: Biosynthesis, physiological function and metabolic engineering. *Front. Plant Sci.* 7. doi: 10.3389/FPLS.2016.01898

Lunn, J. E., Delorge, I., Figueroa, C. M., Van Dijck, P., and Stitt, M. (2014). Trehalose metabolism in plants. *Plant J.* 79, 544–567. doi: 10.1111/TPJ.12509

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10. doi: 10.14806/ej.17.1.200

Matthew, S. P., Mohandas, A., Shareef, S. M., and Nair, G. M. (2006). Biocultural diversity of the endemic 'Wild jack tree' on the malabar coast of south India. *Ethnobot Res. Appl.* 4, 025–040.

Mazet, O., Rodríguez, W., and Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor. Popul Biol.* 104, 46–58. doi: 10.1016/j.tpb.2015.06.003

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity (Edinb)* 116, 362–371. doi: 10.1038/hdy.2015.104

Meenu, M. T., Kaul, G., Shukla, M., Radhakrishnan, K. V., and Chopra, S. (2021). Cudraflavone c from artocarpus hirsutus as a promising inhibitor of pathogenic, multidrug-resistant s. aureus, persisters, and biofilms: A new insight into a rational explanation of traditional wisdom. *J. Nat. Prod.* 84, 2700–2708. doi: 10.1021/ACS.JNATPROD.1C00578/ASSET/IMAGES/LARGE/NP1C00578_0007.JPEG

Meera, M., Ruckmani, A., Saravanan, R., and Lakshmipathy prabhu, R. (2018). Anti-inflammatory effect of ethanolic extract of spine, skin and rind of jack fruit peel - a comparative study. *Nat. Prod. Res.* 32, 2740–2744. doi: 10.1080/14786419.2017.1378200

Mehrotra, R. C., Prakash, U., and Bande, M. B. (1984). Fossil woods of lophopetalum and artocarpus from the deccan intertrappean beds of mandla district, Madhya pradesh, India. *J. Palaeosciences* 32, 310–320. doi: 10.54991/jop.1984.1385

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2021). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36, 5516–5518. doi: 10.1093/BIOINFORMATICS/BTAA1022

Merow, C., Smith, M. J., and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography (Cop.)* 36, 1058–1069. doi: 10.1111/J.1600-0587.2013.07872.X

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/MOLBEV/MSAA015

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., et al. (2014). ENMeval: An r package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models. *Methods Ecol. Evol.* 5, 1198–1205. doi: 10.1111/2041-210X.12261

Myers, N., Mittermeler, R. A., Mittermeler, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nat* 403 (6772), 853–858. doi: 10.1038/35002501

Nadachowska-Brzyska, K., Li, C., Smeds, L., Zhang, G., and Ellegren, H. (2015). Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. *Curr. Biol.* 25, 1375–1380. doi: 10.1016/j.cub.2015.03.047

Naing, A. H., Kyu, S. Y., Pe, P. P. W., Park, K.II, Lee, J. M., Lim, K. B., et al. (2019). Silencing of the phytoene desaturase (PDS) gene affects the expression of fruit-ripening genes in tomatoes. *Plant Methods* 15, 1–10. doi: 10.1186/S13007-019-0491-Z/FIGURES/5

Negi, P., Rai, A. N., and Suprasanna, P. (2016). Moving through the stressed genome: Emerging regulatory roles for transposons in plant stress response. *Front. Plant Sci.* 7. doi: 10.3389/FPLS.2016.01448

Osuri, A. M., Madhusudan, M. D., Kumar, V. S., Chengappa, S. K., Kushalappa, C. G., and Sankaran, M. (2014). Spatio-temporal variation in forest cover and biomass across sacred groves in a human-modified landscape of india's Western ghats. *Biol. Conserv.* 178, 193–199. doi: 10.1016/J.BIOCON.2014.08.008

Patil, A. B., and Vijay, N. (2021). Repetitive genomic regions and the inference of demographic history. *Heredity (Edinb)* 127, 151–166. doi: 10.1038/s41437-021-00443-8

Pauwels, L., and Goossens, A. (2011). The JAZ proteins: A crucial interface in the jasmonate signaling cascade. *Plant Cell* 23, 3089. doi: 10.1105/TPC.111.089300

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/MOLBEV/MSP077

Primack, R. B. (1987). Relationships among flowers, fruits, and seeds. *Annual review of ecology and systematics* 18, 409–430. doi: 10.1146/ANNUREV.ES.18.110187.002205

Ragone, D. (2018). Breadfruit–artocarpus altilis (Parkinson) fosberg. *Exot Fruits*, 53–60. doi: 10.1016/B978-0-12-803138-4.00009-5

Ramakrishnan, M., Satish, L., Kalendar, R., Narayanan, M., Kandasamy, S., Sharma, A., et al. (2021). The dynamism of transposon methylation for plant development and stress adaptation. *Int. J. Mol. Sci.* 22, 11387. doi: 10.3390/IJMS222111387

Ren, C., Wang, J., Xian, B., Tang, X., Liu, X., Hu, X., et al. (2020). Transcriptome analysis of flavonoid biosynthesis in safflower flowers grown under different light intensities. *PeerJ* 8, e8671. doi: 10.7717/peerj.8671

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., et al. (2003). The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic Acids Res.* 31, 224–228. doi: 10.1093/NAR/GKG076

Rothery, P. (1979). A nonparametric measure of intraclass correlation. *Biometrika* 66, 629–639. doi: 10.1093/BIOMET/66.3.629

Rui, Q., Tan, X., Liu, F., Li, Y., Liu, X., Li, B., et al. (2021). Syntaxin of plants31 (SYP31) and SYP32 is essential for golgi morphology maintenance and pollen development. *Plant Physiol.* 186, 330–343. doi: 10.1093/PLPHYS/KIAB049

Sahu, S., Liu, M., Yssel, A., Kariba, R., Muthemba, S., Jiang, S., et al. (2019). Draft genomes of two artocarpus plants, jackfruit (A. heterophyllus) and breadfruit (A. altilis). *Genes (Basel)* 11, 27. doi: 10.3390/genes11010027

Saidou, M., and Zhang, Z. (2022). The l-type lectin-like receptor kinase gene TaLecRK-IV.1 regulates the plant height in wheat. *Int. J. Mol. Sci.* 23, 8208. doi: 10.3390/IJMS23158208

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/BIOINFORMATICS/19.2.301

Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics* 204, 1207–1223. doi: 10.1534/GENETICS.116.190223/-/DC1

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14. doi: 10.1093/nar/gkv318

Sherman-Broyles, S., Boggs, N., Farkas, A., Liu, P., Vrebalov, J., Nasrallah, M. E., et al. (2007). S locus genes and the evolution of self-fertility in arabidopsis thaliana. *Plant Cell* 19, 94–106. doi: 10.1105/TPC.106.048199

Shumskaya, M., and Wurtzel, E. T. (2013). The carotenoid biosynthetic pathway: Thinking in all dimensions. *Plant Sci.* 208, 58. doi: 10.1016/J.PLANTSCI.2013.03.012

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/BIOINFORMATICS/BTV351

Smit, A. F. A., Hubley, R., and Green, P. RepeatMasker Open-4.0.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353.

Solanki, S., Bhardwaj, R., Vasudeva, R., Chourey, S., and Archak, S. (2020). Biochemical composition of pulp and seed of wild jack (Artocarpus hirsutus lam.) fruit. *Plant Foods Hum. Nutr.* 75, 659–660. doi: 10.1007/S11130-020-00849-5/FIGURES/1

Srivastava, R. (1998). Fossil wood of artocarpus from warkalli formation of kerala coast, India. *Phytomorphology* 48, 391–397.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/NAR/GKI458

Sun, Y., Qiao, Z., Muchero, W., and Chen, J. G. (2020). Lectin receptor-like kinases: The sensor and mediator at the plant cell surface. *Front. Plant Sci.* 11. doi: 10.3389/FPLS.2020.596301/BIBTEX

Tambat, B., Rajanikanth, G., Ravikanth, G., Shaanker, R. U., Ganeshaiah, K. N., and Kushalappa, C. G. (2005). Seedling mortality in two vulnerable tree species in the sacred groves of Western ghats, south India. *Curr. Sci.* 88, 350–352.

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annual review of plant biology* 65, 225–257. doi: 10.1146/ANNUREV-ARPLANT-050312-120229

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/NAR/GKX391

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, populus trichocarpa (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Upadhyay, N., Kar, D., and Datta, S. (2020). A multidrug and toxic compound extrusion (MATE) transporter modulates auxin levels in root to regulate root development and promotes aluminium tolerance. *Plant Cell Environ.* 43, 745–759. doi: 10.1111/PCE.13658

Van Der Lee, R., Wiel, L., Van Dam, T. J. P., and Huynen, M. A. (2017). Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* 45, 10634–10648. doi: 10.1093/nar/gkx704

Venkat, A., and Muneer, S. (2022). Role of circadian rhythms in major plant metabolic and signaling pathways. *Front. Plant Sci.* 13. doi: 10.3389/FPLS.2022.836244/BIBTEX

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wang, Y., Chantreau, M., Sibout, R., and Hawkins, S. (2013). Plant cell wall lignification and monolignol metabolism. *Front. Plant Sci.* 4. doi: 10.3389/FPLS.2013.00220/BIBTEX

Wang, Y., Liang, W., and Tang, T. (2018). Constant conflict between gypsy LTR retrotransposons and CHH methylation within a stress-adapted mangrove genome. *New Phytol.* 220, 922–935. doi: 10.1111/NPH.15209

Wang, R., Yang, Y., Jing, Y., Segar, S. T., Zhang, Y., Wang, G., et al. (2021). Molecular mechanisms of mutualistic and antagonistic interactions in a plant–pollinator association. *Nat. Ecol. Evol.* 5 (7), 974–986. doi: 10.1038/s41559-021-01469-1

Wang, R., Zhang, X., Shi, Y.-S., Li, Y.-Y., Wu, J., He, F., et al. (2020). Habitat fragmentation changes top-down and bottom-up controls of food webs. *Ecology* 101, e03062. doi: 10.1002/ECY.3062

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., and Scheffler, K. (2015). RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* 32, 820–832.

Williams, E. W., Gardner, E. M., Harris, R., Chaveerach, A., Pereira, J. T., and Zerega, N. J. C. (2017). Out of Borneo: biogeography, phylogeny and divergence date estimates of artocarpus (Moraceae). *Ann. Bot.* 119, 611–627. doi: 10.1093/AOB/MCW249

Xavier, T. F., Kannan, M., Lija, L., Auxillia, A., Rose, A. K. F., and Kumar, S. S. (2014). Ethnobotanical study of kani tribes in thoduhills of kerala, south India. *J. Ethnopharmacol.* 152, 78–90. doi: 10.1016/J.JEP.2013.12.016

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., et al. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. doi: 10.1093/NAR/GKZ333

Xu, Z., and Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/MOLBEV/MSM088

Yang, J., Duan, G., Li, C., Liu, L., Han, G., Zhang, Y., et al. (2019). The crosstalks between jasmonic acid and other plant hormone signaling highlight the involvement of jasmonic acid as a core component in plant response to biotic and abiotic stresses. *Front. Plant Sci.* 10. doi: 10.3389/FPLS.2019.01349/BIBTEX

Yao, T., Feng, K., Xie, M., Barros, J., Tschaplinski, T. J., Tuskan, G. A., et al. (2021). Phylogenetic occurrence of the phenylpropanoid pathway and lignin biosynthesis in plants. *Front. Plant Sci.* 12. doi: 10.3389/FPLS.2021.704697/BIBTEX

Zerega, N. J. C., Supardi, M. N. N., and Motley, T. J. (2010). Phylogeny and recircumscription of artocarpeae (Moraceae) with a focus on artocarpus. *Syst. Bot.* 35, 766–782. doi: 10.1600/036364410X539853

Zhang, X., Wang, G., Zhang, S., Chen, S., Wang, Y., Wen, P., et al. (2020). Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell* 183, 875–889.e17. doi: 10.1016/J.CELL.2020.09.043

Zhao, D., Yu, Y., Shen, Y., Liu, Q., Zhao, Z., Sharma, R., et al. (2019). Melatonin synthesis and function: Evolutionary history in animals and plants. *Front. Endocrinol. (Lausanne)* 10. doi: 10.3389/FENDO.2019.00249/BIBTEX

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/BIOINFORMATICS/BTT476

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 December 2022
DOI 10.3389/fpls.2022.990927

OPEN ACCESS

EDITED BY
Fang Du,
Beijing Forestry University, China

REVIEWED BY
Filipa Monteiro,
University of Lisbon, Portugal
Wataru Ishizuka,
Hokkaido Research
Organization, Japan

*CORRESPONDENCE
Keiko Kitamura
kitamq@ffpri.affrc.go.jp

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 11 July 2022
ACCEPTED 16 November 2022
PUBLISHED 15 December 2022

CITATION
Kitamura K, Namikawa K, Tsuda Y, Kobayashi M and Matsui T (2022) Possible northern persistence of Siebold's beech, *Fagus crenata,* at its northernmost distribution limit on an island in Japan Sea: Okushiri Island, Hokkaido.
Front. Plant Sci. 13:990927.
doi: 10.3389/fpls.2022.990927

Possible northern persistence of Siebold's beech, *Fagus crenata,* at its northernmost distribution limit on an island in Japan Sea: Okushiri Island, Hokkaido

Keiko Kitamura[1*†], Kanji Namikawa[2], Yoshiaki Tsuda[3†], Makoto Kobayashi[4] and Tetsuya Matsui[5,6]

[1]Hokkaido Research Centre, Forestry and Forest Products Research Institute, Sapporo, Japan, [2]Biological Laboratory, Hokkaido University of Education, Sapporo, Japan, [3]Sugadaira Montane Research Center, University of Tsukuba, Ueda, Japan, [4]Department of Education and Culture, Echigo-Matsunoyama Museum of Natural Science, Tokamachi, Japan, [5]Center of Biodiversity and Climate Change, Forestry and Forest Products Research Institute, Tsukuba, Japan, [6]Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

Siebold's beech, *Fagus crenata*, is widely distributed across the Japanese Archipelago and islands in Japan Sea. Similar to the northern limit of the geographical distribution of *F. crenata* on the mainland of Hokkaido, the northern limit of the distribution of *F. crenata* on islands in the Japan Sea is observed on Okushiri Island (ca 42°*N*). To understand the genetic relationships of *F. crenata* on Okushiri Island, we examined chloroplast (cp) DNA haplotypes and 11 nuclear microsatellite (SSR) loci among 1,838 individuals from 44 populations from Okushiri Island, mainland Hokkaido, and the northern part of the Tohoku region on Honshu Island. We identified 2 cpDNA haplotypes, which represent not only populations on the Japan Sea coast but also those on the Pacific coast and this suggested the Okushiri Island populations might not be formed by single colonization. Genetic diversity of the Okushiri Island populations of nuclear SSR was not lower than the mainland and the STRUCTURE analysis revealed the Okushiri Island individuals were admixed between Hokkaido and Tohoku clusters. Approximate Bayesian computation inferred that divergence between Tohoku and Hokkaido, and admixture between two populations which generated Okushiri populations occurred before the last glacial maximum (LGM), that is, 7,890 (95% hyper probability density (HPD): 3,420 – 9,910) and 3,870 (95% HPD: 431– 8,540) generations ago, respectively. These inferences were well supported by a geological history which suggested an isolation of Okushiri Island from Hokkaido started prior to the Middle Pleistocene. We discuss the possible persistence of *F. crenata* during the last glacial maximum on northern islands in the Japan Sea such as Okushiri Island.

KEYWORDS

the northernmost geographic range, chloroplast DNA, nuclear SSR, refugia, structure

Introduction

Plant populations on islands differ from mainland populations in terms of ecological and demographic processes such as colonization events, population persistence, and expansion because of their geographical isolation. From a genetic perspective, island populations are assumed to have less diversity than their mainland counterparts due to evolutionary processes such as founder effects, limited gene dispersal, and small population size (Barrett, 1996; Frankham, 1996; Frankham, 1997; Takayama et al., 2013; Stuessy et al., 2014; Takayama et al., 2015; Stuessy, 2020). For example, the initial colonizing event inevitably involves a genetic bottleneck which determines the loss of gene diversity in founder populations (Barrett, 1996). Therefore, newly founded populations carry less gene diversity than source populations; however, repeated gene flow from source populations can compensate for a lack of gene diversity and contribute to long-term population persistence and expansion (Alsos et al., 2015). Theoretically, long-distance gene flow is less frequent than short-distance gene flow, so that spatial isolation and small population size may lead to an erosion of genetic variation and increased interpopulation differentiation (Young et al., 1996). In general, geographical isolation is a major contributor to shallow gene diversity (DeJoode and Wendel, 1992; Chiang et al., 2006; Takayama et al., 2013; Alsos et al., 2015).

On the contrary, several studies of terrestrial plant species revealed that island populations have substantial genetic diversity in spite of their isolated geographic locations which might restrict frequent gene flow (Chiang et al., 2006; Rosas Escobar et al., 2011; Désamoré et al., 2012; García-Verdugo et al., 2013; García-Verdugo et al., 2015). This is thought to be due to the fact that the genetic diversity of island populations is not only related to founder effects but also by different geographic and biological conditions and contexts (Stuessy et al., 2014), such as adaptation and maladaptation to the new environment (St Clair and Howe, 2007; Kuparinen et al., 2010) and chronic changes of gene immigration (Elleouet and Aitken, 2019; Kitamura and Nakanishi, 2021).

Another interesting issue regarding island populations is the colonization history during the species range expansion (Taberlet et al., 1998; Magri et al., 2006; Alsos et al., 2015). Compared to mainland populations, the routes and frequency of dispersal to islands are restricted due to geographic isolation, and long-distance gene dispersal represents the initial stage in the colonization process.

Fagus crenata Blume (2n = 24) is monoecious and allogamous, with barochorous and synzoochorous by rodents and birds (Miguchi, 1996). The geographic range of *F. crenata* extends across the Japanese Archipelago, from southern Hokkaido to Kyushu (Peters, 1997), with sharp geographic clines in genetic diversity (Tomaru et al., 1997; Hiraoka and Tomaru, 2009) and morphology (Hagiwara, 1977; Ishii et al., 2018) observed. Our previous study of genetic structure at the northernmost distribution range on mainland Hokkaido discovered a decline in gene diversity toward the northern edge of populations (Kitamura et al., 2015). In addition to the mainland distribution, natural populations of *F. crenata* are observed on off-shore islands in the Japan Sea, with Okushiri Island (ca. 42°N) representing the northernmost distribution of *F. crenata* among such islands (Tatewaki, 1948). The island is reported to have been separated from the mainland in the Middle Pleistocene and never became reattached to the mainland thereafter (Ohshima, 1980). The northward expansion of *F. crenata* after the Last Glacial Maximum (LGM) started about 6,000 years BP on the mainland (Kito and Takimoto, 1999); however, the origins of the northernmost island population of *F. crenata* is still unknown. Long-distance vs. short-distance gene dispersals are driving forces to develop genetic diversity within populations. As tracers of gene flow, neutral molecular markers provide useful information, although care needs to be taken in separating historical connectivity from current gene flow. A combination of maternally inherited chloroplast (cp) genome for angiosperms and bi-parental inherited nuclear genome provides relevant information such as the colonization history because comparisons of uniparentally (cp) vs. biparentally (nuclear) inherited loci can reveal differences in dispersal between seeds and/or pollen (Petit and Excoffier, 2009). Moreover, recent advances in genetic-based population demographic inference make it possible to investigate a colonization history of species from different ancestral lineages, considering time scales (Tsuda et al., 2015).

The aims of this study are: 1) to evaluate the genetic structure of *F. crenata* on Okushiri Island by 44 populations from northern distribution using nuclear and cpDNA variation, 2) to infer the species' population demography on the island, and 3) to discuss the colonization and northern expansion dynamics of *F. crenata*, considering a possibility of northern persistence of the species during the LGM.

Materials and methods

Study sites

Okushiri Island is located 16 km west of the mainland of Hokkaido in the island-mainland northernmost distribution area of *F. crenata* (Figure 1). This Island is a small island of 143 square km, where forest covers 80% of the island, yet the *F. crenata* forest covers approximately 60% of the forested area (Namikawa et al., 2021), and its stand volume is around 60 cubic meters per hectare (Tatewaki, 1948).

In terms of the vegetation classification, the beech forests on Okushiri Island showed diverse floristic features from subarctic to warm temperate elements which is partly due to the warm current (Namikawa et al., 2021). The major vegetation type of

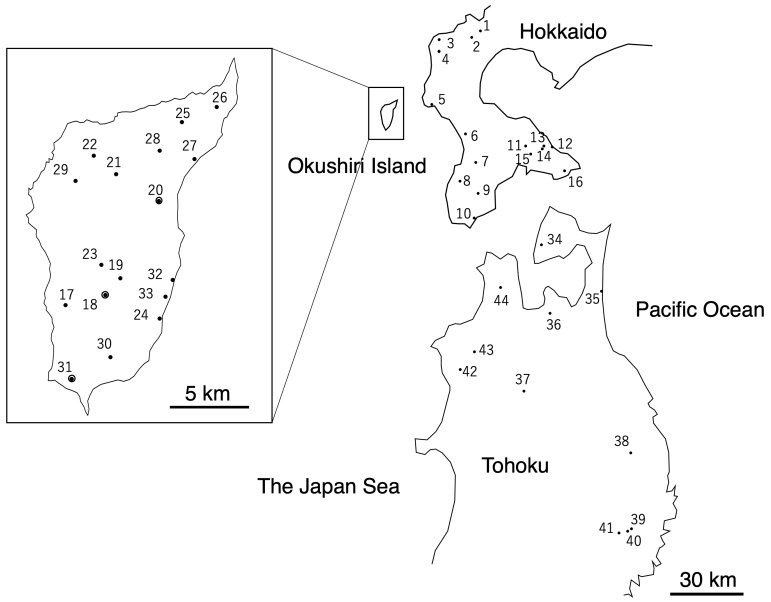


FIGURE 1
Locations of *Fagus crenata* populations examined in this study (Supplementary Table 1) Populations 18, 20, and 31 were referred in the Discussion.

beech forest on the island includes species that frequently appear in beech forests on the Japan Sea coast of Hokkaido and northern Tohoku. Another vegetation type found on Okushiri Island is associated with the southern part of the Pacific coast of Hokkaido, which has a short snow-cover period. Thus, despite the small size of the island, regional differences in species composition seen in mainland Hokkaido and Tohoku were condensed in the beech forests on Okushiri Island (Namikawa et al., 2021).

We selected 17 populations from across the *F. crenata* forests on the island, 16 populations from the mainland of Hokkaido, and 11 from the northern Tohoku region of Honshu as study sites, covering most of the species' distribution in these areas (Figure 1, Supplementary Table 1). To cover most of the natural distribution, study populations on Okushiri Island were selected from both east- and west-facing slopes of the island, as well as coastal and inland habitats. Populations from mainland Hokkaido and the northern Tohoku region were selected across the contrasting climate conditions of the Pacific coast that shows the dry and low temperature winter, and the Japan Sea coast, which experiences heavy snow fall in winter.

DNA extraction, chloroplast DNA SNPs and microsatellite genotyping

Mature trees of diameter at breast height (DBH) > 20 cm (11 to 65 trees per population) were chosen at random, and leaves were collected from a total of 1,838 individuals from 44 populations (Supplementary Table 1). Collected leaves were kept in a cooling box with ice until they could be stored in the refrigerator or freezer in the laboratory. Total DNA was extracted with the DNeasy Plant Mini Kit (Qiagen K. K., Tokyo, Japan) from 50 mg of leaf sample that was ground into a powder using liquid nitrogen and a Multi-Beads Shocker cell disruptor (Yasui Kikai Co. Ltd., Osaka, Japan).

It has been revealed that there are two chloroplast (cp)DNA haplotypes in these regions, that is, A (GenBank: AB046492.1) and B (GenBank: AB046493.1) (Fujii et al., 2002). We determined the cpDNA haplotypes of each population by SNPs according to the method described by Katai et al. (2014); Takahashi et al. (2022). The first polymerase chain reaction (PCR) was performed to amplify *trn*K fragment by 0.2 μM final concentration of each forward (5'-TTATTCTTAGCGGATCGGTCCA-3') and reverse (5'-CCGTGCTTGCATCTTTCATTG-3') primer using Multiplex PCR Kit (Qiagen K. K., Tokyo, Japan) containing 10 to 30 ng of genomic DNA in a final volume of 6 μl. Amplification condition was 95°C for 15 min., 35 cycles of [94°C for 30 sec., 56°C for 90 sec., 72°C for 2 min.], 72°C for 10 min, then hold at 4°C. Amplicons were treated with 0.8 U of *Exo*I (Takara Bio Inc., Shiga, Japan) and 2 U of Shrimp Alkaline Phosphatase (SAP, Takara Bio Inc., Shiga, Japan) with *Exo*I buffer (Takara Bio Inc., Shiga, Japan) to remove unincorporated primers and dNTPs for 1 hour at 35°C followed by 15 min. at 75°C for inactivation of enzyme, then hold at 4°C. The single-base primer extension was performed using SNaPshot Multiplex Kit (Thermo Fisher Scientific K. K., Tokyo, Japan) and *trn*K1754 primer (5'-T^{14}CTAGCATTTGACTCCGCACCACTGAAG -3'). The total volume of the reaction mix was 7 μl which contained 1 μl of

SNaPshot Master Mix, 0.7 μl of 2 μM primer, 1 μl of BigDye Terminator 5 x Sequencing buffer (Thermo Fisher Scientific K. K., Tokyo, Japan), and 2 μl of the first PCR product. The extension was performed under condition of 25 cycles of [96°C for 10 sec., 50°C for 5 sec., 60°C for 30 sec.], then hold at 4°C. SNaPshot extension reaction was treated with 1 U of SAP (Takara Bio Inc., Shiga, Japan) to remove unincorporated primer and ddNTPs for 1 hour at 35°C followed by 15 min. at 75°C for inactivation of enzyme, then hold at 4°C. The fragments were detected with ABI PRISM 3130*xl* Genetic Analyzer (Thermo Fisher Scientific K. K., Tokyo, Japan) using POP-7 polymer and 36 cm capillary. An aliquot of SNaPshot reaction (1 μl) was mixed with 9.8 μl of Hi-Di Formamide (Thermo Fisher Scientific K. K., Tokyo, Japan) and 0.2 μl of GeneScan 120 LIZ dye Size Standard (Thermo Fisher Scientific K. K., Tokyo, Japan). SNPs were determined by GENESCAN for Windows (Thermo Fisher Scientific K. K., Tokyo, Japan). We first analyzed four individuals from each population to determine cpDNA haplotypes. When four individuals from the same population showed the same haplotypes, the cpDNA haplotype of the population was determined to be monomorphic. When different haplotypes were detected from four individuals from the same population, we further analyzed 12 to 22 individuals from that population to determine the cpDNA haplotype ratio of the population. The cpDNA haplotype are fixed in most of the populations of this species, and we analyzed the limited number of individuals to detect at least 6% of polymorphism.

Eleven nuclear microsatellite (SSR) primer pairs were employed to identify genotypes of all individuals for the following loci: mfc2, mfc12 (Tanaka et al., 1999), FS1-03, FS4-46 (Pastorelli et al., 2003), sfc7, sfc18, sfc36, sfc378, sfc1063, sfc1105, and sfc1143 (Asuka et al., 2004), according to the method described in (Kitamura et al., 2015). PCR was performed with the Multiplex PCR Kit (Qiagen K. K., Tokyo, Japan) containing 10 to 30 ng of genomic DNA in a final volume of 10 μl. Amplification condition was 95°C for 15 min., 35 cycles of [94°C for 30 sec., 57°C for 90 sec., 72°C for 1 min.], 60°C for 30 min, then hold at 15°C. PCR products (1 μl) was mixed with 10 μl Hi-Di Formamide (Thermo Fisher Scientific K. K., Tokyo, Japan) and 0.15 μl of GeneScan 600 LIZ dye Size Standard v2.0 (Thermo Fisher Scientific K. K., Tokyo, Japan) for estimating DNA fragment sizes. The length of amplified fragments was analyzed using the ABI PRISM 3130*xl* Genetic Analyzer using POP-7 polymer, 36 cm capillary, and GENESCAN for Windows (Thermo Fisher Scientific K. K., Tokyo, Japan).

The number of individuals used for analyses are shown in Supplementary Table 1.

Data analyses

The genetic diversity parameters within each population were evaluated by determining the expected heterozygosity (H_E) (Nei 1987), allelic richness (R_S) (El Mousadik and Petit, 1996), and the fixation index (F_{IS}) (Wright 1965) using the FSTAT 2.9.3 computer program (hereafter, FSTAT) (Goudet, 2001). Significant deviations in F_{IS} values from 0 were estimated by 95% confidence intervals (CI) obtained through 999 permutations of bootstrapping using GENODIVE (Meirmans and Van Tienderen, 2004). The degree of genetic differentiation among the populations was evaluated by Nei's G'_{ST}. The number of alleles and the effective number of alleles within a region were calculated by GENODIVE. Frequency of null allele were calculated by CERVUS (Marshall et al., 1998).

To determine the coancestry composition among populations and admixtures, we employed multilocus model-based cluster analysis using STRUCTURE 2.3.4 (hereafter, STRUCTURE) (Pritchard et al., 2000). All of the runs consisted of 30,000 Markov chain Monte Carlo (MCMC) generations, after a burn-in period of 70,000 iterations. We confirmed that these settings for the length of the MCMC generations and burn-in period are sufficient to obtain valid genetic structures by comparing the results from initial runs with those involving longer MCMC generations. This analysis was based on the LOCPRIOR model described by Hubisz et al. (2009), an admixture model, and the correlated allele frequencies model (hereafter, the F-model) described by Falush et al. (2003). Ten runs were performed for each value of K, ranging from 1 to 10; as we confirmed that the genetic structure among population could be detected at $K < 10$ in the pilot run. CLUMPAK server (Kopelman et al., 2015) was employed to evaluate the probability of the data [LnP (D)] for each K and to calculate ΔK according to the method described by Evanno et al. (2005), and to evaluate multimodality among runs and major clustering patterns at each K.

In addition, principal coordinate analysis (PCoA) of 44 populations based on the allele frequencies of 11 SSR loci, as well as Mantel test for isolation by distance between the pairwise genetic distance and geographic distances were conducted by GENODIVE (Meirmans and Van Tienderen, 2004). Phylogenetic networks among populations on Okushiri Island was obtained from neighbor-net split network method based on pairwise genetic distance using SplitsTree4 (Huson and Bryant, 2006).

Inference of past population demographic history using the approximate bayesian computation

To infer past demographic history of colonization, we employed the Approximate Bayesian Computation (ABC) approach and used software DIYABC v2.1 (Cornuet et al., 2008; Cornuet et al., 2014). To make the scenarios in the ABC analysis simple, we focused on three representative populations: (1) the mainland of Hokkaido, (2) Okushiri Island, and (3)

Tohoku region. These representative populations were made as follows. According to the results of the above-mentioned STRUCTURE analysis, three clusters were detected that corresponded to these three areas at K = 3. Then the 50 individuals which showed the highest ancestry within each cluster were selected, and named to Pop1 (the mainland of Hokkaido), Pop2 (Okushiri Island) and Pop3 (Tohoku region). These three representative populations, that is a total of 150 individuals, were examined six simple population demography scenarios (Figure 2).

Since DIYABC requires a population which can be traced back to an ancestral population, we set Pop3 to be the ancestor because the Tohoku region showed higher genetic diversity than the mainland of Hokkaido (Takahashi et al., 1994; Hiraoka and Tomaru, 2009) (see also, Table 1). This treatment is reasonable according to Tsuda et al. (2015). In these scenarios, t1 to t6 represents the time scale, measured by generation time, N1 to N3 represents the effective population size of the corresponding populations (Pop1, 2 and 3) and ra to rd represents the ratio of admixture. Although the DIYABC basically does not take gene flow after the split into the scenario, a recent study by Chapuis et al. (2020) modified the DIYABC to allow symmetrical admixture for two population data and we extended this approach to make it more flexible for multiple population data, making it possible to put directional admixture (e.g., from Pop1 to 2, from Pop2 to 1) in the scenarios. Thus, this directional admixture can be analogues to gene flow. The inferred value of admixture parameter (r#) from Pop1 to Pop2 could be considered as the relative amount of gene flow at t# and 1-r# was the value of relative amount of gene flow within a parental population for the admixture (gene flow).

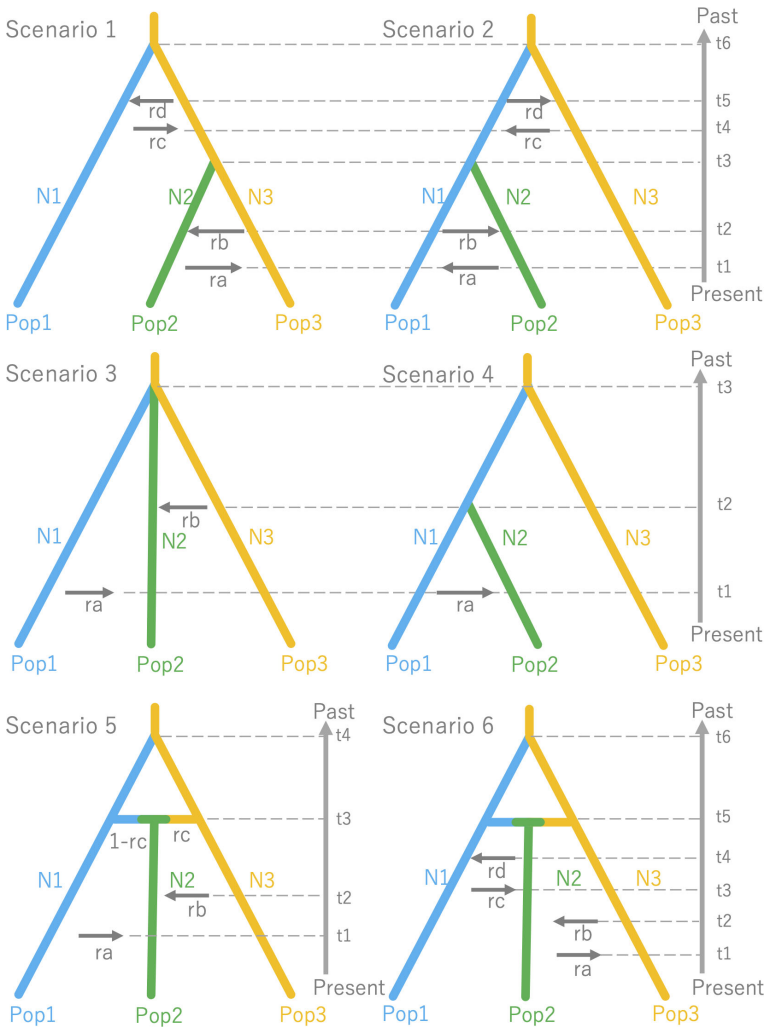


FIGURE 2
Six simple population demography scenarios used for Approximate Bayesian Computation (ABC).

TABLE 1 Genetic parameters for each region.

Region	H_E	Effective number of alleles	G'_{ST}	F_{IS}	Number of alleles
Hokkaido	0.776 (0.681 - 0.858)	5.639 (3.872 - 7.759)	0.016 (0.012 - 0.022)	0.010 (-0.008 - 0.031)	22.273 (17.000 - 28.455)
Okushiri Island	0.778 (0.685 - 0.857)	5.332 (3.804 - 7.006)	0.023 (0.016 - 0.034)	0.012 (-0.007 - 0.042)	22.273 (16.909 - 27.818)
Tohoku	0.797 (0.722 - 0.862)	5.795 (4.019 - 7.841)	0.036 (0.014 - 0.079)	0.033 (0.003 - 0.076)	23.455 (17.545 - 30.182)

95% confidence intervals are in parentheses.

Scenario 1: Hierarchical split model I: Pop2 was merged to Pop3 at t3 and Pop1 was merged to Pop3 at t6, assuming populations on the Okushiri Island were split from Tohoku region. The directional admixture from Pop2 to Pop3 (ra), Pop3 to Pop2 (rb), from Pop1 to Pop3 (rc), and from Pop3 to Pop1 (rd) were set at t1, t2, t4, and t5, respectively. As we did not set any orders of these time scale for the admixtures, the time scale order relied on the inference (e.g., t2 could be earlier than t1) and it was the same in the following Scenarios (2-6).

In Scenarios 3, 4, and 5, the unidirectional admixture was adopted whilst populations on Okushiri Island were assumed to be too small to influence mainland populations by emigrant genes.

Scenario 2: Hierarchical split model II: Pop2 was merged to Pop1 at t3 and Pop1 was merged to Pop3 at t6, assuming populations on Okushiri Island were split from the mainland of Hokkaido. Similar to Scenario 1, the directional admixtures between populations after the splitting were set.

Scenario 3: Simple split model with admixture: all 3 populations diverged at the same time at t3, allowing admixture from Pop1 to Pop2 (ra) and Pop3 to Pop2 (rb) at t1 and t2, respectively.

Scenario 4: Hierarchical split model III: The hierarchical split pattern was similar to Scenario 2. However, we assumed only one directional admixture from Pop1 to Pop2 (ra) at t1.

Scenario 5: Isolation with admixture model I: Pop1 was split from Pop3 at t4 and Pop2 was created by admixtures from Pop3 to Pop2 (rc), and from Pop1 to Pop2 (1-rc) at t3. The secondary directional admixture from Pop1 to Pop2 (ra) and from Pop3 to Pop2 were set at t1 and t2, respectively.

Scenario 6: Isolation with admixture model II: Although this scenario was similar to Scenario 5, we assumed all possible combinations of secondary directional admixture from Pop2 to Pop3 (ra), from Pop3 to Pop2 (rb), from Pop1 to Pop2 (rc), from Pop2 to Pop1 (rd), at t1, 2, 3 and 4, respectively.

The prior distributions of these parameters were shown in Supplementary Table 2. We employed the higher mutation rate from the generalized stepwise mutation model (GSM) (Estoup et al., 2002) and the lower rate for single nucleotide indels (SNI) as mutation models of SSRs. To summarize the observed and simulated data, the mean values for expected heterozygosity (H_E), the number of alleles (A), and the variance of allele size were used as summary statistics for each of the three populations. H_E, A, variance of allele size, classification index (analogues to genotype likelihood) and F_{ST} were the summary statistics for each of the population pairs. One million simulations were run for each scenario. After all the simulations had been run, the most likely scenario was determined by comparing the posterior probabilities using the logistic regression method. The goodness of fit of the scenario was assessed by the option 'model checking' with PCA in DIYABC, which measures the discrepancy between data sets simulated with the prior distributions of parameters and the observed data as well as data sets from the posterior predictive distribution in the scenario.

Results

cpDNA polymorphism

Two haplotypes were distinguished; namely, A (GenBank: AB046492.1) and B (GenBank: AB046493.1) (Fujii et al., 2002), among the populations studied (Supplementary Tables 1, 3, Figure 3). The population was determined to be fixed to either haplotype, when the first analysis of 4 individuals showed the same haplotypes. An additional 12 to 22 individuals were analyzed at populations with mixed haplotypes. Most of the populations were fixed to either haplotype; 16 populations to haplotype A and 23 populations to haplotype B. Most of the populations in Hokkaido were fixed to haplotype A, but most of those on Okushiri Island and in Tohoku were fixed to haplotype B. The other 5 populations included both haplotypes, that is, 1 population in Hokkaido, 3 populations on Okushiri Island, and 1 population in Tohoku. However, the ratio of haplotype A/B differed among populations (Supplementary Table 1, Figure 3). The lower representative haplotype in each region showed specific localization; namely, haplotype B in the southeastern part of Hokkaido, haplotype A in the southeastern part of Okushiri Island, and haplotype A in the northwestern part of Tohoku (Figure 3).

10.3389/fpls.2022.990927

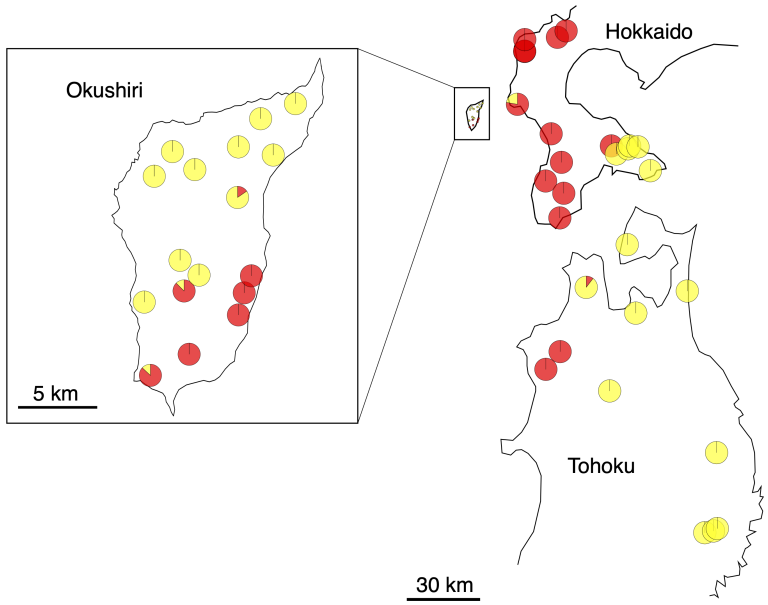


FIGURE 3
Pie diagrams of cpDNA haplotypes. Red and yellow represent haplotype A and B, respectively.

Nuclear gene diversity by SSR

The total heterozygosity (H_T) for all 44 populations was 0.783, and the genetic differentiation for all populations (G'_{ST}) was 0.026. The H_E values of Hokkaido, Okushiri Island, and Tohoku were 0.776 (values for each population ranges from 0.711 to 0.792), 0.778 (from 0.726 to 0.792), and 0.797 (from 0.731 to 0.800), respectively (Table 1, Supplementary Table 1). The effective number of alleles of Hokkaido, Okushiri Island, and Tohoku were 5.639 (each population value ranges from 4.811 to 6.338), 5.332 (from 4.576 to 6.377), and 5.795 (from 4.848 to 6.809), respectively (Table 1, Supplementary Table 1). The G'_{ST} for Hokkaido, Okushiri Island, and Tohoku were 0.016, 0.023, and 0.036, respectively (Table 1). F_{IS} of Hokkaido and Okushiri Island did not show significant deviation from zero, while that of Tohoku was positively deviated from zero (Table 1). The positively significant F_{IS} was observed at one population from Hokkaido and Okushiri Island, and 4 populations from Tohoku; while negatively significant F_{IS} was observed at one population from Okushiri Island (Supplementary Table 1). There were no significant regional differences in genetic diversity parameters, such as H_E, the effective number of alleles, G'_{ST}, and F_{IS}, among Hokkaido, Okushiri Island, and the Tohoku region (Table 1, Supplementary Table 1, Figure 4). The allelic richness ranged from 7.153 to 8.007, from 6.677 to 8.045, and from 7.056 to 8.325, for Hokkaido, Okushiri Island, and Tohoku, respectively (Supplementary Table 1). The number of alleles for Hokkaido, Okushiri Island, and Tohoku were 22.273 (each population value ranged from 7.182 to 13.364), 22.273 (from 7.818 to 13.636), and 23.455 (from 11.091 to 14.364), respectively (Table 1, Supplementary Table 1). High frequency of null allele (0.278) along with a significant deviation from Hardy-Weinberg equilibrium was estimated in mfc12 at population 8 (Supplementary Table 4), however, we believe that it has little influence for the analysis.

Coancestry composition among population

Although the results from STRUCTURE analysis revealed that ΔK was the highest when $K = 2$ (Supplementary Figure 1), the Ln P(D) increased with K. However, variances of Ln P(D) values among runs were larger after $K = 4$. As regards ΔK, even though a hierarchical population scenario is not likely, the highest value can be detected at the uppermost hierarchical level of the population structure (Evanno et al., 2005). Similarly, whilst the highest ΔK is often the case of $K = 2$, further population structure could be revealed in $K > 2$ (Pritchard et al., 2000). In this study, the second highest ΔK was obtained at $K = 4$, whereas the bar plots of $K = 3$ and 4 show similar population structures. Therefore, $K = 3$ could be still a meaningful clustering with the variance of Ln P(D) being smaller than $K > 4$.

We compiled the individual coancestry to each population and drew pie diagrams on geographical map (Figures 5A, B).

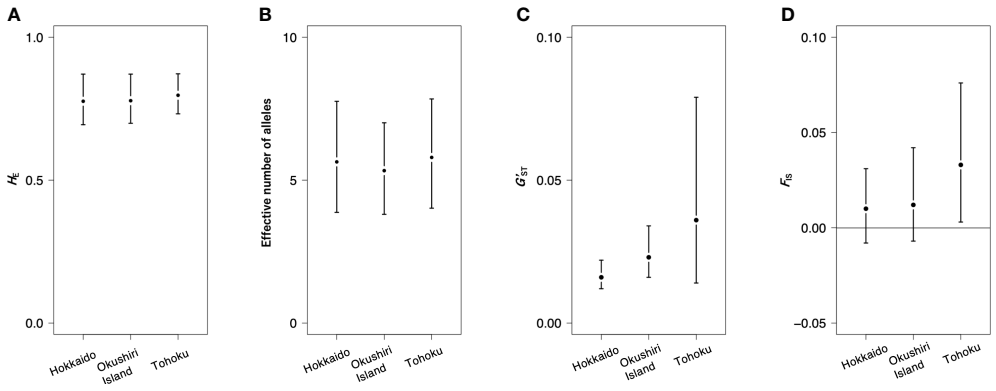


FIGURE 4
Genetic diversity parameters for 3 regions. **(A)** H_E, **(B)** effective number of alleles, **(C)** G'_{ST}, and **(D)** F_{IS}.

When $K = 2$ (Figure 5A), the F values (analogue to F_{ST} between assumed common ancestral population and cluster) were 0.0156 and 0.0275 for Cluster I and II, respectively. The Cluster I was dominant in Tohoku and Cluster II in both Hokkaido and Okushiri Island. Cluster II was more dominant in Okushiri Island than Hokkaido.

When $K = 3$ (Figure 5B), the F values were 0.0125, 0.0122 and 0.0141 for Cluster I, II, and III, respectively. Three different clusters dominated the three different regions; that is, Custer I, II, and III for Tohoku, Hokkaido, and Okushiri Island, respectively. Unrooted neighbor-joining cladograms among the 3 clusters based on F_{ST} genetic distance matrix showed a

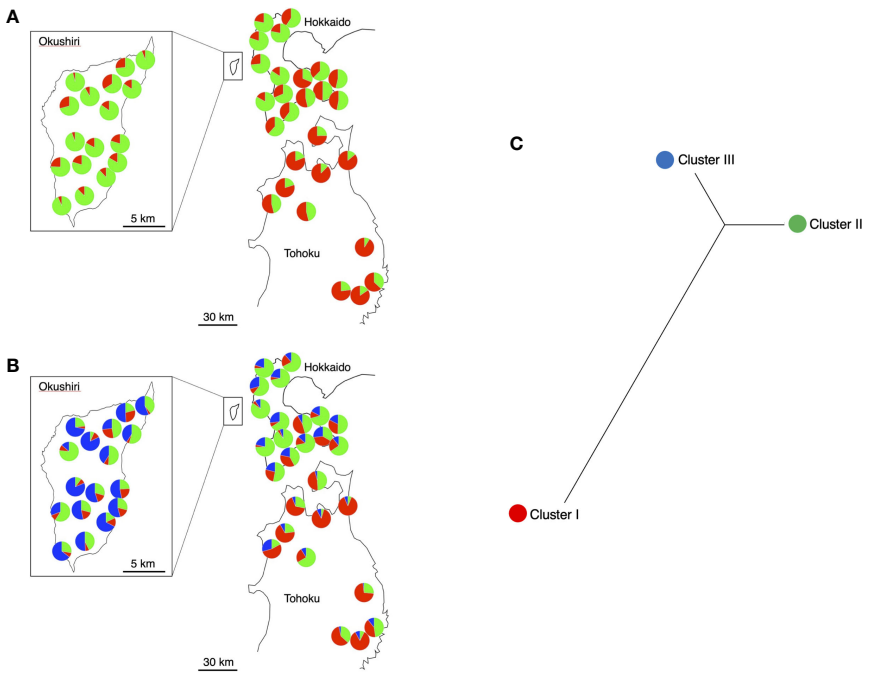


FIGURE 5
Pie diagrams of STRUCTURE results. The different colors indicate the different ancestral clusters. **(A)** $K = 2$, red and green indicate Cluster I and II, respectively. **(B)** $K = 3$, red, green, and blue indicate Cluster I, II, and III, respectively. **(C)** Relationships among the 3 clusters in $K = 3$ based on F_{ST}.

10.3389/fpls.2022.990927

closer relationship between Cluster II and III than between Cluster II or III and Cluster I (Figure 5C).

Principle coordinate analysis, isolation-by-distance, and within-island phylogenetic network

The results of principle coordinate analysis (PCoA) based on the correlation matrix are shown in Figure 6. The first coordinate substantially split the populations in Tohoku from those in Hokkaido and on Okushiri Island. The second coordinate weakly divided populations on Okushiri Island from those in Hokkaido. Also, the second coordinate isolated one population (site 16) located at the tip of the southeastern peninsula of Hokkaido. The cpDNA haplotype differences were reflected in the PCoA diagram (Figure 6); populations fixed with haplotype A (red symbols in Figure 6) were plotted at the center and those with haplotype B (black) scattered on the periphery of the coordinate plane.

Significant isolation-by-distance was detected between study sites (Mantel test by 1,000 permutations; $R^2 = 0.198$, $p = 0.001$) (Figure 7).

A phylogenetic network among populations within Okushiri Island was revealed by neighbor-net split tree (Figure 8). The tree distinguished two geographic localities within the island; southern and northwestern populations. The existence of cpDNA haplotype A on the island was associated with phylogenetic network.

Inference of past population demographic history using the approximate bayesian computation

In DIYABC, the highest posterior probability was found in Scenario 5 (Isolation with admixture model I) (Table 2, Supplementary Figure 2). The value (0.5388, 95% CI: 0.5305 - 0.5471) was much higher than other scenarios and the 95% CI of Scenario 5 was not overlapped with other scenarios. For Scenario 5, the median values of effective population size of N1 (Pop1), N2 (Pop2), and N3 (Pop 3) were 7,820 (95% hyper probability density (HPD): 4,600-9,640), 5,680 (95% HPD; 2,150-8,640) and 9,180 (95% HPD; 6,450-9,970), respectively (Supplementary Table 5). The median values of t1, t2, t3, and t4 were 1,780 (95%HPD: 65.8 -7,500), 4,380 (95%HPD: 998-8,730), 3,870 (95%HPD: 431-8,540) and 7,890 (95%HPD: 3,420-9,910) (Supplementary Table 5). The median values of ra, rb and rc were 0.615 (95%HPD: 0.0585-0.983), 0.615 (95%HPD: 0.0402-0.982) and 0.505 (95%HPD: 0.0272-0.975), respectively (Supplementary Table 5). The median values of the mean mutation rate of SSR and SNI were 8.68 $\times 10^{-4}$ (95% HPD; 4.98×10^{-4}-9.99×10^{-4}) and 8.69 $\times 10^{-7}$ (95% HPD; 1.65×10^{-8}-8.50×10^{-6}), respectively (Supplementary Table 5). The median

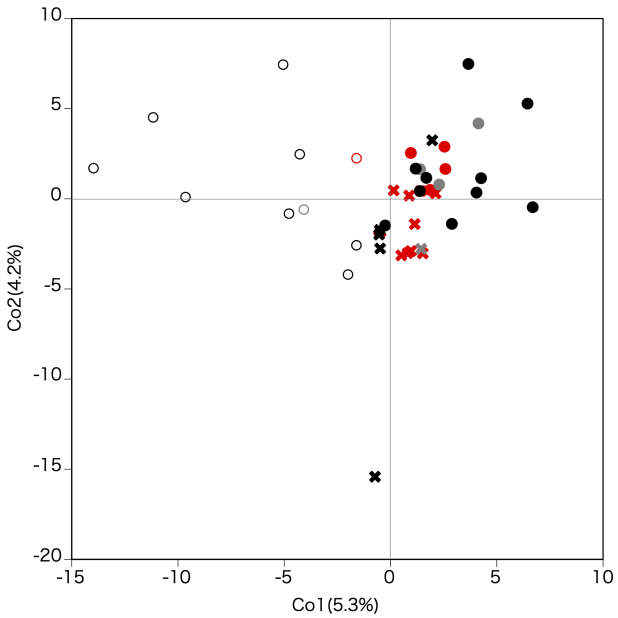


FIGURE 6
Principle coordinate analysis based on the correlation matrix. The first and second axes represent the first and second main components, respectively. Contribution rates of each axis are given in parentheses. Crosses, filled circles, and open circles indicate Hokkaido, Okushiri Island, and Tohoku, respectively. Red, black, and grey indicate cpDNA haplotype A, B, and mixed A and B populations, respectively.

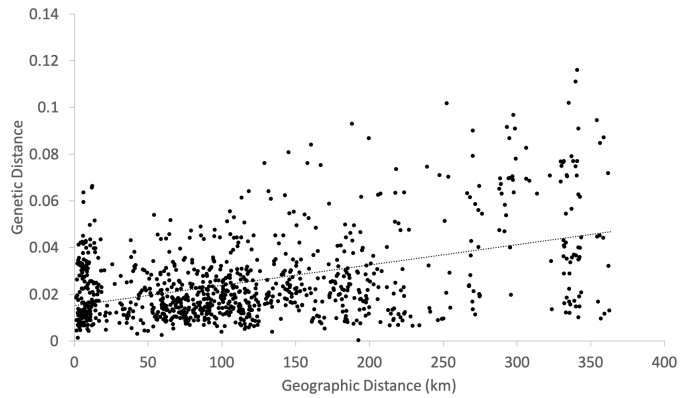


FIGURE 7
Isolation-by-distance between populations. Genetic distance matrix is based on $F_{ST}/(1 - F_{ST})$. Mantel test by 1,000 permutation detected positive correlation between genetic and geographic distances (y = 0.015 + 9E-05x, R^2 = 0.198, p = 0.001).

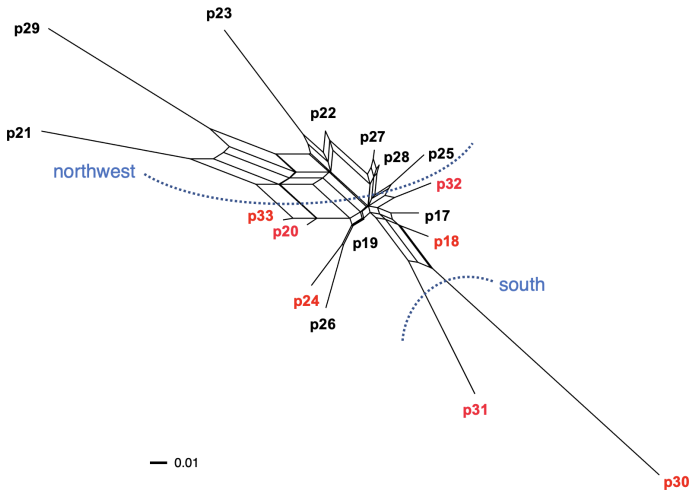


FIGURE 8
Phylogenetic network among populations within Okushiri Island based on pairwise genetic distance. Population numbers are identical to Figure 1 and Supplementary Table 1. Two distinguished localities, south and northwest, are indicated by blue dotted lines. Population in red letters include cpDNA haplotype A.

TABLE 2 Posterior probability of each scenario and its 95% confidence interval based on the logistic estimate by DIYABC.

Scenario	Posterior probability	95% CI (lower - upper)
1	0.0058	0.0029 - 0.0087
2	0.0050	0.0022 - 0.0079
3	0.2201	0.2138 - 0.2265
4	0.1726	0.1670 - 0.1782
5	0.5388	0.5305 - 0.5471
6	0.0576	0.0536 - 0.0616

value of mean P, the parameter of the geometric distribution to generate multiple stepwise mutations was 0.244 (95%HPD; 0.117-0.300) (Supplementary Table 5). Only one of the 27 summary statistics of the simulated data was significantly different from the observed data and the PCA yielded a large cloud of data from the prior and observed datasets, centered around a small cluster from the posterior predictive distribution (Supplementary Figure 3), suggesting a good fit of the scenario together with high posterior probability.

Discussion

Comparable amount of genetic diversity on Okushiri Island to Hokkaido and Tohoku region

The calculated values of H_T of nuclear SSRs of this study (0.711-0.800, mean value: 0.736) were lower than that for the whole geographical range of *F. crenata* (0.793-0.873, mean value: 0.839) (Hiraoka and Tomaru, 2009). This is because the northern *F. crenata* populations show lower genetic diversity than the geographically central populations (Hiraoka and Tomaru, 2009; Kitamura et al., 2015). Our previous study on the northernmost *F. crenata* populations on mainland Hokkaido revealed a decline in genetic diversity at the northern edge of the species distribution (Kitamura et al., 2015). Although Okushiri Island represents the northernmost distribution of *F. crenata* on islands in the Japan Sea, our data did not show any decline in gene diversity (Table 1; Figure 4). The northernmost populations on mainland Hokkaido were relatively new founder populations at the northern front of its expansion (Kitamura et al., 2015; Kitamura and Nakanishi, 2021). On the other hand, the high genetic diversity on Okushiri Island indicated that the *F. crenata* populations on that island were not new founders, but might be populations that have persisted for a long time. This was also supported by the fact that *F. crenata* on Okushiri Island were well differentiated with a similar level of G_{ST} as Hokkaido and the Tohoku region (Figure 4). Thus, although the island is surrounded by ocean, which is a strong practical barrier against gene flow, the effective population size of *F. crenata* remained large enough to compensate for the isolation from the mainland population.

In general, the loss of genetic diversity in island populations can be explained by founder effect during population foundation, limited gene flow and genetic drift due to geographical isolation and small population size (Barrett, 1996; Frankham, 1997). However, our nuclear SSR results revealed that populations on Okushiri Island showed comparable amount of genetic diversity when we focused on Hokkaido and the Tohoku region (Figures 4A–D). As a result, the common assumption that island populations have lower genetic diversity than mainland populations (Barrett, 1996; Frankham, 1997) was not applicable to the *F. crenata* populations on Okushiri Island. García-Verdugo et al. (2015) have reported similar results that the nuclear markers do not tend to show lower genetic variation in island populations than in mainland ones.

Relevant amount of genetic diversity on Okushiri Island is also due to the fact that this island is covered with deciduous-broadleaved forest including a substantial population of *F. crenata* (Tatewaki, 1948; Namikawa et al., 2021). Outcrossing by wind-pollination of the species might compensate for the topographic gene flow barriers such as mountain ridges and valleys (Austerlitz et al., 2004; Hu et al., 2008). When genetic connectivity was secured, a large population size might prevent the loss of gene diversity on the island. Also, it is probable that ancestral alleles were retained within the population (Su et al., 2018) as the species has long life span of 250 years on average in Hokkaido (Kitamura et al., 2007) and the oldest tree on Okushiri Island was confirmed to be 374 years by annual ring counts (T. Matsui, M. Kobayashi, and K. Kitamura, unpublished data). Another supportive evidence of persistence of *F. crenata* on Okushiri Island is that the number of private alleles of SSR on Okushiri Island was 16 which was comparative to 14 in Hokkaido. Therefore, the levels of genetic diversity among *F. crenata* on Okushiri Island might be attributable to the large population size, which allows a significant degree of gene flow, and long demographic history, rather than new founder populations from a few individuals. Indeed, although the ABC-based demographic inference showed that the effective population size of Okushiri Island population was smaller than Hokkaido and Tohoku region, the inferred median value was 5,680 (Supplementary Table 5), which could be a criterion at least to maintain genetic diversity.

CpDNA genetic differentiation among *F. crenata* populations in Okushiri Island, Hokkaido, and Tohoku region

Our results from cpDNA haplotypes clearly separated Hokkaido from the Tohoku region (Supplementary Table 1, Figure 3). A primary difference between *F. crenata* in Hokkaido and the Tohoku region was supported by phytosociology, which categorized the different vegetation types in *F. crenata* forests in Hokkaido and the Japan Sea side of Honshu (Hukusima et al., 1984). Furthermore, the vegetational component of several *F. crenata* populations on the Pacific Ocean side of Hokkaido and the northwestern peninsula of Tohoku were classified into the same category as that on Okushiri Island (Namikawa et al., 2021), which was compatible with the existence of the same haplotype B on Okushiri Island and the Pacific Ocean side of southern Hokkaido (Figure 3). The areas are also characterized by short snow-cover period (Namikawa et al., 2021). These vegetational data might afford circumstantial evidence to

support the existence of relict *F. crenata* populations on Okushiri Island. In addition, Takahashi et al. (2022) revealed that the leaf area of *F. crenata* with haplotype B was significantly smaller than that of haplotype A individuals, which might be adapted to the dry winter environment, namely short snow-cover period.

Our cpDNA haplotype data suggested that a primary differentiation between Hokkaido and the Tohoku region. *Fagus crenata* on Okushiri Island was dominated by the Tohoku lineage yet showed partial establishment of the Hokkaido lineage in the southeastern populations. In addition, nuclear SSR showed coancestry lineages of Okushiri Island belong to Hokkaido lineages at $K = 2$ of the STRUCTURE analysis. The pollen immigration occurred at the northernmost *F. crenata* population at least 12 km in distance (Kitamura and Nakanishi, 2021). This suggested that gene flow by pollen might occur between Okushiri Island and Hokkaido, where the nearest distance between the two regions is 16 km.

Organelle DNA, such as cpDNA, is maternally inherited and dispersed only by seeds (Mogensen, 1996). Thus, the spatial structure of cpDNA haplotypes might indicate the maternal genetic lineages. Previous studies of organelle DNA diversity of *F. crenata* revealed single or closely related lineages in the northern distribution range (Koike et al., 1998; Tomaru et al., 1998; Fujii et al., 2002; Okaura and Harada, 2002). We focused on local populations in the northern distribution range in this study and observed two haplotypes which can be distinguished by a single SNP difference. These different haplotypes appear to indicate the existence of at least two ancestral origins in the northern distribution range (Hewitt, 1999). Haplotype B was shared among northwestern Okushiri Island and the majority of the Tohoku region, while haplotype A was shared among southeastern Okushiri Island and Hokkaido. Hence, the northwestern populations on Okushiri Island might be more closely related to those in the Tohoku region in this study, while those in southeastern Okushiri Island are closely related to the populations in Hokkaido. This spatial structuring of cpDNA haplotypes might reflect the different expansion routes by seed into Okushiri Island. Like the former studies which characterized refugia of this species by cpDNA haplotypes (Okaura and Harada, 2002; Katai et al., 2011; Katai et al., 2017), different haplotypes on the island might indicate the existence of refugia.

Ooi (2016) resolved the species distribution history by an extensive pollen record. According to this, the most recent increase of *Fagus* pollen in Hokkaido observed after period 5 ka (thousand years ago), with its dominance expanding northward. This is in concordance of the species increase along the Japan Sea coast of Tohoku reflecting the increase of snowfall in this area. However, during the late-glacial (from 14 ka to 12 ka), *Fagus* pollen occurred in southern Hokkaido (Ooi, 2016), an area that corresponded to populations 5, and 11 to 16, which shared haplotype B. Moreover, Hoshino (2001) revealed that *F. crenata* have occurred continuously on Okushiri Island from 6,650 ± 120 years BP to the present. Being so, the species might have existed prior to the most recent northward increase along the Japan Sea coast. Given that the haplotype B is of an earlier prosperous lineage, this lineage, namely relict of *F. crenata*, has been sustained within Okushiri Island. Then, the other lineage, which should be represented by haplotype A, increased dominance from the relict source on the island (Ooi, 2016) or immigrated from the Japan Sea coast of Hokkaido. Possible seed dispersal agents of *F. crenata* were spotted nutcrackers (*Nucifraga caryocatactes*) (Vander Wall and Balda, 1977; Kobayashi and Watanabe, 2003; Nishi and Bekku, 2015) and jays (*Garrulus glandarius brandtii*) (Johnson and Adkisson, 1985). Nishi and Bekku (2015) estimated that the spotted nutcracker flew 10.5 km for hoarding seeds, and Kitamura and Nakanishi (2021) revealed seed migration of *F. crenata* at least 12 km in distance in Hokkaido. Also, European nutcrackers carry seeds as far as 22 km (Vander Wall and Balda, 1977). Thus, a 16 km-long seed dispersal between Hokkaido and Okushiri Island likely occurs. In addition, the cpDNA haplotype mixture such as populations 18, 20, and 31, might have resulted from recent seed dispersal on the island, given that the haplotype A lineage increased dominance more recently than the haplotype B lineage.

Demographic history and northern persistence of *F. crenata* predated the LGM

The result of the demographic inference suggested that the scenario of the isolation with admixture and secondary admixture was the most likely to explain the genetic structure of *F. crenata* populations examined in this study (Figure 5). The STRUCTURE analysis by nuclear SSR genotype revealed different coancestries between Hokkaido (including Okushiri Island) and the Tohoku region when $K = 2$ (Figure 5A), and a distinct difference between Hokkaido and Okushiri Island when $K = 3$ (Figure 5B). Moreover, PCoA separated the Tohoku region at the first coordinate and secondary separation between Hokkaido and Okushiri Island (Figure 6). Although a clear admixture pattern was not detected in Okushiri Island populations at $K = 2$, this might be due to recent secondary gene flow from both Hokkaido and Tohoku populations.

The highest probability was obtained from Scenario 5 (Table 2, Figure 2, Supplementary Figure 2). Although time scales for the demographic events were inferred in this study, transformation of generation time to year is still challenging in tree species as they are long-lived with overlapping of generations (Tsuda et al., 2015). Therefore, generation time assumption causes uncertainty in the inference. However, ecological information is still useful to consider generation time, and a feasible substitution for the generation time can be

the earliest reproductive age of the species in the northern range. There are several observations for the earliest reproductive age of *F. crenata* of the northern distribution. In the Kuromatsunai Depression, naturally dispersed *F. crenata* in the open environment have flowered at 20 and 21 years of age (K. Kitamura and H. Saito, personal observations 2013). An ornamental *F. crenata* tree at Kuromatsunai Depression flowered at 18 years after the juvenile plantation (H. Saito, personal observations, 2013). Provenance tests of *F. crenata* at two nurseries of Tokyo University (Chichibu, Saitama, Japan and Furano, Hokkaido, Japan) revealed that the northern originated *F. crenata* flowered 15 and 17 years after juvenile plantation at the earliest (S. Goto and M. Takahashi, personal observations 2013). On the other hand, the age of reproductive trees at the northernmost population was estimated to be 80 years (Kitamura and Nakanishi, 2021). From these facts, we assumed two different generation time; 20 years at the earliest and 100 years at the fully matured status for *F. crenata* in this study (Table 3). Thus, the time scales of t1 (admixture from Pop1 to Pop2), t2 (admixture from Pop3 to Pop2), t3 (admixture to generate Pop2) and t4 (splitting time of Pop1 from Pop3) were inferred 35,600 (95% HPD: 1,316 - 150,000), 87,600 (95% HPD: 19,960 - 174,600), 77,400 (95% HPD: 8,620-170,800) and 157,800 (95% HPD: 68,400-198,200) years BP, respectively, under the assumption of generation time of 20 years. When we assumed 100 years for the generation time, they are 178,000 (95% HPD: 6,580 - 750,000), 438,000 (95% HPD: 99,800 - 873,000), 387,000 (95% HPD: 43,100-854,000) and 789,000 (95% HPD: 342,000-991,000) years BP, respectively. The results suggested the divergence between Pop1 (Hokkaido) and Pop3 (Tohoku) predated the LGM even under the assumption of a short generation time of 20 years and even when we considered the 95% HPDs. The divergence time was much earlier when we assumed 100 years for the generation time. Moreover, the demographic inference suggested that the Okushiri Island population was generated by admixture between Hokkaido and Tohoku populations 77,400 (95% HPD: 8,620-170,800) and 387,000 (95% HPD: 43,100-854,000) years BP, under the assumption of the generation time of 20 and 100 years, respectively. The inferred rc value (0.505) suggested that both parental populations equally contributed to the admixture to Okushiri Island population. Notably, asymmetrical secondary admixtures to Okushiri Island population from both parental populations was inferred in this study. However, as the median value of t2 (4,380), time of the secondary admixture from Tohoku to Okushiri Island populations, was slightly larger than t3 (3,870) and their 95% HPDs were overlapped, it is difficult to discuss the time scale of this secondary admixture in detail. On the other hand, the secondary asymmetrical admixture from Hokkaido to Okushiri Island populations was inferred to be 35,600 (95% HPD: 1,316 - 150,000) years BP, under the assumption of the generation time of 20 years, suggesting recent gene flow to Okushiri Island population from Hokkaido, probably in relation to the demographic episode during the last ice age or the LGM. Although it is difficult to estimate locations where ancestors of Hokkaido and Tohoku populations distributed in the past and the secondary admixture could not be a single event nor at a specific time but rather by continuous gene flow over time, these finding shed a new light on past population demography of the northern populations of *F. crenata*, and these results were different from the previous studies which discussed post-LGM northward expansion to Hokkaido from Tohoku region (Tsukada, 1982; Tomaru et al., 1997; Fujii et al., 2002).

Recent studies of distribution shift of species based on palaeoecological and/or genetic data revealed the existence of cryptic refugia and northern persistence of forest tree species in more northerly latitudes than initially expectations which assumed southern refugia in lower altitude, not only in Europe and North America but also Japan (Petit et al., 2002; McLachlan and Clark, 2004; Tsuda and Ide, 2005; Petit et al., 2008; Provan and Bennett, 2008; Liepelt et al., 2009; Parducci et al., 2012; Tsuda et al., 2015; Tsuda et al., 2016; Tsuda et al., 2017). Bradshaw et al. (2010) reviewed long-term distribution dynamics of 3 *Fagus* species; namely, *F. sylvatica*, *F. grandifolia*, and *F. crenata*, and indicated the recent northward spread from several outlying founder populations as well as the location of glacial refugia (Kito and Takimoto, 1999; McLachlan et al., 2005; Magri et al., 2006; Magri, 2008). This hypothesis of northern persistence of cold-tolerant tree species was well supported by species distribution model (SDM) (Svenning et al., 2008) and a combined analysis of the

TABLE 3 Demographic parameters of Scenario 5 obtained by DIYABC.

Parameter	Generations	Generation time	
		20 years	100 years
t1	1,780 (95%HPD: 65.8-7,500)	35,600 (95%HPD: 1,316 - 150,000)	178,000 (95%HPD: 6,580 - 750,000)
t2	4,380 (95%HPD: 998 - 8,730)	87,600 (95%HPD: 19,960 - 174,600)	438,000 (95%HPD: 99,800 - 873,000)
t3	3,870 (95%HPD: 431-8,540)	77,400 (95%HPD: 8,620-170,800)	387,000 (95%HPD: 43,100-854,000)
t4	7,890 (95%HPD: 3,420-9,910)	157,800 (95%HPD: 68,400-198,200)	789,000 (95%HPD: 342,000-991,000)

SDM and genetic based-demographic inference (Tsuda et al., 2015). Indeed, a number of palynological studies suggest that there must have been cryptic refugia on the Japan Sea side of Hokkaido (Nakamura and Tsukada, 1960; Nakamura, 1968; Uemura and Takeda, 1987; Kito, 2015; Ooi, 2016). Actually, Tsuda et al. (2015) showed a high possibility of persistence of *Betula maximowicziana*, a birch species that often distributes in cool temperate forests with *F. crenata*, during the LGM mainly in the western part of Hokkaido as well as the northern part of the Tohoku region. The distribution of *F. crenata* in the Pleistocene interglacial periods expanded further north on the Japan Sea side and east to the Pacific Ocean side of Hokkaido (Yano, 1972). Palynological studies revealed that *F. crenata* existed in Hokkaido (Sakaguchi, 1989; Kito and Takimoto, 1999; Kito and Ohkuro, 2012) and Okushiri Island (Yamanoi, 1992; Hoshino, 2001) before LGM. Therefore, *F. crenata* might have been distributed further north and east than the present geographic margin in Hokkaido before the straits divided the Japanese Archipelago (Ito, 1987; Ooi, 2016). Also, phytogeographical evidence of the disjunctive distribution of plants on Okushiri Island and the eastern part of Hokkaido supported the notion that temperate species were distributed further north and east than the present geographic margin (Tatewaki, 1954). Moreover, the Okushiri Strait separated Okushiri Island from Hokkaido prior to the Middle Pleistocene, and then the Tsugaru Strait divided Hokkaido from Honshu in the Riss-Würm interglacial period (Ohshima, 1980). These straits could be indeed physical barriers to prevent long-distance gene flow between regions. Thus, it appears that Okushiri Island, besides the Japan Sea coast of Hokkaido, was a possible persistence of *F. crenata* in the northern range. The similar amount of genetic diversity in the northern populations from Tohoku to Hokkaido regions might be mirrored by long-term persistence of populations as well as no experience of severe bottlenecks by the post-LGM colonization. Indeed, we did not detect any recent bottlenecks in the examined populations in the pilot run of the ABC inference in this study (data not shown), while newly founded northern marginal populations showed the effects of genetic drift (Kobayashi et al., 2013; Kitamura et al., 2015; Kitamura and Nakanishi, 2021). Moreover, *F. crenata* on Okushiri Island showed a comparative number of private alleles of SSR to Hokkaido, which indicated the existence of relict on this island.

Finally, the results of this study coincided with palaeoecological and vegetation studies as well as geology of the Okushiri Island, and supported the existence of possible persistence in the northern distribution of *F. crenata* with secondary admixture from the both parental populations in Hokkaido and Tohoku region. This means Okushiri Island included cryptic refugia of *F. crenata* which persisted for a longer period than expected from the post-LGM expansion.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

Conceptualization, KK and YT. Methodology, KK and YT. Formal analysis, KK and YT. Investigation KK, KN, TM, and MK. Writing – original draft preparation, KK. Writing- review and editing, KK, KN, YT, TM, and MK. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by JSPS KAKENHI (JP17K07852 and JP20K06152) and Core-to-Core Program (Asia-Africa Science Platforms: JPJSCCB20220007) from the Japan Society for the Promotion of Science and the 27th Pro Natura Fund Grant Program from the Pro Natura Foundation Japan.

Acknowledgments

We thank M. Ooue, M. Takahashi, H. Saitou, T. Nagamitsu, A. Nakanishi, and A. Takazawa for their assistance in the field and laboratory work.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.990927/full#supplementary-material

References

Alsos, I. G., Ehrich, D., Eidesen, P. B., Solstad, H., Westergaard, K. B., Schönswetter, P., et al. (2015). Long-distance plant dispersal to north Atlantic islands: Colonization routes and founder effect. *AoB Plants* 7 (1), 1–19. doi: 10.1093/aobpla/plv036

Asuka, Y., Tani, N., Tsumura, Y., and Tomaru, N. (2004). Development and characterization of microsatellite markers for *Fagus crenata* blume. *Mol. Ecol. Notes* 4 (1), 101–103. doi: 10.1046/j.1471-8286.2003.00583.x

Austerlitz, F., Dick, C. W., Dutech, C., Klein, E. K., Oddou-Muratorio, S., Smouse, P. E., et al. (2004). Using genetic markers to estimate the pollen dispersal curve. *Mol. Ecol.* 13 (4), 937–954. doi: 10.1111/j.1365-294X.2004.02100.x

Barrett, S. C. H. (1996). The reproductive biology and genetics of island plants. *Philos. Trans. R. Soc. London B* 351, 725–733. doi: 10.1098/rstb.1996.0067

Bradshaw, R. H. W., Kito, N., and Giesecke, T. (2010). Factors influencing the Holocene history of *Fagus*. *For. Ecol. Manage.* 259 (11), 2204–2212. doi: 10.1016/j.foreco.2009.11.035

Chiang, Y. C., Hung, K. H., Schaal, B. A., Ge, X. J., Hsu, T. W., and Chiang, T. Y. (2006). Contrasting phylogeographical patterns between mainland and island taxa of the *Pinus luchuensis* complex. *Mol. Ecol.* 15 (3), 765–779. doi: 10.1111/j.1365-294X.2005.02833.x

Cornuet, J. M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., et al. (2014). DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30 (8), 1187–1189. doi: 10.1093/bioinformatics/btt763

Cornuet, J. M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J. M., Balding, D. J., et al. (2008). Inferring population history with *DIY ABC*: A user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24 (23), 2713–2719. doi: 10.1093/bioinformatics/btn514

DeJoode, D. R., and Wendel, J. F. (1992). Genetic diversity and origin of the Hawaiian islands cotton, *Gossypium tomentosum*. *Am. J. Bot.* 79 (11), 1311–1311. doi: 10.2307/2445059

Désamoré, A., Laenen, B., González-Mancebo, J. M., Jaén Molina, R., Bystriakova, N., Martinez-Klimova, E., et al. (2012). Inverted patterns of genetic diversity in continental and island populations of the heather. *Erica scoparia* s.l. *J. Biogeography* 39 (3), 574–584. doi: 10.1111/j.1365-2699.2011.02622.x

Elleouet, J. S., and Aitken, S. N. (2019). Long-distance pollen dispersal during recent colonization favors a rapid but partial recovery of genetic diversity in *Picea sitchensis*. *N. Phytol.* 222 (2), 1088–1100. doi: 10.1111/nph.15615

El Mousadik, A., and Petit, R. J. (1996). High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) skeels] endemic to Morocco. *Theor. Appl. Genet.* 92 (7), 832–839. doi: 10.1007/BF00221895

Estoup, A., Jarne, P., and Cornuet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* 11, 1591–1604. doi: 10.1046/j.1365-294X.2002.01576.x

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* 14 (8), 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.1093/genetics/164.4.1567

Frankham, R. (1996). Relationship of genetic variation to population size in wildlife. *Conserv. Biol.* 10 (6), 1500–1508. doi: 10.1046/j.1523-1739.1996.10061500.x

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations? *Heredity* 78 (3), 311–327. doi: 10.1038/hdy.1997.46

Fujii, N., Tomaru, N., Okuyama, K., Koike, T., Mikami, T., and Ueda, K. (2002). Chloroplast DNA phylogeography of *Fagus crenata* (Fagaceae) in Japan. *Plant Systematics Evol.* 232 (1-2), 21–33. doi: 10.1007/s006060200024

García-Verdugo, C., Calleja, J. A., Vargas, P., Silva, L., Moreira, O., and Pulido, F. (2013). Polyploidy and microsatellite variation in the relict tree *Prunus lusitanica* l.: How effective are refugia in preserving genotypic diversity of clonal taxa? *Mol. Ecol.* 22 (6), 1546–1557. doi: 10.1111/mec.12194

García-Verdugo, C., Sajeva, M., La Mantia, T., Harrouni, C., Msanda, F., and Caujapé-Castells, J. (2015). Do island plant populations really have lower genetic variation than mainland populations? Effects of selection and distribution range on genetic diversity estimates. *Mol. Ecol.* 24 (4), 726–741. doi: 10.1111/mec.13060

Goudet, J. (2001). *FSTAT, a program to estimate and test gene diversities and fixation indices* (Lausanne, Switzerland: University of Lausanne).

Hagiwara, S. (1977). Buna ni mirareru youmenseki no kurain ni tsuite [Clines of leaf area of beech]. *Soc. Study Species Biol. / Shuseibutsugaku Kenkyu* 1, 39–51.

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* 68 (1-2), 87–112. doi: 10.1006/bijl.1999.0332

Hiraoka, K., and Tomaru, N. (2009). Genetic divergence in nuclear genomes between populations of *Fagus crenata* along the Japan Sea and Pacific sides of Japan. *J. Plant Res.* 122 (3), 269–282. doi: 10.1007/s10265-009-0217-9

Hoshino, F. (2001). Okushirito chuobu no kahun bunseki [Fossil pollen analysis in central Okushiri Island]. *Hoppo Sanso* 18, 53–58.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9 (5), 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

Hukusima, T., Nashimoto, M., and Watanabe, I. (1984). Phytosociological studies on the beech forest in Hokkaido, Japan. *Tech. Bull. Faculty Horticulture Chiba Univ.* 33, 117–131.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23 (2), 254–267. doi: 10.1093/molbev/msj030

Hu, L.-J., Uchiyama, K., Shen, H.-L., Saito, Y., Tsuda, Y., and Ide, Y. (2008). Nuclear DNA microsatellites reveal genetic variation but a lack of phylogeographical structure in an endangered species, *Fraxinus mandshurica*, across north-east China. *Ann. Bot.* 102 (2), 195–205. doi: 10.1093/aob/mcn074

Ishii, H. R., Horikawa, S.i., Noguchi, Y., and Azuma, W. (2018). Variation of intra-crown leaf plasticity of *Fagus crenata* across its geographical range in Japan. *For. Ecol. Manage.* 429, 437–448. doi: 10.1016/j.foreco.2018.07.016

Ito, K. (1987). *Vegetation of Hokkaido* (Sapporo: Hokkaido University Press).

Johnson, W. C., and Adkisson, C. S. (1985). Dispersal of beech nuts by blue jays in fragmented landscape. *Am. Midland Nat.* 113 (2), 319–324. doi: 10.2307/2425577

Katai, H., Takahashi, M., Hiraoka, K., Yamada, S., and Tomaru, N. (2017). Indigenous genetic lineages of *Fagus crenata* found in the Izu Peninsula suggest that there was one of refugia for the species during the last glacial maximum. *J. For. Res.* 18 (5), 418–429. doi: 10.1007/s10310-012-0368-8

Katai, H., Takahashi, M., Hiraoka, K., Yamada, S., Yamamoto, S., Kato, K., et al. (2011). Inference of genetic lineages of *Fagus crenata* populations in Shizuoka Prefecture based on chloroplast DNA and nuclear microsatellite variations. *J. Japanese For. Soc.* 93 (2), 73–78. doi: 10.4005/jjfs.93.73

Katai, H., Yamada, S., Hiraoka, K., Hoshikawa, T., Tomaru, N., and Takahashi, M. (2014). Genetic lineage and diversity of *Fagus crenata* trees planted in Shizuoka Prefecture. *For. Genet. Tree Breed.* 3, 101–110.

Kitamura, K., Kobayashi, M., and Kawahara, T. (2007). Age structure of wind-felled canopy trees for Siebold's beech (*Fagus crenata*) in the northernmost population in Karibayama, Hokkaido. *J. For. Res.* 12 (6), 467–472. doi: 10.1007/s10310-007-0026-8

Kitamura, K., Matsui, T., Kobayashi, M., Saitou, H., Namikawa, K., and Tsuda, Y. (2015). Decline in gene diversity and strong genetic drift in the northward-expanding marginal populations of *Fagus crenata*. *Tree Genet. Genomes* 11. doi: 10.1007/s11295-015-0857-y. Article Number 36.

Kitamura, K., and Nakanishi, A. (2021). Recovery process of genetic diversity through seed and pollen immigration at the northernmost leading-edge population of *Fagus crenata*. *Plant Species Biol.* 36 (3), 489–502. doi: 10.1111/1442-1984.12332

Kito, N. (2015). Expansion of beech forest in Tohoku and Hokkaido since Last Glacial. *Japanese J. For. Environ.* 57 (2), 69–74. doi: 10.18922/jjfe.57.2_69

Kito, N., and Ohkuro, Y. (2012). Vegetation response to climatic oscillations during the last glacial-interglacial transition in northern Japan. *Quaternary Int.* 254, 118–128. doi: 10.1016/j.quaint.2011.06.050

Kito, N., and Takimoto, F. (1999). Population growth and migration rate of *Fagus crenata* during the Holocene in southwestern Hokkaido, Japan. *Quaternary Res.* 38 (4), 297–311.

Kobayashi, M., Kitamura, K., Matsui, T., and Kawano, S. (2013). Genetic characteristics reflecting the population size and disturbance regime of Siebold's beech (*Fagus crenata* blume) populations at the northernmost distribution. *Silvae Genetica* 62 (1-2), 1–7. doi: 10.1515/sg-2013-0001

Kobayashi, M., and Watanabe, S. (2003). Stand structure of the northern bound population of *Fagus crenata*, located at Tsubamenosawa, Hokkaido, Japan. *Bull. Geo-environmental Sci.* 5, 1–23.

Koike, T., Kato, S., Shimamoto, Y., Kitamura, K., Kawano, S., Ueda, K., et al. (1998). Mitochondrial DNA variation follows a geographic pattern in Japanese beech species. *Botanica Acta* 111 (1), 87–91. doi: 10.1111/j.1438-8677.1998.tb00682.x

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inferences across *K*. *Mol. Ecol. Resour.* 15 (5), 1179–1191. doi: 10.1111/1755-0998.12387

Kuparinen, A., Savolainen, O., and Schurr, F. M. (2010). Increased mortality can promote evolutionary adaptation of forest trees to climate change. *For. Ecol. Manage.* 259 (5), 1003–1008. doi: 10.1016/j.foreco.2009.12.006

Liepelt, S., Cheddadi, R., de Beaulieu, J. L., Fady, B., Gömöry, D., Hussendörfer, E., et al. (2009). Postglacial range expansion and its genetic imprints in *Abies alba* (Mill.) - a synthesis from palaeobotanic and genetic data. *Rev. Palaeobotany Palynology* 153 (1-2), 139–149. doi: 10.1016/j.revpalbo.2008.07.007

Magri, D. (2008). Patterns of post-glacial spread and the extent of glacial refugia of European beech (*Fagus sylvatica*). *J. Biogeography* 35 (3), 450–463. doi: 10.1111/j.1365-2699.2007.01803.x

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., et al. (2006). A new scenario for the Quaternary history of European beech populations: Palaeobotanical evidence and genetic consequences. *N Phytol.* 171 (1), 199–221. doi: 10.1111/j.1469-8137.2006.01740.x

Marshall, T. C., Slate, J., Kruuk, L. E. B., and Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* 7 (5), 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

McLachlan, J. S., and Clark, J. S. (2004). Reconstructing historical ranges with fossil data at continental scales. *For. Ecol. Manage.* 197 (1-3), 139–147. doi: 10.1016/j.foreco.2004.05.026

McLachlan, J. S., Clark, J. S., and Manos, P. S. (2005). Molecular indicators of tree migration capacity under rapid climate change. *Ecology* 86 (8), 2088–2098. doi: 10.1890/04-1036

Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* 4 (4), 792–794. doi: 10.1111/j.1471-8286.2004.00770.x

Miguchi, H. (1996). Dynamics of beech forest from the view point of rodents ecology - ecological interactions of the regeneration characteristics of *Fagus crenata* and rodents. *Japanese J. Ecol.* 46, 185–189.

Mogensen, H. L. (1996). The hows and whys of cytoplasmic inheritance in seed plants. *Am. J. Bot.* 83 (3), 383–404. doi: 10.1002/j.1537-2197.1996.tb12718.x

Nakamura, J. (1968). Palynological aspects of the Quaternary in Hokkaido. V. pollen succession and climatic change since the upper Pleistocene. *Research Reports of Kochi University. Natural Sci.* 17 (3), 39–51.

Nakamura, J., and Tsukada, M. (1960). Palynological aspects of the Quaternary in Hokkaido. I. the Oshima Peninsula. *Research Reports of Kochi University. Natural Sci.* 9, 117–138.

Namikawa, K., Kitamura, K., Matsui, T., and Ishikawa, Y. (2021). The comparative study on the species composition of *Fagus crenata* Blume forests in Okushiri Island of southwestern Hokkaido and its surrounding area. *Vegetation Sci.* 38, 175–190.

Nishi, N., and Bekku, Y. S. (2015). The spotted nutcracker hoarding seeds of Japanese white pine in Mt. Fuji where Japanese stone pine is not distributed. *Strix* 31, 113–123.

Ohshima, K. (1980). Recording the Late-Quaternary sea-level change on the topographic feature of the straits of the Japanese islands. *Quaternary Res. Tokyo* 19 (1), 23–37. doi: 10.4116/jaqua.19.23

Okaura, T., and Harada, K. (2002). Phylogeographical structure revealed by chloroplast DNA variation in Japanese beech (*Fagus crenata* Blume). *Heredity* 88, 322–329. doi: 10.1038/sj.hdy.6800048

Ooi, N. (2016). Vegetation history of Japan since the last glacial based on palynological data. *Japanease J. Historical Bot.* 25 (1-2), 1–101.

Parducci, L., Jørgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., et al. (2012). Glacial survival of boreal trees in northern Scandinavia. *Science* 335 (6072), 1083–1086. doi: 10.1126/science.1216043

Pastorelli, R., Smulders, M.J.M., Van'T Westende, W.P.C., Vosman, B., Giannini, R., Vettori, C., et al (2003). Characterization of microsatellite markers in *Fagus sylvatica* L. and *Fagus orientalis* Lipsky. *Mol. Ecol. Notes* 3 (1), 76–78. doi: 10.1046/j.1471-8286.2003.00355.x

Peters, R. (1997). *Beech forests* (Dordrecht: Kluwer Academic Publishers).

Petit, R. J., Brewer, S., Bordács, S., Burg, K., Cheddadi, R., Coart, E., et al. (2002). Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *For. Ecol. Manage.* 156 (1-3), 49–74. doi: 10.1016/S0378-1127(01)00634-X

Petit, R. J., and Excoffier, L. (2009). Gene flow and species delimitation. *Trends Ecol. Evol.* 24 (7), 386–393. doi: 10.1016/j.tree.2009.02.011

Petit, R. J., Hu, F. S., and Dick, C. W. (2008). Forests of the past: A window to future changes. *Science* 320 (5882), 1450–1452. doi: 10.1126/science.1155457

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Provan, J., and Bennett, K. D. (2008). Phylogeographic insights into cryptic glacial refugia. *Trends Ecol. Evol.* 23 (10), 564–571. doi: 10.1016/j.tree.2008.06.010

Rosas Escobar, P., Gernandt, D. S., Piñero, D., and Garcillán, P. P. (2011). Plastid DNA diversity is higher in the island endemic guadalupe cypress than in the continental tecate cypress. *PloS One* 6 (1), e16133–e16133. doi: 10.1371/journal.pone.0016133

Sakaguchi, Y. (1989). Some pollen records from Hokkaido and Sakhalin. *Bull. Department Geography Univ. Tokyo* 21, 1–17.

St Clair, J. B., and Howe, G. T. (2007). Genetic maladaptation of coastal Douglas-fir seedlings to future climates. *Global Change Biol.* 13 (7), 1441–1454. doi: 10.1111/j.1365-2486.2007.01385.x

Stuessy, T. F. (2020). The importance of historical ecology for interpreting evolutionary processes in plants of oceanic islands. *J. Systematics Evol.* 58 (6), 751–766. doi: 10.1111/jse.12673

Stuessy, T. F., Takayama, K., López-Sepúlveda, P., and Crawford, D. J. (2014). Interpretation of patterns of genetic variation in endemic plant species of oceanic islands. *Botanical J. Linn. Soc.* 174 (3), 276–288. doi: 10.1111/boj.12088

Su, J., Yan, Y., Song, J., Li, J., Mao, J., Wang, N., et al. (2018). Recent fragmentation may not alter genetic patterns in endangered long-lived species: Evidence from *Taxus cuspidata*. *Front. Plant Sci.* 871. doi: 10.3389/fpls.2018.01571

Svenning, J. C., Normand, S., and Kageyama, M. (2008). Glacial refugia of temperate trees in Europe: Insights from species distribution modelling. *J. Ecol.* 96 (6), 1117–1127. doi: 10.1111/j.1365-2745.2008.01422.x

Taberlet, P., Fumagalli, L., Wust-Saucy, A. G., and Cosson, J. F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.* 7 (4), 453–464. doi: 10.1046/j.1365-294x.1998.00289.x

Takahashi, M., Goto, S., Fukuda, Y., and Watanabe, A. (2022). Utility of chloroplast DNA haplotype data for ecological restoration using *Fagus crenata* seedlings in case of incomplete seed source information availability. *Ecol. Res.* doi: 10.1111/1440-1703.12351

Takahashi, M., Tsumura, Y., Nakamura, T., Uchida, K., and Ohba, K. (1994). Allozyme variation of *Fagus crenata* in northeastern Japan. *Can. J. For. Res.* 24 (5), 1071–1074. doi: 10.1139/x94-141

Takayama, K., Lopez-Sepulveda, P., Greimler, J., Crawford, D. J., Penailillo, P., Baeza, M., et al. (2015). Relationships and genetic consequences of contrasting modes of speciation among endemic species of Robinsonia (Asteraceae, senecioneae) of the Juan Fernandez Archipelago, Chile, based on AFLPs and SSRs. *N. Phytol.* 205 (1), 415–428. doi: 10.1111/nph.13000

Takayama, K., Sun, B. Y., and Stuessy, T. F. (2013). Anagenetic speciation in Ullung Island, Korea: Genetic diversity and structure in the island endemic species, *Acer takesimense* (Sapindaceae). *J. Plant Res.* 126 (3), 323–333. doi: 10.1007/s10265-012-0529-z

Tanaka, K., Nakamura, T., and Tsumura, Y. (1999). Development and polymorphism of microsatellite markers for *Fagus crenata* and the closely related species, *F. Japonica*. *Theor. Appl. Genet.* 99 (1-2), 11–15. doi: 10.1007/s001220051203

Tatewaki, M. (1948). Buna no hokugenkai [Northern limit of *Fagus crenata]*. *Ecol. Rev. / Seitaigaku Kenkyu* 11, 46–51.

Tatewaki, M. (1954). Disjunctive distribution of the flowering plants in Hokkaido, Japan. *Bull. Soc. Plant Ecol.* 3 (4), 250–170.

Tomaru, N., Mitsutsuji, T., Takahashi, M., Tsumura, Y., Uchida, K., and Ohba, K. (1997). Genetic diversity in *Fagus crenata* (Japanese beech): Influence of the distributional shift during the late-Quaternary. *Heredity* 78 (3), 241–251. doi: 10.1038/hdy.1997.38

Tomaru, N., Takahashi, M., Tsumura, Y., Takahashi, M., and Ohba, K. (1998). Intraspecific variation and phylogeographic patterns of *Fagus crenata* (Fagaceae) mitochondrial DNA. *Am. J. Bot.* 85 (5), 629–636. doi: 10.2307/2446531

Tsuda, Y., Chen, J., Stocks, M., Källman, T., Sønstebø, J. H., Parducci, L., et al. (2016). The extent and meaning of hybridization and introgression between Siberian spruce (*Picea obovata*) and Norway spruce (*Picea abies*): Cryptic refugia as stepping stones to the west? *Mol. Ecol.* 25 (12), 2773–2789. doi: 10.1111/mec.13654

Tsuda, Y., and Ide, Y. (2005). Wide-range analysis of genetic structure of *Betula maximowicziana*, a long-lived pioneer tree species and noble hardwood in the cool temperate zone of Japan. *Mol. Ecol.* 14 (13), 3929–3941. doi: 10.1111/j.1365-294X.2005.02715.x

Tsuda, Y., Nakao, K., Ide, Y., and Tsumura, Y. (2015). The population demography of *Betula maximowicziana*, a cool-temperate tree species in Japan, in relation to the last glacial period: Its admixture-like genetic structure is the result of simple population splitting not admixing. *Mol. Ecol.* 24 (7), 1403–1418. doi: 10.1111/mec.13123

Tsuda, Y., Semerikov, V. L., Sebastiani, F., Vendramin, G. G., and Lascoux, M. (2017). Multispecies genetic structure and hybridization in the *Betula* genus across Eurasia. *Mol. Ecol.* 26 (2), 589–605. doi: 10.1111/mec.13885

Tsukada, M. (1982). Late-quaternary shift of *Fagus* distribution. *Botanical Magazine Tokyo* 95 (2), 203–217. doi: 10.1007/BF02488586

Uemura, S., and Takeda, Y. (1987). Phytogeographical study on the distribution of the main temperate plants composing the natural forests in Hokkaido, Japan. *Papers Plant Ecol. Taxonomy to Memory Dr. Satoshi Nakanishi*, 259–269.

Vander Wall, S. B., and Balda, R. P. (1977). Coadaptations of the Clark's Nutcracker and the piñon pine for efficient seed harvest and dispersal. *Ecol. Monogr.* 47 (1), 89–111. doi: 10.2307/1942225

Yamanoi, T. (1992). Palyno-flora of middle Miocene sediments of Okushiri Island, southwest Hokkaido. *Japanese J. Palynology* 38 (2), 106–115.

Yano, M. (1972). On the plant remains from the fossil elephant bearing bed in Tokachi Plain, Hokkaido. *Assoc. Geological Collaboration Japan /Chikyu Kagaku* 26 (1), 12–19.

Young, A., Boyle, T., and Brown, T. (1996). The population genetic consequences of habitat fragmentation for plants. *Trends Ecol. Evol.* 11 (10), 413–418. doi: 10.1016/0169-5347(96)10045-8

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fpls.2022.1000489

OPEN ACCESS

EDITED BY
Rong Wang,
East China Normal University, China

REVIEWED BY
Zhen-Hua Zhang,
China National Rice Research Institute (CAAS), China
Junyin Deng,
East China Normal University, China

*CORRESPONDENCE
Kean-Jin Lim
keanjin.lim@zafu.edu.cn
Zhengjia Wang
wzhj21@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 22 July 2022
ACCEPTED 12 December 2022
PUBLISHED 04 January 2023

CITATION
Jin H, Yang Z, Luo J, Li C, Chen J, Lim K-J and Wang Z (2023) Comprehensive identification and analysis of circRNAs during hickory (*Carya cathayensis* Sarg.) flower bud differentiation.
Front. Plant Sci. 13:1000489.
doi: 10.3389/fpls.2022.1000489

Comprehensive identification and analysis of circRNAs during hickory (*Carya cathayensis* Sarg.) flower bud differentiation

Hongmiao Jin†, Zhengfu Yang†, Jia Luo, Caiyun Li, Junhao Chen, Kean-Jin Lim* and Zhengjia Wang*

State Key Laboratory of Subtropical Silviculture, College of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou, Zhejiang, China

Flower bud differentiation represents a crucial transition from vegetative growth to reproductive development. *Carya cathayensis* (hickory) is an important economic species in China, with a long juvenile period that hinders its commercial development. In recent years, circular RNAs (circRNAs) have been widely studied and identified as sponges for miRNA regulation of mRNA expression. However, little is known regarding the role of circRNAs in flower buds. In this study, we sequenced circRNAs at three developmental stages (undifferentiated, differentiating, and fully differentiated) in both female and male buds. A total of 6,931 circRNAs were identified in the three developmental stages and 4,449 and 2,209 circRNAs were differentially expressed in female and male buds, respectively. Gene ontology demonstrated that many circRNA host genes participated in various processes, for example, cellular and intracellular pH regulation. Function annotation identified 46 differentially expressed circRNAs involved in flowering regulation, with 28 circRNAs found only in female buds, 4 found only in male buds, and 11 found in both female and male buds. A circRNA-miRNA-mRNA network was predicted based on 13 flowering-related circRNAs and their seven putative interacting miRNAs to describe the regulatory mechanism. Our preliminary results demonstrated a potential involvement of circRNA in bud differentiation. They provided a preliminary theoretical basis for how circRNA might participate in flower development in hickory, perhaps in woody plants.

KEYWORDS

hickory, flowering, ceRNA network, circular RNA, circRNA

1 Introduction

Hickory (*Carya cathayensis* Sarg.) is a woody plant species with prominent economic value for its nuts and oil. It is mainly distributed in the Tianmu Mountains, at the junction of Zhejiang and Anhui provinces (Yang et al., 2015). It is a monoecious tree with a long juvenile period and different developmental times for the male and female flowers. The female flower buds differentiate in mid-March and then develop into flower organs in mid-April. By contrast, vegetative growth of the male flower buds occurs in mid-late April, with differentiation completed in early May (Huang et al., 2006; Huang et al., 2007; Huang et al., 2013). This type of development is known as heterodichogamy, a mechanism that avoids inbreeding, and is common in 21 genera in 13 families, such as *Acer* in the Aceraceae and *Cyclocarya* and *Juglans* in the Juglandaceae (Kikuchi et al., 2009). However, this process results in low pollination, and therefore low fruit bearing rates, which hinders the hickory industry (Fukuhara and Tokumaru, 2014; Chen et al., 2019). Traditional breeding methods, such as natural variation selection and cross-breeding, have been applied to shorten the juvenile period and promote fruit-bearing. Nevertheless, hybrid breeding usually takes 5-10 years, and the effects are usually unsatisfactory. Therefore, understanding the regulatory mechanisms of male and female flower development at the molecular level might aid in solving some of these breeding problems.

The process of flowering has received intensive study. To date, the external and internal factors understood to regulate flower development fall into five major pathways associated with photoperiod, vernalization, autonomous, gibberellin (GA), and sucrose (Roldán et al., 1999; Mouradov et al., 2002). Transcription factors (TFs) have been reported to participate in at least two of these pathways and are called integrators of flowering regulation. In *Arabidopsis*, *FLOWERING LOCUS C (FLC)*, a central repressor of flowering first identified in the primordial cell, promotes the formation of the floral meristem (FM) and influences flowering by vernalization (Whittaker and Dean, 2017). *SHORT VEGETATIVE PHASE (SVP)*, which belongs to MADS-box family, interacts with *FLC* and binds to the CArG motif of the *FLOWERING LOCUS T (FT)* promoter to suppress *FT* expression (Jeong et al., 2007).

In addition to TFs, non-coding RNAs (ncRNAs) also play a role in flower regulation. For example, the vernalization-mediated epigenetic repression of *FLC* requires a long intronic noncoding RNA (lncRNA), COLD ASSISTED INTRONIC NONCODING RNA (COLDAIR) (Heo and Sung, 2011). Similarly, COOLAIR, another lncRNA identified in *Arabidopsis*, arises from the 3' end of *FLC* in an antisense direction relative to *FLC* (Chekanova, 2015). Both lncRNAs repress the expression of *FLC via* an epigenetic mechanism to regulate flowering in *Arabidopsis* (Šťorchová, 2017). *SVP* changes the expression of one of the ambient temperature-responsive miRNAs, micro RNA172 (miR172), and a subset of its target genes (Lee et al., 2010).

Research has shown that only 2% of the transcripts of a genome are translated into protein. A vast amount of the transcriptome is ncRNAs, including ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), micro RNAs, short interfering RNAs (siRNAs), circular RNAs (circRNAs), and long-coding RNAs (Jandura and Krause, 2017). At present, functional studies of ncRNAs have mainly been conducted in the medical field, largely in cancer studies (Mattick and Makunin, 2006; Wang et al., 2014). Research on ncRNAs in plants has received relatively little attention until recently. Substantial research has now demonstrated that ncRNAs play an indispensable role in plant functions, including responses to abiotic stress, plant development, and fruit development (D'Ario et al., 2017; Correa et al., 2018; Tong et al., 2018; Waititu et al., 2020; Gelaw and Sanan-Mishra, 2021). For example, overexpression of osa-miR5506, a miRNA, resulted in pleiotropic abnormalities, including defects in ovary development, confirming a vital role for osa-miR5506 in regulating floral number and female gametophyte production (Chen et al., 2021). Similarly, circRNAs were revealed to act as miRNA sponges and to form competing endogenous RNA (ceRNA) networks to inhibit miRNA activities (Hansen et al., 2013; Zhang et al., 2017a).

CircRNAs were first discovered and experimentally verified in the 1970s. They were first considered a by-product of mistaken translation, but are now regarded as a class of covalently closed-loop RNAs characterized by their 3' end and 5' ends (Nigro et al., 1991). Four mechanisms are proposed to explain the generation of circRNA: back-splicing, intron-driven cyclization of complementary sequences, complementary cyclization driven by different introns from the same single gene, and exon cyclization regulated by RNA binding proteins (Sun et al., 2016). CircRNAs can be produced from exons, introns, or intergenic regions, but those produced from exons have received the most attention (Li et al., 2018).

The functions of circRNAs include transcriptional regulation, miRNA sponging, and translation into proteins. Some circRNAs may influence biological processes by regulating the expression of their parental genes (Zhou et al., 2020). For example, in *Arabidopsis*, circSEP3, which is derived from an exon of *SEPALLATA3*, regulates the splicing of its cognate mRNA by forming R-loop to affect floral development (Conn et al., 2017). In tea plants, the abundances of circRNAs were positively correlated with the mRNA transcript level of their parental genes and were considered to play a role in leaf development (Tong et al., 2018). In rice, overexpression of a linear Os08circ16564 construct reduced the expression of its parental gene in the leaf and panicle (Lu et al., 2015). By contrast, the circRNAs that act as sponges must possess rich miRNA binding sites or show high expression in the cytoplasm (Ashwal-Fluss et al., 2014). At present, only a few circRNAs from plants (5% in Arabidopsis and 6.6% in rice) have been demonstrated to contain miRNA binding

sites (Ye et al., 2015). In Arabidopsis, circRNA biogenesis is altered by heat stress, leading to the suggestion that circRNAs may participate in heat stress responses through circRNA-mediated ceRNA networks. Recent research on rice has identified 11 circRNAs that were predicted to act as miRNA sponges that functioned in flag leaf senescence through the formation of circRNA-miRNA-mRNA ceRNA networks (Pan et al., 2018; Zhou et al., 2020; Huang et al., 2021). In *Brassica campestris*, a ceRNA and miRNA-mRNA network containing the circRNA A02:23507399|23531438 was hypothesized to act as a miRNA sponge for the mRNAs unconservative_A06_21945 and unconservative_Scaffold000096_42992 to regulate the expression of Bra002275 and the biosynthesis of tryphine and sporopollenin (Liang et al., 2019).

In the present study, we explored the role of circRNAs in regulating hickory flower development by examining six circRNA libraries obtained from the transcriptome of female and male flowers. At different developmental stages, we observed different circRNA expression profiles. After annotating and predicting the target miRNAs, we proposed possible circRNA-miRNA-mRNA regulatory networks that could be involved in hickory flower development. Our preliminary results shed light on the potential role of circRNAs in hickory flower bud development.

2 Materials and methods

2.1 Plant materials and sample collection

Male and female flower buds were collected from 15-year-old asexually propagated hickory trees growing in the nursery orchard of Zhejiang A&F University (lat. 30°15'N, long. 119° 43'E), Zhejiang Province, China. Female flower buds were collected at the F1 undifferentiated stage (early March, 2016), the F2 differentiating stage (late March, 2016), the F3 fully differentiation stage (April, 2016). Male flower buds were collected at the M1 undifferentiated stage (April, 2016), the M2 differentiating stage (May, 2016) and the M3 fully differentiation stage (June, 2016) (Huang et al., 2006; Huang et al., 2007; Huang et al., 2013). All samples were immediately immersed in liquid nitrogen and stored at -80 °C until RNA extraction.

2.2 Total RNA extraction, library construction, and sequencing

Total RNA was isolated from female and male hickory buds at different developmental stages using a modified CTAB method (Lim et al., 2016) combined with TRIzol reagent (Invitrogen, Grand Island, NY, USA). A total of 3 μg of total RNA per sample was used as the starting material for the RNA sequencing (RNA-seq) libraries preparation. Ribosomal RNA depletion and RNA-seq library preparation were performed as described in Li and colleagues (Li et al., 2022).

A total of six libraries were sequenced on an Illumina Hiseq 2500 platform. After sequencing, the low-quality and adapters sequences were removed from the raw data using NGS QC Toolkit 2.3.3 software (Patel and Jain, 2012) to obtain clean data.

2.3 Identification of circRNAs

For circRNA recognition, find_circ and CIRI2 software were used to identify circRNA from female and male flower buds (Memczak et al., 2013; Gao et al., 2018). The find_circ utilized bowtie2 reference matching to extract 20-nt anchor sequences as a seed sequence from each read end that match the reference sequence. Each pair of anchor sequences was compared to the reference sequence again. If the 5' end of the anchor sequence matched the reference sequence (A3 and A4 for the start and end sites, respectively), and the 3' end of the anchor sequence matched upstream of that site (A1 and A2 for the start and end sites, respectively), and a splice site was present between A2 and A3, then that read was identified as a candidate circRNA. The candidate circRNA with a read count greater than or equal to two was used as the identified circRNA. CIRI2 searched for a paired chiastic clipping (PCC) signal, and a paired end mapping (PEM) signal, with a GT-AG signal first. It then filtered the candidate circRNAs based on the global comparison, the reads support of circRNAs, and the annotation information to identify junction reads. Those with a read number larger than two were selected as the identified circRNA (Gao et al., 2015; Ye et al., 2015; Tong et al., 2018; Zhang et al., 2019; Li et al., 2020; Philips et al., 2020; Yang et al., 2020; Jiang et al., 2021). The results that appeared in both find_circ and CIRI2 were selected as candidate circRNAs (Xu et al., 2018; Liu et al., 2019; Li et al., 2020; Philips et al., 2020; Yang et al., 2020).

2.4 Bayesian hierarchical clustering analysis

The newly identified circRNAs were mapped against the hickory genome (Huang et al., 2019), and the mapped count tables of female and male flower bud libraries were obtained using featureCounts software (version 1.20.6, Liao et al., 2014). The mapped count tables were then loaded into the edgeR (version 3.2.4, Robinson et al., 2009) R session (version 3.5.0). The raw counts were normalized using the trimmed mean of M-value method to obtain normalized counts per million (CPM). The average CPM and relative mean CPM counts were calculated, and the expression level of circRNA was set to 1, as described by Lim et al. (2021). Bayesian hierarchical clustering analysis was carried out using Spline Cluster (version 2002,

Heard et al., 2006), with the default parameters, except that prior precision and normalize targets were set to 1 e-45 and 0, respectively. Clusters were visualized using SplineCluster (Lim et al., 2021).

2.5 Functional enrichment analysis

An expression level of circRNA greater than or equal to two CPM was defined as differential expression in both female and male flower bud data sets (Singh et al., 2014). The differently expressed (DE) circRNAs of female and male flower buds were subjected to subsequent analysis. We performed GO enrichment analysis for the host genes of DE circRNAs to better understand how circRNA may be involved in flower development. GOseq (version 2.12) was used to annotate the function of the parent genes of the differentially expressed circRNAs' host genes (Young et al., 2010) using the Wallenius non-central hypergeometric distribution method (Kang and Liu, 2015; Bedre et al., 2019; Lipka et al., 2019; Li et al., 2021). The Benjamini Hochberg method was used to correct the p-value, with a smaller value being more significant. To further understand how circRNAs may be involved in bud differentiation, we proposed a regulatory network of flowering-related genes and combined the expression profiles of circRNAs characterized by different expressions in the three differentiation stages.

2.6 Prediction of target miRNAs site

To explore the regulatory role circRNAs might play as ceRNAs, the functions circRNA in bud differentiation were examined by predicting the target miRNAs of differentially expressed circRNAs in undifferentiated and differentiated stages and the corresponding mRNAs of miRNA. The miRNA binding sites of the circRNA were predicted using miRanda (version 3.3a, Enright et al., 2003). The predicted results of miRNA binding sites were used to deduce the circRNA-miRNA regulatory networks using Cytoscape software (version 3.8.2) with default parameters (Shannon et al., 2003).

2.7 Real-time quantitative RT-PCR (qRT-PCR) analysis

The female hickory flower buds corresponding to undifferentiated, differentiating, and fully differentiated stages were collected in early March, late March, and April, 2021, respectively. The male flower buds at the same differentiation stages were collected in April, May, and June 2021, respectively. Total RNA was isolated as described above. A total of nine flowering-related circRNAs were selected from the differentially expressed circRNAs for RNA-seq verification. The hickory histone sequence was used as an internal reference gene. The relative expression level was calculated using the $2^{-\Delta Ct}$ method (de Vos et al., 2017). All primers were designed using primer3plus (https://primer3plus.com/) and were listed in Supplementary Table S13.

3 Results

3.1 Identification of circRNAs in female and male hickory flower buds

The regulatory mechanism of circRNA during hickory flower bud differentiation was revealed using RNA-sequencing (RNA-seq) to mine circRNAs in female (F) and male (M) flower buds at different developmental stages. All the acquired transcriptomes were divided into six libraries: F1 and M1 for the undifferentiated stage, F2 and M2 for the differentiating stage, and F3 and M3 for fully differentiated stage. Mapping to the hickory genome, as shown in Supplementary Table S1, results showed that, averagely the male flower bud libraries had a higher unique mapping rate than the female flower bud libraries.

Between the three developmental stages, an average of 7,472 and 4,381 circRNAs were identified in female and male flower buds. Among all the identified candidate circRNAs, 83.63% in the female buds and 84.19% in the male buds were from exons (Supplementary Table S3). Our results showed that most of the candidate circRNAs were from exons, with only about 5% derived from introns in both female and male flower buds (Supplementary Figure S1A). We noticed that over 85% of the circRNAs were shorter than 5000 nt, and nearly 1% were between 5000 nt and 10000 nt in length. The number of circRNAs shorter than 500 nt peaked in all six circRNA libraries. In general, 60.34% of circRNAs were less than 1500 nt in length (Supplementary Figure S1B).

Chromosomal localization analysis of the candidate circRNA showed that contig232, contig233, contig234, contig241, contig244, and contig264 transcribed the largest number of circRNAs in female buds. By contrast, the largest number of circRNAs transcribed in male buds were in contig232, contig233, contig234, contig237, contig241, and contig244 (Figure 1).

3.2 The expression profiles of circRNAs in developing hickory flowers

A total of 6,931 circRNAs were identified (Supplementary Figure S2). The 6,172 circRNAs in female buds and 3,349 circRNAs in male buds contained mapped counts. Further analysis of these circRNAs demonstrated the expression of 2,336 circRNAs in F1, 5,145 in F2, and 2,280 in F3 during the

10.3389/fpls.2022.1000489

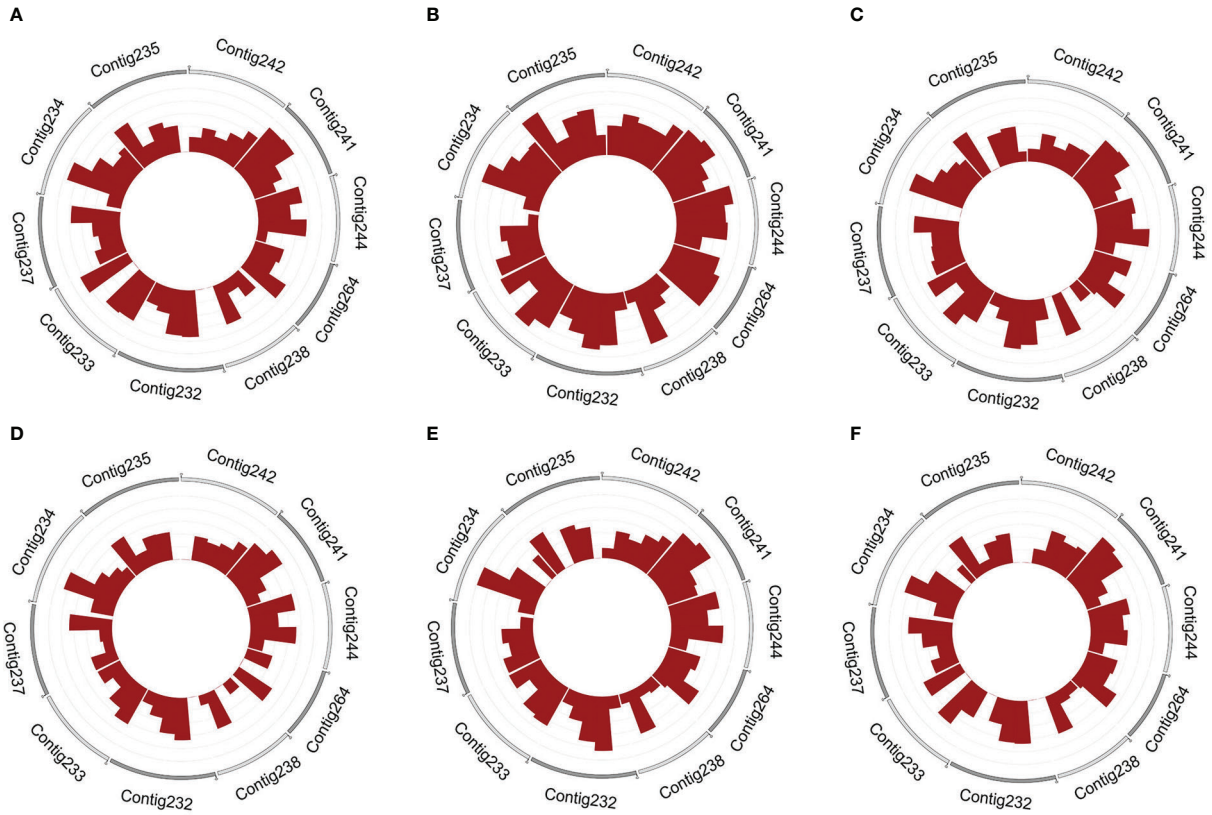


FIGURE 1
The density of circRNAs in female and male flower buds on different contigs. The 10 most densified distributed contigs in female and **(A–C)** and male **(D–F)** flower buds, respectively. The outer ring takes 10 contigs for display. A contig with more rings has a higher density of circRNAs.

development of female flower buds. Of these, 484 circRNAs were only expressed in F1, while 636 circRNAs participated in the development process of female flower buds. During the development of male flower buds, a total of 1,434 circRNAs were expressed during the M1 stage, 2,229 circRNAs were involved in the flower bud differentiation process during the M2 stage, and 1854 circRNAs were expressed during the M3 stage. A total of 697 circRNAs were involved in the whole progress of male flower bud differentiation and development (Supplementary Tables S4, S5).

We performed clustering analysis with the Bayesian approach to reveal the expression profile of 6,931 circRNA of female and male flower buds using CPM counts. The clustering analysis showed that circRNAs in female flower buds were divided into 20 clusters (Figure 2A) with several expression profiles. In the first profile, in clusters 2-6, the expression was a constitutively increasing expression from F1 to F3, whereas clusters 7-9, 12, 14, and 20 showed peak expression in F2. By contrast, clusters 10 and 11 had their lowest expression in F2. CircRNAs in clusters 13 and 15-19 were consistently downregulated. Only circRNAs in cluster 1 had a relatively constant expression profile, and they may be involved in maintaining basic biological functions. The 4,449 differentially expressed circRNAs (CPM >2) in female flower buds were distributed in clusters 2-4 and 17-20 (Supplementary Table S6).

Clustering analysis also grouped the circRNAs of male flower buds into 15 clusters (Figure 2B).Upregulation profiles across all stages were observed in clusters 2-3 and 12-13, whereas clusters 4, 9-10, and 15 shared a similar profile, with the peak expression at the M2 stage, and clusters 5 and 1 shared a similar profile with peak expression at M2. Clusters 5 and11 had the opposite profile, showing the lowest expression at M2. Clusters 6-8 and 14 shared the downregulation profiles. The circRNA in cluster 1 shared the same profile as that in the female flower. A total of 2,209 circRNAs (CPM >2) were differentially expressed during male bud development and were distributed in clusters 2, 7, and 13-15 (Supplementary Table S7).

3.3 Enrichment analysis of differentially expressed circRNA host genes

Female flower bud circRNAs in cluster 2 were enriched with biological process terms including 1,3-beta-D-glucan synthase

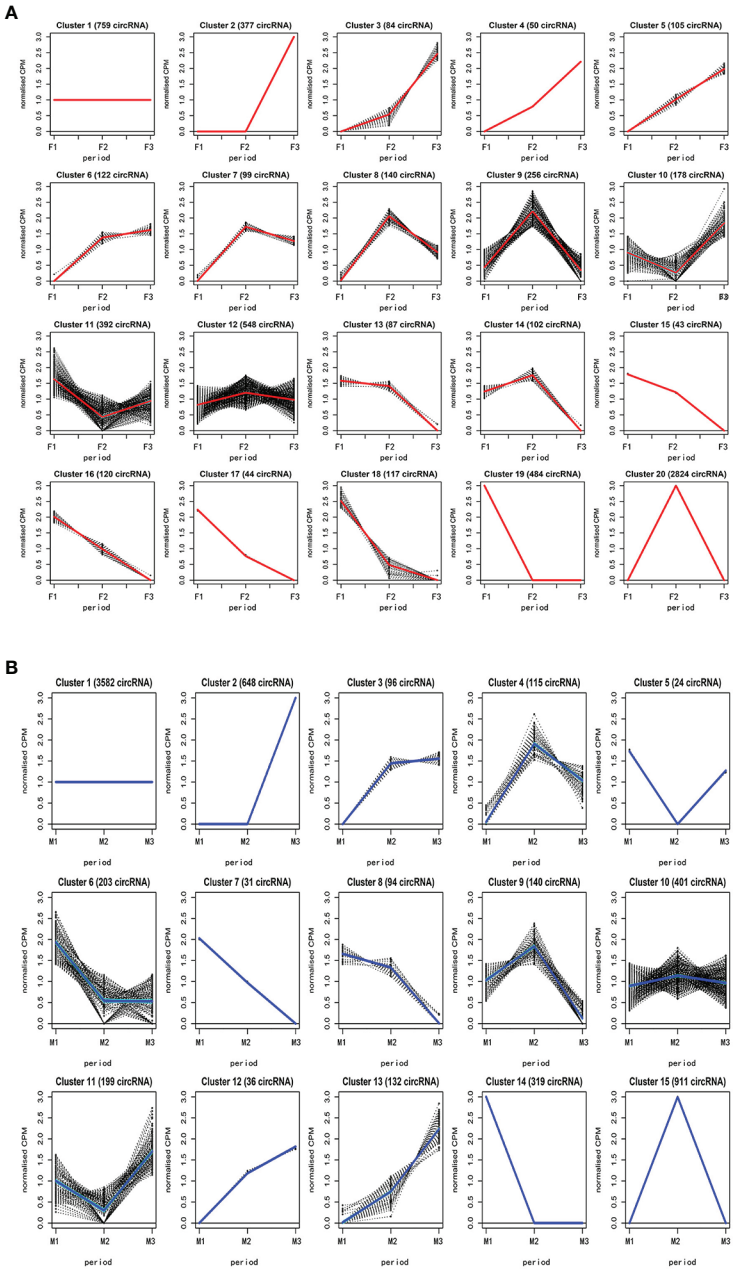


FIGURE 2
The hierarchical clustering analysis of circRNAs in female and male buds. The circRNAs in female hickory flower buds were divided into 20 clusters **(A)** while the circRNAs of male flower buds were grouped into 15 clusters **(B)**. The analysis was performed with the Bayesian approach to reveal the expression profile of circRNAs in female and male bud using relative mean counts per million (CPM) counts. F, female; M, male; F1 and M1, undifferentiated stage; F2 and M2, differentiating stage; F3 and M3, fully differentiated stage.

complex (GO:0000148), 1,3-beta-D-glucan synthase activity (GO:0003834), (1->3)-beta-D-glucan metabolic process (GO:0006074), and (1->3)-beta-D-glucan biosynthetic process (GO: 0006075). Positive regulation of the biological process (GO:0048518) was also enriched. CircRNAs in cluster 4 were enriched by many molecular function terms, including thiol-dependent ubiquitin-specific protease activity (GO:0004843), ubiquitin-like protein-specific protease activity (GO:0019783), cysteine-type peptidase activity (GO:0008234), and serine-type endopeptidase activity (GO:0004252). Several biological processes, such as protein phosphorylation (GO:0006468) and peptidyl-serine modification (GO:00182095),

were enriched in cluster 17. Terms enriched in cluster 19 were diverse and included purine nucleotide binding (GO:0001883), guanyl nucleotide binding (GO:0019001), and mRNA metabolic processes (GO:0016071). Terms in cluster 20 were more complex (Supplementary Table S8).

In male flower buds, cellular components, such as vacuolar membrane (GO:0005774), vacuolar part (GO:0044437), and vacuolar organization (GO:0007033), were enriched in cluster 2. Terms concerning beta-D-glucan, such as GO:0000148, GO:0003834, GO:0006074, and GO:0006075, were also enriched. Interestingly, phosphate-related terms, including positive regulation of phosphorylation (GO: 0042327 and positive regulation of phosphate metabolic processes (GO: 0045937), were enriched in cluster 2. Cluster 13 was enriched with ubiquitin-related terms, such as ubiquitin-like protein-specific protease activity (GO: 0019783), thiol-dependent ubiquitinyl hydrolase activity (GO:0036459), and ubiquitinyl hydrolase activity (GO:0101005). In cluster 15, we noted that regulation of cellular pH (GO:0030641) and regulation of intracellular pH (GO:0051453) were enriched (Supplementary Table S9).

3.4 Functional analysis of differentially expressed circRNAs

Due to the imperfection of the reference genome, a large number of circRNAs were not annotated. Therefore, we first performed a functional homology search on the obtained circRNAs that were not annotated in the genome. The results identified a total of 46 circRNAs that may be involved in female and male flowering processes. In general, more circRNAs involved in flowering were identified in female flowers; four were specifically expressed in male buds and 11 were expressed in both female and male buds (Supplementary Table S10). Our results showed that certain circRNAs were derived from *MADS-box* genes. Novel_circ_0002474 was predicted to be derived from *SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1)*, and novel_circ_0005951, novel_circ_0005952 together with novel_circ_0005953 were thought to be products of *GIGANTEA (GI)*. Novel_circ_0004945 was from a host gene encoding the PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1 (PIE1) protein. Four circRNAs were derived from *HISTONE MONOUBIQUITINATION1* (*HUB1)*; only novel_circ_0005820 was found in both female and male buds. Novel_circ_0005815 and novel_circ_0005823 were from the same parental gene, *HUB1*. Their similar expression profiles suggested that they may be functionally redundant. The *GI*-related circRNAs were novel_circ_0005951, novel_circ_0005952, and novel_circ_5953, and all three circRNAs were differentially expressed in F1 and F2 in the female buds. Their expression profiles hint that they may be functionally redundant. Novel_circ_0008461 and novel_circ_0008442 were from genes encoding SALT INDUCED ZINC FINGER PROTEIN1 (SIZ1), Novel_circ_0008461 had significantly rich expression in the third stage in flower buds (Figure 3).

3.5 Interaction of differentially expressed circRNA-miRNA-mRNA networks

Studies have shown that circRNAs can perform regulatory functions by adsorbing specific miRNAs that function in multiple life processes (Olesen and Kristensen, 2021). We identified 1,397 differentially expressed circRNAs targeting 17 miRNAs and regulating 22 mRNAs in the F1 compared with the F2 stages. By contrast, comparison of the M1 and M2 stages revealed 53 differentially expressed circRNA targeting two miRNAs with two mRNAs. The process was more complex in female buds, as 135 circRNAs that interacted with miR169i and miR169r were only present in female buds. (Lee et al., 2010) We noted that 127 circRNAs interacted with pct-miR399f, and 235 circRNAs interacted with pct-miR396e-3p; these are involved in plant development and were found only in female buds (Supplementary Tables S11, 12). (Liebsch and Palatnik, 2020) Our regulatory network (Figure 4) using the differentially expressed circRNA and targeted miRNAs predicted that the axes novel_circ_novel_0002857-novel_23-CCA0686S0035/CCA1456S0031/CCA0883S0013/CCA0903S0022 presented up-down-up trends, suggesting that novel_circ_0002857 may act as the ceRNA of novel_23. Similarly, novel_circ_0010874 may have the same function as novel_circ_0002857, although they are derived from different genes. Both novel_circ_0005823 and novel_circ_0005105 have binding sites for ptc-miR167e targeting CCA1519S0026, and the expression profile indicated a putative sponge role for these two circRNAs.

3.6 Real-time quantitative RT-PCR (qRT-PCR)

We randomly selected nine circRNAs and compared their expressions to those determined by RNA-seq. The qRT-PCR results were in good agreement with the RNA-seq results (Figure 5, Supplementary Table 14).

4 Discussion

4.1 Identification and characterization of circRNAs in hickory floral development

In this study, we identified 6,931 circRNAs in hickory flowers by RNA-seq. Many circRNAs have been observed in other plant species, such as *Arabidopsis thaliana* (6,012), *Oryza sativa* (12,307), and *Zea mays* (1,199) (Ye et al., 2015; Han et al.,

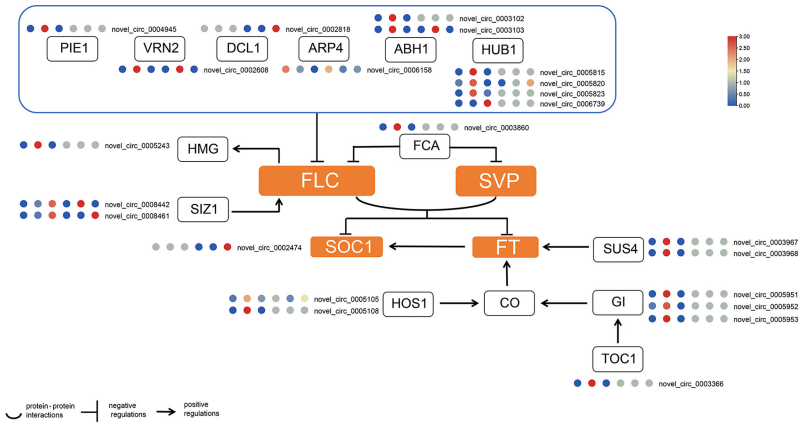


FIGURE 3
Transcript abundance of flowering-related circRNAs in hickory buds. The circRNA expression level was presented in relative mean counts per million (CPM). The six plots from leftmost to the rightmost represent F1, F2, F3, M1, M2 and M3 respectively. The scale bar is the relative mean of CPM. The square denotes genes or proteins, while the orange represents integrator gens or proteins. *PEI1, PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1; HUB1, HISTONE MONOUBIQUITINATION1; VRN2, VERNALIZATION 2; DCL1, DICER-LIKE 1; ABH1, ABA HYPERSENSITIVE 1; ARP4, ACTIN-RELATED PROTEIN 4; HMG, HIGH MOBILITY GROUP; FCA, FLOWERING CONTROL LOCUS A; SIZ1, SALT INDUCED ZINC FINGER PROTEIN1; FLC, FLOWERING LOCUS C; SVP, SHORT VEGETATIVE PHASE; SOC1, SUPPRESSOR OF OVEREXPRESSION OF CO 1; FT, FLOWERING LOCUST T; SUS4, SUCROSE SYNTHASE 4; HOS1, HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENES 1; CO,CONSTANS; GI, GIGANTEA; TOC1,TIMING OF CAB EXPRESSION 1.*

2020). Our characterization of the circRNAs revealed more circRNAs in female buds. In both female and male buds, circRNAs from exons had a more prominent role, as observed in many other plants (Wang et al., 2022), suggesting that exonic circRNAs may serve as regulators in cells (Li et al., 2015). Of the differentially expressed circRNAs, 3,538 were explicitly expressed in female buds, while 1,298 circRNAs were only expressed in male buds. Female and male buds shared 911 expressed circRNAs. These results support that circular RNAs in female shoots might be involved in more diverse biological processes.

4.2 Functional analysis of differently expressed circRNA

The parental genes of circRNAs in clusters 2 to 4 in female buds generally presented an upregulation trend. Interestingly,

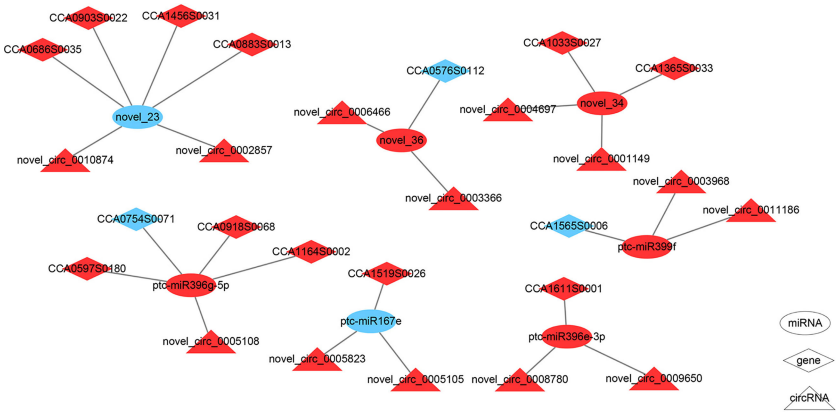


FIGURE 4
The circRNA-miRNA-mRNA regulation network in F1 compared with F2. The ellipses, diamonds, and triangles denote miRNAs, genes, and circRNAs, respectively. The red denotes upregulation and the blue denotes downregulation.

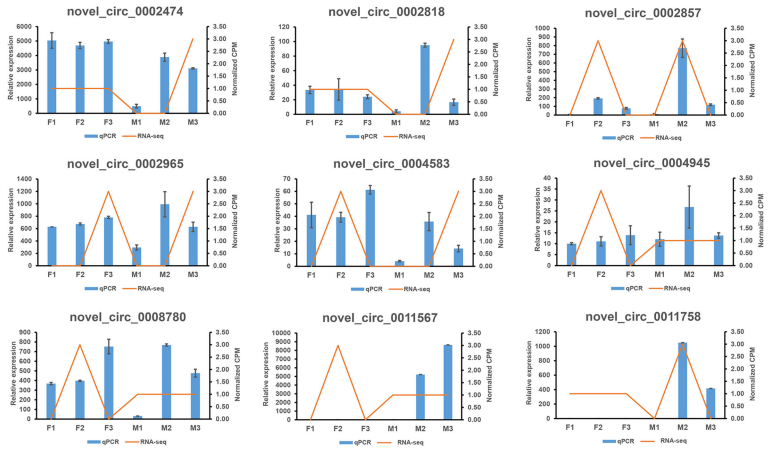


FIGURE 5
The comparison of real-time quantitative RT-PCR (qRT-PCR) and RNA sequencing (RNA-seq) for 9 circRNAs. The two results were in good agreement. The blue columns represented the results of the q RT-PCR, and the orange lines denoted for the results of the RNA-seq.

the most enriched GO terms were positive regulators of biological processes (GO:0048518), suggesting that these circRNAs might contribute to cell development (Ishikawa et al., 1995). Cluster 20 showed peak expression in F2, with enrichment terms of ATPase activity (GO:0016887), nuclear transport (GO:0051169), enzyme binding (GO:0019899), and protein import (GO:0017038). The circRNAs generated from these parental genes may be involved in cell growth (Du et al., 2015). In male flower buds, members of cluster 2 showed expression beginning in M2 and were enriched with a batch of glucan-related GO terms, including 1,3-beta-D-glucan synthase complex (GO:0000148), 1,3-beta-D-glucan synthase activity (GO:0003834), (1->3)-beta-D-glucan metabolic process (GO:0006074), and (1->3)-beta-D-glucan biosynthetic process (GO:0006075). The results indicated possible activation of wall assembly by circRNAs (Yoshimi et al., 2017). Furthermore, members in cluster 15 had high expression in M2 but were silent in M1 and M3. We noted that regulation of cellular pH (GO:0030641) and regulation of intracellular pH (GO:0051453) were enriched in this cluster. Intracellular pH control is vital for many cellular behaviors, such as enzyme activity, protein degradation, and organelle activities (Ishaque and Al-Rubeai, 1998).

We traced circRNAs back to the parental genes and found that one parental gene can generate multiple circRNAs that can exhibit the same or different expression profiles. In this study, we noted that *SUS4*, *GI*, *HUA1*, and *HOS1* generated several circRNAs, while *HUB1* produced four circRNAs, with different expression profiles. Novel_circ_0005815 and novel_0005823 peaked in F2, while novel_circ_0006739 and novel_circ_0005820 peaked in F3. The different expression profiles indicated that they might have different functions (Conn et al., 2017).

4.3 The circRNA-miRNA-mRNA network in hickory floral development

CircRNA has been identified as a member of ceRNAs due to its abundance of conserved miRNA binding sites or miRNA response elements (MREs). Studies have illustrated that circRNA may act as ceRNAs (Cortés-López and Miura, 2016; Zhang et al., 2017b; Zhong et al., 2018). The ceRNA hypothesis proposes that RNAs sharing multiple MREs could have efficient crosstalk, forming a large-scale network in the transcriptome (Salmena et al., 2011). Therefore, we predicted the occurrence of different interactions among circRNA-miRNA-mRNA in F1 than in F2 and in M1 compared with M2. Among all the differentially expressed flower-related circRNAs, we only obtained 13 circRNAs, seven miRNA targets, and 14 mRNAs, all in F1 and F2. In our study, the expression level of circRNAs, such as novel_circ_0002857, novel_circ_0005105, and novel_circ_0005823, increased during female bud development, while their target miRNAs decreased (Figure 4, Supplementary Tables S11 and S12), thereby leading to the promotion of targeted mRNAs. In brief, these circRNAs might act as miRNA sponges to regulate the expression of target mRNAs. The targeted miRNAs include miR169, miR396 and miR399 all were thought to play a role in flowering regulation (Lee et al., 2010; Xu et al., 2016; Liebsch and Palatnik, 2020). Recent research has demonstrated that these correlations may result from co-expression or mutual exclusivity in subpopulations in complex tissues; therefore, the expression profiles need confirmation (Zhang et al., 2019).

The results of expression profile analysis, host gene annotation, and circRNA-miRNA-mRNA prediction indicate that, among all four circRNAs derived from *HUB1*, only novel_circ_0005823 was predicted to have a miRNA binding

site. Novel_circ_0005815 was derived from cluster 20 and had the highest expression among the four circRNAs, while novel_circ_0005823 was from cluster 9. Hence, we hypothesized that these circRNAs might play roles in different mechanisms. The other putative sponge of miR167 was novel_circ_0005105, which was grouped in cluster 9, while its sibling, novel_circ_0005108, was clustered in cluster 20 and was predicted to have a pct-miR396g-5p binding site. The target of miR167 was CCA1519S0026, annotated as a putative FT-like gene by NCBI. We hypothesized that novel_circ_0005823 and novrl_circ_0005105 may act as sponges for miR167, thereby contributing to the abundance of CCA1519S0026. We proposed that novel_circ_0003968 and novel_circ_0011186 promote the expression of ptc-miR399, which regulates LNC_02115 in the female hickory bud during temperature changes (Li et al., 2022). Nevertheless, more detailed mechanisms may also exist that need further experimental exploration.

5 Conclusions

We collected and constructed transcriptome datasets of female and male flower buds at three different developmental stages. From the transcriptome datasets, we identified 6,931 circRNAs in hickory buds. Characterization analysis indicated that most circRNAs were derived from exons and were less than 5000 nt in length. Expression analysis revealed that most circRNAs were expressed in female buds. In total, 4,449 and 2,209 differentially expressed circRNAs were identified in female and male buds, respectively. We studied the flowering-related candidates from the differentially expressed circRNA host genes and predicted the miRNA binding sites. Based on the ceRNA theory, we noted that novel_circ_0005823 and novel_circ_0005105 might target miR167 and regulate CCA15190026 to influence flowering. We also observed that novel_circ_0002857 and novel_circ_0010874 might serve as miRNA sponges of novel_23, thereby influencing flowering. Our preliminary results shed light on how circRNA might involve in the hickory flower development, perhaps in woody plants.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA820165.

Author contributions

JL, ZW, and K-JL conceived and designed this study. HJ, JL, JC, and K-JL performed research and analyzed data. HJ and JL wrote the manuscript. K-JL and ZY edited and reviewed the writing. ZW and ZY acquired funding. All authors have read and agreed to the published version of the manuscript.

Funding

Key Scientific and Technological Grant of Zhejiang for Breeding New Agricultural Varieties (2021C02066-12), Key research and development project of Zhejiang Province (2021C02054), Ministry of Science and Technology High-End Foreign Expert Introduction Program (G2021016035L), Research and Development Fund of Zhejiang A&F University (W20190248).

Acknowledgments

We thank Tao Qin for the technical assistance on this work.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1000489/full#supplementary-material

SUPPLEMENTARY TABLE 1
Linear alignment of reference genome.

SUPPLEMENTARY TABLE 2
Identification of circRNA in female and male buds.

SUPPLEMENTARY TABLE 3
The genomic feature of identified circRNAs.

SUPPLEMENTARY TABLE 4
The expression of circRNAs in female buds.

SUPPLEMENTARY TABLE 5
The expression of circRNAs in male buds.

SUPPLEMENTARY TABLE 6
The clusters of circRNAs in female buds.

SUPPLEMENTARY TABLE 7
The clusters of circRNAs in male buds.

SUPPLEMENTARY TABLE 8
Gene ontology analysis in female buds.

SUPPLEMENTARY TABLE 9
Gene ontology analysis in male buds.

SUPPLEMENTARY TABLE 10
Flower-related differentially expressed circRNAs.

SUPPLEMENTARY TABLE 11
The prediction of circRNA-miRNA-mRNA in F1 compared with F2.

SUPPLEMENTARY TABLE 12
The prediction of circRNA-miRNA-mRNA in M1 compared with M2.

SUPPLEMENTARY TABLE 13
Primers for qRT-PCR.

SUPPLEMENTARY TABLE 14
The original data and calculation of qRT-PCR.

SUPPLEMENTARY FIGURE 1
Characterizations of circRNA in female and male flower buds. **(A)** The distribution of circRNAs. The circRNAs derived from exons were indicated in red, and those from intergenic and introns were shown in green and blue, respectively. **(B)** The circRNA length distribution of six libraries. The female (F) and male (M) bud samples were collected at undifferentiated (F1, M1), differentiation (F2, M2), and differentiation completed (F3, M3) stages.

SUPPLEMENTARY FIGURE 2
The flowchart of circRNA identification and analysis.

References

Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., et al. (2014). CircRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* 56, 55–66. doi: 10.1016/j.molcel.2014.08.019

Bedre, R., Irigoyen, S., Schaker, P. D.C., Monteiro-Vitorello, C. B., Da Silva, J. A., and Mandadi, K. K. (2019). Genome-wide alternative splicing landscapes modulated by biotrophic sugarcane smut pathogen. *Scientific Reports* 9, 8876. doi: 10.1038/s41598-019-45184-1

Chekanova, J. A. (2015). Long non-coding RNAs and their functions in plants. *Curr. Opin. Plant Biol.* 27, 207–216. doi: 10.1016/j.pbi.2015.08.003

Chen, Z., Li, Y., Li, P., Huang, X., Chen, M., Wu, J., et al. (2021). Mircrorna profiles of early rice inflorescence revealed a specific mirna5506 regulating development of floral organs and female megagametophyte in rice. *Int. J. Mol. Sci.* 22, 6610. doi: 10.3390/ijms22126610

Chen, X., Mao, X., Huang, P., and Fang, S. (2019). Morphological characterization of flower buds development and related gene expression profiling at bud break stage in heterodichogamous cyclocarya paliurus (Batal.) lljinskaja. *Genes (Basel)* 10, 818. doi: 10.3390/genes10100818

Conn, V. M., Hugouvieux, V., Nayak, A., Conos, S. A., Capovilla, G., Cildir, G., et al. (2017).). a circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through r-loop formation. *Nat. Plants* 3, 17053. doi: 10.1038/nplants.2017.53

Correa, J. P., de, O., Silva, E. M., and Nogueira, F. T. S. (2018). Molecular control by non-coding RNAs during fruit development: From gynoecium patterning to fruit ripening. *Front. Plant Sci.* 871. doi: 10.3389/fpls.2018.01760

Cortés-López, M., and Miura, P. (2016). Emerging functions of circular RNAs. *Yale J. Biol. Med.* 89, 527–537.

D'Ario, M., Griffiths-Jones, S., and Kim, M. (2017). Small RNAs: Big impact on plant development. *Trends Plant Sci.* 22, 1056–1068. doi: 10.1016/j.tplants.2017.09.009

de Vos, L., Gevensleben, H., Schröck, A., Franzen, A., Kristiansen, G., Bootz, F., et al. (2017). Comparison of quantification algorithms for circulating cell-free DNA methylation biomarkers in blood plasma from cancer patients. *Clin. Epigenet.* 9, 125. doi: 10.1186/s13148-017-0425-4

Du, J., Cao, C., and Jiang, L. (2015). Genome-scale genetic screen of lead ion-sensitive gene deletion mutations in saccharomyces cerevisiae. *Gene* 563, 155–159. doi: 10.1016/j.gene.2015.03.018

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in drosophila. *Genome Biology* 5, R1 doi: 10.1186/gb-2003-5-1-r1

Fukuhara, T., and Tokumaru, S. I. (2014).). inflorescence dimorphism, heterodichogamy and thrips pollination in platycarya strobilacea (Juglandaceae). *Ann. Bot.* 113, 467–476. doi: 10.1093/aob/mct278

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: An efficient and unbiased algorithm for *de novo* circular RNA identification. *Genome Biol.* 16, 4. doi: 10.1186/s13059-014-0571-3

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. *Brief Bioinform.* 19, 803–810. doi: 10.1093/bib/bbx014

Gelaw, T. A., and Sanan-Mishra, N. (2021). Non-coding RNAs in response to drought stress. *Int. J. Mol. Sci.* 22, 12519. doi: 10.3390/ijms222212519

Han, Y., Li, X., Yan, Y., Duan, M. H., and Xu, J. H. (2020). Identification, characterization, and functional prediction of circular RNAs in maize. *Mol. Genet. Genomics* 295, 491–503. doi: 10.1007/s00438-019-01638-9

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Am. Stat. Assoc.* 101, 18–29. doi: 10.1198/016214505000000187

Heo, J. B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* (1979)331, 76–79. doi: 10.1126/science.1197349

Huang, Y. J., Liu, L. L., Huang, J. Q., Wang, Z. J., Chen, F. F., Zhang, Q. X., et al. (2013). Use of transcriptome sequencing to understand the pistillate flowering in hickory (Carya cathayensis sarg.). *BMC Genomics* 14, 1–18. doi: 10.1186/1471-2164-14-691

Huang, Y. J., Wang, Z. J., Zheng, B. S., Huang, X. M., and Huang, J. Q. (2006). Anatomy of stamen development on carya cathayensis. *J. Zhejiang A&F Univ.* 23, 56–60.

Huang, Y., Xiao, L., Zhang, Z., Zhang, R., Wang, Z., Huang, C., et al. (2019). The genomes of pecan and Chinese hickory provide insights into carya evolution and nut nutrition. *Gigascience* 8, giz036. doi: 10.1093/gigascience/giz036

Huang, Y. J., Xia, G. H., Wang, Z. J., Zheng, B. S., Liang, J. Y., and Huang, J. Q. (2007). Studies on anatomy of development of female flower in carya cathayensis sarg. *Acta Agriculturae Universitatis Jiangxiensis* 29, 723–726. doi: 10.13836/j.jjau.2007147

Huang, X., Zhang, H., Guo, R., Wang, Q., Liu, X., Kuang, W., et al. (2021). Systematic identification and characterization of circular RNAs involved in flag leaf senescence of rice. *Planta* 253, 26. doi: 10.1007/s00425-020-03544-6

Ishaque, A., and Al-Rubeai, M. (1998).). use of intracellular pH and annexin-V flow cytometric assays to monitor apoptosis and its suppression by bcl-2 over-expression in hybridoma cell culture. *J. Immunol. Methods* 221, 43–57. doi: 10.1016/s0022-1759(98)00166-5

Ishikawa, J., Kaisho, T., Tomizawa, H., Lee, B. O., Kobune, Y., INAzAwA, J., et al. (1995). Molecular cloning and chromosomal mapping of a bone marrow stromal cell surface gene, BSTZ, that may be involved in pre-B-Cell growth. *Genomics* 26, 527–534. doi: 10.1016/0888-7543(95)80171-h

Jandura, A., and Krause, H. (2017). The new RNA world: Growing evidence for long noncoding RNA functionality. *Trends Genet.* 33 10, 665–676. doi: 10.1016/j.tig.2017.08.002

Jeong, H. L., Seong, J. Y., Soo, H. P., Hwang, I., Jong, S. L., and Ji, H. A. (2007). Role of SVP in the control of flowering time by ambient temperature in arabidopsis. *Genes Dev.* 21, 397–402. doi: 10.1101/gad.1518407

Jiang, M., Chen, H., Du, Q., Wang, L., Liu, X., and Liu, C. (2021).). genome-wide identification of circular RNAs potentially involved in the biosynthesis of secondary metabolites in salvia miltiorrhiza. *Front. Genet.* 12. doi: 10.3389/fgene.2021.645115

Kang, C., and Liu, Z. (2015). Global identification and analysis of long non-coding RNAs in diploid strawberry Fragaria vesca during flower and fruit development. *BMC Genomics* 16, 815. doi: 10.1186/s12864-015-2014-2

Kikuchi, S., Shibata, M., Tanaka, H., Yoshimaru, H., and Niiyama, K. (2009). Analysis of the disassortative mating pattern in a heterodichogamous plant, acer mono maxim. using microsatellite markers. *Plant Ecol.* 204, 43–54. doi: 10.1007/s11258-008-9564-1

Lee, H., Yoo, S. J., Lee, J. H., Kim, W., Yoo, S. K., Fitzgerald, H., et al. (2010). Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in arabidopsis. *Nucleic Acids Res.* 38, 3081–3093. doi: 10.1093/nar/gkp1240

Liang, Y., Zhang, Y., Xu, L., Zhou, D., Jin, Z., Zhou, H., et al. (2019). CircRNA expression pattern and ceRNA and miRNA-mRNA networks involved in anther development in the CMS line of brassica campestris. *Int. J. Mol. Sci.* 20, 4808. doi: 10.3390/ijms20194808

Liao, Y., Smyth, G. K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Li, S., Chen, L., Xu, C., Qu, X., Qin, Z., Gao, J., et al. (2020). Expression profile and bioinformatics analysis of circular RNAs in acute ischemic stroke in a south Chinese han population. *Sci. Rep.* 10, 10138. doi: 10.1038/s41598-020-66990-y

Liebsch, D., and Palatnik, J. F. (2020). MicroRNA miR396, GRF transcription factors and GIF co-regulators: a conserved plant growth regulatory module with potential for breeding and biotechnology. *Curr. Opin. Plant Biol.* 53, 31–42. doi: 10.1016/j.pbi.2019.09.008

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., et al. (2015). Exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 22, 256–264. doi: 10.1038/nsmb.2959

Li, C., Jin, H., Zhang, W., Qin, T., Zhang, X., Pu, Z., et al. (2022). Whole-transcriptome analysis reveals long noncoding RNAs involved in female floral development of hickory (Carya cathayensis sarg.). *Front. Genet.* 13. doi: 10.3389/fgene.2022.910488

Lim, K. J., Paasela, T., Harju, A., Venäläinen, M., Paulin, L., Auvinen, P., et al. (2016). Developmental changes in scots pine transcriptome during heartwood formation. *Plant Physiol.* 172, 1403–1417. doi: 10.1104/pp.16.01082

Lim, K. J., Paasela, T., Harju, A., Venäläinen, M., Paulin, L., Auvinen, P., et al. (2021). A transcriptomic view to wounding response in young scots pine stems. *Sci. Rep.* 11, 3778. doi: 10.1038/s41598-021-82848-3

Liu, W., Jia, C., Luo, L., Wang, H. L., Min, X. L., Xu, J. H., et al. (2019). Novel circular RNAs expressed in brain microvascular endothelial cells after oxygen-glucose deprivation/recovery. *Neural Regener. Res.* 14, 2104–2111. doi: 10.4103/1673-5374.262589

Li, X., Yang, L., and Chen, L. L. (2018). The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell* 71, 428–442. doi: 10.1016/j.molcel.2018.06.034

Li, X., Fan, J., Luo, S., Yin, L., Liao, H., Cui, X., et al. (2021). Comparative transcriptome analysis identified important genes and regulatory pathways for flower color variation in Paphiopedilum hirsutissimum. *BMC Plant Biology* 21, 495. doi: 10.1186/s12870-021-03256-3

Lipka, A., Paukszto, L., Majewska, M., Jastrzebski, J. P., Panasiewicz, G., and Szafranska, B. (2019). De novo characterization of placental transcriptome in the Eurasian beaver (Castor fiber L.). *Functional & Integrative Genomics* 19, 421–435. doi: 10.1007/s10142-019-00663-6

Lu, T., Cui, L., Zhou, Y., Zhu, C., Fan, D., Gong, H., et al. (2015). Transcriptome-wide investigation of circular RNAs in rice. *RNA* 21, 2076–2087. doi: 10.1261/rna.052282.115

Mattick, J. S., and Makunin, I. v. (2006). Non-coding RNA. *Hum. Mol. Genet.* 15 Spec No, 17–29. doi: 10.1093/hmg/ddl046

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Mouradov, A., Cremer, F., and Coupland, G. (2002). Control of flowering time: Interacting pathways as a basis for diversity. *Plant Cell* 14, 111–130. doi: 10.1105/tpc.001362

Nigro, J. M., Cho, K. R., Fearon, E. R., Kern, S. E., Ruppert, J. M., Oliner, J. D., et al. (1991). Scrambled exons. *Cell* 64, 607–613. doi: 10.1016/0092-8674(91)90244-s

Olesen, M. T. J., and Kristensen, L. S. (2021). Circular RNAs as microRNA sponges : evidence and controversies. *Essays in Biochemistry* 65, 685–696. doi: 10.1042/EBC20200060

Pan, T., Sun, X., Liu, Y., Li, H., Deng, G., Lin, H., et al. (2018). Heat stress alters genome-wide profiles of circular RNAs in arabidopsis. *Plant Mol. Biol.* 96, 217–229. doi: 10.1007/s11103-017-0684-7

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit : A toolkit for quality control of next generation sequencing data *PLoS ONE* 7, e30619. doi: 10.1371/journal.pone.0030619

Philips, A., Nowis, K., Stelmaszczuk, M., Jackowiak, P., Podkowiński, J., Handschuh, L., et al. (2020). Expression landscape of circRNAs in arabidopsis thaliana seedlings and adult tissues. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.576581

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Roldán, M., Gómez-Mena, C., Ruiz-García, L., Salinas, J., and Martínez-Zapater, J. (1999). Sucrose availability on the aerial part of the plant promotes morphogenesis and flowering of arabidopsis in the dark. *Plant J.* 20 5, 581–590. doi: 10.1046/j.1365-313x.1999.00632.x

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: The rosetta stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Singh, U. M., Chandra, M., Shankhdhar, S. C., and Kumar, A. (2014). Transcriptome wide identification and validation of calcium sensor gene family in the developing spikes of finger millet genotypes for elucidating its role in grain calcium accumulation. *PloS One* 9, e103963. doi: 10.1371/journal.pone.0103963

Štorchová, H. (2017). The role of non-coding RNAs in cytoplasmic male sterility in flowering plants. *Int. J. Mol. Sci.* 18, 2429. doi: 10.3390/ijms18112429

Sun, X., Wang, L., Ding, J., Wang, Y., Wang, J., Zhang, X., et al. (2016). Integrative analysis of arabidopsis thaliana transcriptomics reveals intuitive splicing mechanism for circular RNA. *FEBS Lett.* 590, 3510–3516. doi: 10.1002/1873-3468.12440

Tong, W., Yu, J., Hou, Y., Li, F., Zhou, Q., Wei, C., et al. (2018). Circular RNA architecture and differentiation during leaf bud to young leaf development in tea (Camellia sinensis). *Planta* 248, 1417–1429. doi: 10.1007/s00425-018-2983-x

Waititu, J. K., Zhang, C., Liu, J., and Wang, H. (2020). Plant non-coding rnas: Origin, biogenesis, mode of action and their roles in abiotic stress. *Int. J. Mol. Sci.* 21, 1–22. doi: 10.3390/ijms21218401

Wang, D., Gao, Y., Sun, S., Li, L., and Wang, K. (2022). Expression characteristics in roots, phloem, leaves, flowers and fruits of apple circRNA. *Genes (Basel)* 13, 712. doi: 10.3390/genes13040712

Wang, G. Y., Zhu, Y. Y., and Zhang, Y. Q. (2014). The functional role of long non-coding RNA in digestive system carcinomas. *Bull. Cancer* 101, E27–E31. doi: 10.1684/bdc.2014.2023

Whittaker, C., and Dean, C. (2017). The FLC locus: A platform for discoveries in epigenetics and adaptation. *Annual Review of Cell and Developmental Biology* 33, 555–575 doi: 10.1146/annurev-cellbio-100616

Xu, K., Chen, D., Wang, Z., Ma, J., Zhou, J., Chen, N., et al. (2018). Annotation and functional clustering of circRNA expression in rhesus macaque brain during aging. *Cell Discovery* 4, 48. doi: 10.1038/s41421-018-0050-1

Xu, M., Zhu, J., Zhang, M., and Wang, L. (2016). Advances on plant miR169/NF-YA regulation modules. *Yi Chuan* 38, 700–706. doi: 10.16288/j.yczz.15-526

Yang, J., Zhou, F., Xiong, L., Mao, S., Hu, Y., and Lu, B. (2015). Comparison of phenolic compounds, tocopherols, phytosterols and antioxidant potential in Zhejiang pecan [Carya cathayensis] at different stir-frying steps. *LWT - Food Science and Technology* 62, 541–548. doi: 10.1016/j.lwt.2014.09.049

Yang, X., Liu, Y., Zhang, H., Wang, J., Zinta, G., Xie, S., et al. (2020). Genome-wide identification of circular RNAs in response to low-temperature stress in tomato leaves. *Front. Genet.* 11. doi: 10.3389/fgene.2020.591806

Ye, C. Y., Chen, L., Liu, C., Zhu, Q. H., and Fan, L. (2015). Widespread noncoding circular RNAs in plants. *New Phytol.* 208, 88–95. doi: 10.1111/nph.13585

Yoshimi, A., Miyazawa, K., and Abe, K. (2017). Function and biosynthesis of cell wall α-1,3-glucan in fungi. *J. Fungi* 3, 63. doi: 10.3390/jof3040063

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010).). open access METHOD gene ontology analysis for RNA-seq: accounting for selection bias GOseq GOseq is a method for GO analysis of RNA-seq data that takes into account the length bias inherent in RNA-seq. *Genome Biol.* 11, 14. doi: 10.1186/gb-2010-11-2-r14

Zhang, Y., Liang, W., Zhang, P., Chen, J., Qian, H., Zhang, X., et al. (2017b). Circular RNAs: emerging cancer biomarkers and targets. *J. Exp. Clin. Cancer Res.* 36, 152. doi: 10.1186/s13046-017-0624-z

Zhang, X., Ma, X., Ning, L., Li, Z., Zhao, K., Li, K., et al. (2019). Genome-wide identification of circular RNAs in peanut (Arachis hypogaea l.). *BMC Genomics* 20, 653. doi: 10.1186/s12864-019-6020-7

Zhang, F., Zhang, R., Zhang, X., Wu, Y., Li, X., Zhang, S., et al. (2019). Comprehensive analysis of circRNA expression pattern and circRNA-miRNA-mRNA network in the pathogenesis of atherosclerosis in rabbits. *Aging* 10, 2266–2283. doi: 10.18632/aging.101541

Zhang, S., Zhu, D., Li, H., Li, H., Feng, C., and Zhang, W. (2017a). Characterization of circRNA-Associated-ceRNA networks in a senescence-accelerated mouse prone 8 brain. *Mol. Ther.* 25, 2053–2061. doi: 10.1016/j.ymthe.2017.06.009

Zhong, Y., Du, Y., Yang, X., Mo, Y., Fan, C., Xiong, F., et al. (2018). Circular RNAs function as ceRNAs to regulate and control human cancer progression. *Mol. Cancer* 17, 79. doi: 10.1186/s12943-018-0827-8

Zhou, W. Y., Cai, Z. R., Liu, J., Wang, D. S., Ju, H. Q., and Xu, R. H. (2020). Circular RNA: metabolism, functions and interactions with proteins. *Mol. Cancer* 19, 172. doi: 10.1186/s12943-020-01286-3

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 03 April 2023
DOI 10.3389/fpls.2023.1052425

OPEN ACCESS

EDITED BY
Fang Du,
Beijing Forestry University, China

REVIEWED BY
Patrick Von Aderkas,
University of Victoria, Canada
Chunjie Fan,
Research Institute of Tropical Forestry,
Chinese Academy of Forestry, China

*CORRESPONDENCE
Esteban Galeano
esteban.galeano@msstate.edu

SPECIALTY SECTION
This article was submitted to
Plant Bioinformatics,
a section of the journal
Frontiers in Plant Science

RECEIVED 03 October 2022
ACCEPTED 15 March 2023
PUBLISHED 03 April 2023

CITATION
Galeano E and Thomas BR (2023)
Unraveling genetic variation among white
spruce families generated through different
breeding strategies: Heritability,
growth, physiology, hormones
and gene expression.
Front. Plant Sci. 14:1052425.
doi: 10.3389/fpls.2023.1052425

Unraveling genetic variation among white spruce families generated through different breeding strategies: Heritability, growth, physiology, hormones and gene expression

Esteban Galeano [1,2*] and Barb R. Thomas [2]

[1]Department of Forestry, Mississippi State University, Starkville, MS, United States, [2]Department of Renewable Resources, University of Alberta, Edmonton, AB, Canada

Tree improvement programs select genotypes for faster growth, at both early and late stages, to increase yields over unimproved material, and the improvement is frequently attributed to genetic control in growth parameters among genotypes. Underutilized genetic variability among genotypes also has the potential to ensure future gains are possible. However, the genetic variation in growth, physiology and hormone control among genotypes generated from different breeding strategies has not been well characterized in conifers. We assessed growth, biomass, gas exchange, gene expression and hormone levels in white spruce seedlings obtained from three different breeding strategies (controlled crosses, polymix pollination, open pollination) using parents grafted into a clonal seed orchard in Alberta, Canada. A pedigree-based best linear unbiased prediction (ABLUP) mixed model was implemented to quantify variability and narrow-sense heritability for target traits. The levels of several hormones and expression of gibberellin-related genes in apical internodes were also determined. Over the first two years of development, the estimated heritabilities for height, volume, total dry biomass, above ground dry biomass, root:shoot ratio and root length, varied between 0.10 and 0.21, with height having the highest value. The ABLUP values showed large genetic variability in growth and physiology traits both between families from different breeding strategies, and within families. The principal component analysis showed that developmental and hormonal traits explained 44.2% and 29.4% of the total phenotypic variation between the three different breeding strategies and two growth groups. In general, controlled crosses from the fast growth group showed the best apical growth, with more accumulation of indole-3-acetic acid, abscisic acid, phaseic acid, and a 4-fold greater gene expression of *PgGA3ox1* in genotypes from controlled crosses versus those from open pollination. However, in some cases, open pollination from the fast and slow growth groups showed the best root development, higher water use efficiency (iWUE and $\delta^{13}C$) and more accumulation of zeatin and isopentenyladenosine. In

conclusion, tree domestication can lead to trade-offs between growth, carbon allocation, photosynthesis, hormone levels and gene expression, and we encourage the use of this phenotypic variation identified in improved and unimproved trees to advance white spruce tree improvement programs.

KEYWORDS

Conifers, tree improvement, phenotypic variation, domestication, selection, gas exchange, phytohormones

Introduction

White spruce (*Picea glauca* (Moench) Voss) is one of the most widely distributed tree species in Canada, and with tree improvement programs across the country (Galeano et al., 2021). In Alberta, Canada, improvement programs for white spruce mainly use open-pollinated material, however, this strategy is not sufficient to explore the full potential of productivity of white spruce plantations (Flewelling, 2008; Sklar, 2012; FGRMS, 2016; Government of Alberta, 2018; Galeano et al., 2021). In order to advance improvement in white spruce, the adoption of controlled crosses needs to be considered as it typically results in better productivity than open pollination in most cases, and this approach has been the backbone of plant domestication (Leakey, 2017a). Forest tree domestication is defined as the production of individual genotypes or clones exhibiting traits that are desirable to foresters through an active and conscious process of selection, testing, and breeding, becoming increasingly important in supplying timber, fiber, fuel, food, resistance to pests/diseases and to abiotic stressors for future generations (Pearsall, 2008; Harfouche et al., 2012; Gepts, 2014; Leakey, 2014; Leakey, 2017a). Tree domestication also has a key role to play in the mitigation of climate change and conservation of natural resources (Leakey, 2017b). The first plant breeding occurred with the domestication of wild wheat, barley, and lentils around 13,000-10,000 years ago, but the first successful selections and crosses were made with cereals in the Fertile Crescent in Eurasia (today known as Iraq, Iran, Turkey, Lebanon, and Israel) between 9,000 and 7,000 BCE (Pearsall, 2008; Abbo and Gopher, 2020). Tree domestication was first reported in date palm through hand pollination, in approximately 700 BCE (Singh et al., 2021).

Taking advantage of polymix pollination and controlled crosses, uncovers hidden variation not visible using open pollination, and can thereby maximize the benefits possible from the better trees in a population (Kanowski and Borralho, 2004). The genetic variation of trees within genotypes resulting from different breeding strategies is unknown, and trees or families with outstanding performance are commonly not included with an open pollination strategy. Genetic variation with substantial ranges has been observed, for juvenile morphological and physiological traits of spruce and pine seedlings, between seed orchards, regions, families, somatic clones, provenances and among trees within the same family (Li et al., 1993; Rweyongeza et al., 2005; Carles et al., 2012; Quesada et al., 2017; Prud'Homme et al., 2018). The main traits of interest in tree breeding programs usually involve height, diameter, wood properties, and disease and insect resistance, which are considered quantitative traits and influenced by the action of multiple genes and the environment, but undoubtedly, other traits of increasing interest in forestry are biomass, photosynthesis, water use efficiency, and $\delta^{13}C$ in particular, with the increased impact of climate change (Quesada et al., 2017). Genetic variation contributes to phenotypic plasticity and, therefore, constitutes an essential factor in the adaptive capacity of trees and subsequent forest productivity (Grattapaglia et al., 2009; Verta et al., 2013; Plomion et al., 2016). Conifers have shown considerable capacity for rapid local adaptation even with their long lifespans and extensive gene flow due to wind pollination (Six et al., 2021). Furthermore, growth performance and architectural patterns of conifers established in the early stages are indicative of how their development will be, decades later, when the plant has increased substantially in size (Raj et al., 2006). Different patterns in gene expression and hormone levels (e.g. auxins, cytokinins, abscisic and gibberellic acid) allow for contrasting performance of growth and physiology of trees within families, contributing to functional innovations, genetic variation and may facilitate adaptation following environmental change (Verta et al., 2013; Park et al., 2014).

Auxins function primarily in stem elongation by promoting cell growth, and indole-3-acetic acid (IAA) is the major naturally occurring auxin (Dilworth et al., 2017). Cytokinins promote cell division and differentiation, budbreak, increase tolerance to drought stress, and inhibit cell elongation; the most naturally occurring cytokinins are trans-zeatin (tZ), dihydrozeatin (dhZ) and isopentenyladenosine (iPR) (Persson et al., 1994; Cline and Harrington, 2007; Lulsdorf et al., 2013; Dilworth et al., 2017). Abscisic acid (ABA) prevents the activation of axillary buds, inhibits elongation of internodes and seed germination, and it is responsible for the plant response to abiotic stresses (Arney and Mitchell, 1969; Dilworth et al., 2017; Shu et al., 2018). Gibberellic acid (GA) promotes stem elongation, flowering, leaf expansion, and seed germination (Dilworth et al., 2017; Shu et al., 2018). Many hormones are present in plants and plant developmental processes are controlled by complex interactions. For example, the cross-talk between auxin and cytokinin controls the apical dominance and functions mainly in the newly formed buds of the current year's growth, while the cross-talk between ABA and GA antagonistically regulates root initiation, stem and hypocotyl elongation, and the interaction between auxins and GA regulates plant growth (Cline

and Harrington, 2007; Shu et al., 2018; Akhtar et al., 2020). Furthermore, previous studies have found differences in hormone levels and GA-related genes between individuals and families of spruce and pine open-pollinated plant material, identifying these hormones as potential biomarkers for growth in trees (Kayal et al., 2011; Park et al., 2014; Galeano and Thomas, 2020).

Despite the availability of different breeding strategies, open pollination is still the dominant practice in Alberta's tree improvement programs. We believe there is a need for genetic and phenotypic data to inform our understanding of the potential of the genetic variability when using different breeding strategies to support selection of material and genotypes and increase the genetic gain. So far, there are no studies characterizing the genetic variation of growth and physiological traits in white spruce individuals coming from different breeding strategies. Also, there are no studies analyzing the influence and changes of gene expression and hormones in spruce trees from different breeding strategies, and use of these tools in conifer tree improvement. The objectives of the present study are to: (a) evaluate the genetic variation in growth and physiology among families from three different breeding strategies (i.e. open pollination, polymix pollination, controlled cross pollination), (b) analyze the interactions between growth, physiology, hormones, and gibberellin-related genes between breeding strategies and growth groups (fast and slow).

Materials and methods

Selection of genotypes for breeding

This study was conducted using seeds from the G1 clonal seed orchard, as part of the G1 white spruce Controlled Parentage Program (CPP) that began in 1979 (John, 2011). The G1 clonal seed orchard is located near Grande Prairie, Alberta, Canada (lat. 55°03'51" N, long. 119°16'24" W, 720 elevation) and designed to produce seed for the G1 region in north-west central Alberta (FGRMS, 2016).The G1 clonal seed orchard was established from ramets, or grafts, collected from parent trees located in wild stands between 1988-2005, and consisted of 151 'founders' (with ~12 ramets per selected parent tree (clone)). Based on the growth measurements from four progeny trials, the orchard manager ranked genotypes based on the breeding values for height (%) at a rotation age of 80, and based on the rankings, conducted three roguing. The first roguing removed six genotypes (all ramets) in fall of 2009, the second roguing removed eight genotypes (all ramets) in spring of 2010, and the third roguing removed 84 genotypes in spring of 2018 (Supplementary Material 1A), leaving 53 founder genotypes considered as "superior" eliminating 65% of the original 151 genotypes in the orchard by summer 2018 (John, 2011). From the 53 remaining genotypes, we selected six genotypes as females and nine different genotypes as males (Supplementary Material 1B), to perform and compare three different breeding strategies: controlled cross pollination (CC), polymix pollination (PM), and open pollination (OP).

Controlled cross pollination, polymix pollination and open pollination

During summer 2018, we performed a polymix pollination (PM) and controlled cross pollination (CC), and collected seeds from open pollination (OP), with the same females used for all three breeding strategies. Controlled crosses were performed using a disconnected factorial design (Isik et al., 2017) with 15 elite parents grouped into three different clusters based on breeding values for tree height obtained from two progeny trials at ages 14 and 31. Each of the six females were also pollinated using a mix of pollen (polymix) from the nine selected males, and open-pollinated cones which were also collected from the same female trees for comparison. Male and female individuals were inspected for phenological stages of reproductive structures (strobili) in the first week of May 2018 (Supplementary Material 2). Reciprocal crosses were not performed. Branches selected for male strobili were marked for pollen collection. Female branches selected were covered with pollination bags (PBS 3D.75, 158mm wide x 750mm tall x 158mm deep with a UV stable vinyl window 100mm x 250mm on front, non-woven polyester, PBS International, Scarborough, United Kingdom) on May 10, 2018, prior to pollen release based on monitoring, and based on the female conelets not being receptive yet. The PBS 3D.75 bags were chosen based on Heine et al. (2020), which showed that PBS bags were more suitable than kraft bags for conelet survival when breeding loblolly pine trees. The bags were sealed with plastic ties, and sanitary pads were placed in between the bags and branches where the ties were placed to avoid damage to the branch and prevent pollen contamination (Supplementary Material 2). Approximately 50 male conelets were collected on May 17, 2018 from each of the nine selected clones. Next, male conelets were placed into a 20 cm diameter industrial funnel (Scepter Canada Inc., Scarborough, Canada) which allowed for separation of the pollen from the conelet and sieved through a 1.18 mm sieve (Royal Selangor Co). An equal portion of pollen (~ 1 g) from each male was combined and used for the PM. Either individually or in the pollen mix, pollen was placed into individual long nozzle squeeze bottles with a 13G size purple blunt needle (2.4 mm outer diameter x 1.8 mm inner diameter x 0.3 wall thickness) affixed to the tip. For the controlled crosses, one bottle was assigned for each individual pollen source. Seven days after bagging the branches with female conelets, pollen was applied directly to the 1-10 conelets inside each pollination bag (May 17-18 2018), by puncturing the pollination bags with the needle. This process was repeated every other day for a total of three applications to ensure each female conelet was exposed to the pollen during the receptive period. The puncture holes were resealed with black electrician's tape and the bags remained on the branches for four weeks during fertilization (Bonner et al., 2009). The bags were removed by mid-June 2018 after conelets were no longer receptive, and the conelets continued to develop on the branches with periodic examinations through June and July 2018. Two OP cones were cut from different trees every week to evaluate maturity. All cones used in this study were collected on August 13-14, 2018 (Supplementary Material 2).

Twenty-seven individual seedlots were collected from the CC and nine seedlots from the PM (without replication) with a range of 1-60 cones per seedlot. In addition, approximately 20 open-pollinated cones were randomly selected from each of the nine females used in the controlled crosses, representing the OP seedlots. Cones were placed in metal trays to dry for two weeks with intermittent stirring (Gärtner et al., 2011) at the Thomas Lab, University of Alberta. The dried, opened cones were taken to the Alberta Tree Improvement and Seed Center in Smoky Lake, AB, Canada, for seed extraction. Seed extraction and de-winging of seeds were done manually during September 2018 (Supplementary Material 2). The extracted seeds were sorted in two steps, first by size, through a Canadian Standard Sieve size 16 (1.18 mm) (W.S. Tyler Company of Canada Ltd., St. Catharines, Ontario, Canada) to remove the smallest seeds which were discarded. Then, wings and waste were subsequently removed using a size 8 (2.36 mm) sieve. Second, the seeds were sorted by weight through a series of blowers to remove empty and aborted seeds (Supplementary Material 2). Extracted seeds were placed into 10 x 15 cm metal bags metallic bags (ULINE, Milton, Ontario, Canada) and stored in a -20°C freezer until grown.

Experimental design and condition of plants at the greenhouse

The experiment was a randomized complete block design (Supplementary Material 3) with a total of 18 families, 10 blocks and three trees/family/block. To assess the first objective (evaluate the genetic variation of growth and physiology among families from three different breeding strategies), we categorized all genotypes into two breeding value rankings based on height (%): mid BV (2.5 to 6.0), high BV (6.0 to 12) (Supplementary Material 1). To assess the second objective (analyze the interactions between growth, physiology, hormones, and gibberellin-related genes with the breeding strategies and growth groups), we took the two families with the lowest BVs from the mid BV ranking (genotypes 193 and 122, BV=3.9), and we called them the "slow growth" group (Table 1; Supplementary Material 1).We then took the two families with the highest BVs from the high BV ranking (genotypes 927 and 138, BV=8.9), and we called them the "fast growth" group (Table 1; Supplementary Material 1). Seed was removed from the -20°C freezer and pre-treated at 4°C in the fridge for two weeks, after which a total of 180 seeds were sown (10 per family, each in 10 blocks, 18 families) on January 16, 2019 at a commercial forest nursery (Bonnyville, Alberta, Canada). Seedlings began emerging on February 1, 2019 and were grown until November 29, 2019 (10 months), and subsequently lifted, packed and stored at -2°C on December 1, 2019 at Bonnyville, Alberta, Canada. On January 20, 2020 seedlings were removed from cold storage and planted into 2L pots filled with Sunshine Mix #4 (Sungro, Vilna, Alberta, Canada) and grown at the Biological Sciences greenhouse, University of Alberta, for their second growing season, between January 28 and April 28, 2020 (Day 92 of the experiment). Plants were grown under natural light supplemented by cool-white fluorescent lamps (400 μmol m^{-2} s^{-1}) and provided with a 16/8 hour photoperiod and maximum day and night temperatures of 25°C and 18°C, respectively. Seedlings were irrigated three times per week, fertilized using 'pH Reducer' fertilizer (Plant-Prod Solutions Inc, Brampton, Canada) and Iron Chelate/Rexolin (FeEDTA 13.3%) (Yara Company, Regina, Canada) once per week, and fans were used to ensure good air circulation throughout the greenhouse. On April 28, 2020 (Day 92), seedlings were measured and harvested including internodes, stem (RNA extraction and hormone analysis), needles, branches and roots.

TABLE 1 Experimental design of the study, with a total of six genotypes used as females, six breeding value (BV) groups, two growth groups (slow growth as the average of F193 and F122; fast growth as the average of F927 and F138), three breeding strategies [the same female was used for OP (open pollination), PM (polymix pollination), CC (controlled crosses)]. The breeding value of each genotype used as female and male is also included.

Genotypes used as Female (BV)*	BV group	Growth group	Breeding strategies		
			OP**	PM (BV)***	CC [Female X Male (BV)]
F193 (3.93)	Mid BV	Slow growth	**F193** X (53 males + external pollen)	**F193** X 9 males (7.27)	**F193** X M1045 (7.44)
F122 (3.94)	Mid BV	Slow growth	**F122** X (53 males + external pollen)	**F122** X 9 males (7.27)	**F122** X M1047 (8.26)
F754 (5.63)	Mid BV	–	**F754** X (53 males + external pollen)	**F754** X 9 males (7.27)	**F754** X M966 (9.49)
F129 (7.89)	High BV	–	**F129** X (53 males + external pollen)	**F129** X 9 males (7.27)	**F129** X M756 (6.31)
F927 (8.87)	High BV	Fast growth	**F927** X (53 males + external pollen)	**F927** X 9 males (7.27)	**F927** X M991 (6.49)
F138 (8.91)	High BV	Fast growth	**F138** X (53 males + external pollen)	**F138** X 9 males (7.27)	**F138** X M752 (6.70)

*The females (bold) are overlapped between the breeding strategies. **Open pollination includes the 53 genotypes in the G1 orchard and external pollen from other sources (orchards and natural stands). ***Genotypes contributing as males for polymix: 199, 756, 991, 752, 115, 1045, 1002, 1047, 966 (Average BV=7.27).

Measurements of growth, biomass, and gas exchange parameters

Final height was measured with a meter stick and diameter was measured with a digital caliper, for all seedlings (n=180), at Day 92. Volume was calculated with the equation $V=\pi^*(D/2)^{2*}H$ at each of the five measurement points, following Galeano and Thomas (2020). Apical internode length was measured to the nearest 0.1 mm at Day 92 of the experiment. Above ground tissue, which includes branches, needles and stem, was collected and placed into paper bags. Roots were stored at -20°C for two months prior to being thawed and carefully washed. Subsequently, root length was measured to the nearest 0.1 mm and placed back into paper bags. All tissue (above ground tissue and roots) was dried for 48 hours at 60°C, and dry weights were measured using a digital scale, model AV53 (readability 0.001 g, OHAUS Adventurer Pro, Melrose, MA, USA). Gas exchange parameters were measured on all seedlings (n=180) with a CIRAS-3 Portable Photosynthesis System (PP systems, Amesbury, USA) using the conifer cuvette at Day 92 of the experiment. Measurements were taken on needles from the top branch to ensure a uniform phenological stage for all plants. Four traits were measured: photosynthesis (A; μmol CO_2 m^{-2} s^{-1}), transpiration (E; mmol H_20 m^{-2} s^{-1}), stomatal conductance (gs; mol H_20 m^{-2} s^{-1}), and intrinsic water use efficiency (iWUE=A/gs; μmol CO_2 mol^{-1} H_20). After the measurements were concluded, the small branch with needles, used for the gas exchange measurements, was cut off and frozen in liquid nitrogen. The projected needle surface area of the needles (cm^2) was calculated using WinSEEDLE software, version 2004 (Regent Instruments, Quebec, Canada) to correct the gas exchange measurements.

Harvesting of needles for $\delta^{13}C$ analysis

On April 28, 2020 (Day 92), a subsample of 36 trees from three blocks were chosen for harvesting of needles for $\delta^{13}C$ analysis, corresponding to the genotypes 193 and 122, (lowest BV=3.9, both families with same value) and genotypes 927 and 138, (highest BV=8.9, both families with same value) from the three breeding strategies, for a total of 12 families (see Experimental Design above). A total of 100 mg of needles from the top branch were harvested into small paper bags, dried for 72 hours at 65°C, and ground using 25 mL stainless steel metal jars, 20 mm metal balls, with a Qiagen TissueLyser II (Qiagen, Redwood City, CA, USA). Approximately 50 mg of ground sample per tree was sent to the Stable Isotope Lab of InnoTech Alberta (Victoria, BC, Canada) for $\delta^{13}C$ analysis. Samples were analyzed using an established method on a MAT 253 mass spectrometer with Conflo IV interface (Thermo Fisher Scientific, Waltham, MA, USA.), and a Fisons NA1500 EA (Fisons Instruments, Milano, Italy), providing a bulk analysis of carbon discrimination ($\delta^{13}C$). In brief, approximately 1.0 mg of solid sample was weighed into tin capsules and then dropped into a combustion reactor that produces CO_2. The separated CO_2 was then transferred to the mass spectrometer for isotopic measurement. Multiple in-house standards, calibrated relative to international standards, were also run as samples to allow the results to be normalized and reported vs Vienna Pee Dee Belemnite (VPDB) ($\delta^{13}C$).

Harvesting of the apical internode

The same subsample of 36 trees, from three blocks chosen for $\delta^{13}C$, were used for harvesting the apical internode. After manually removing all needles, we used a scalpel to cut and collect at least 1.0 g of the apical internode. The scalpel was cleaned with 50% ethanol and rinsed with distilled water between samples. After cutting, the samples were wrapped in a double-layer of aluminum foil, labelled and immediately frozen by immersion in liquid nitrogen at the greenhouse and stored at -80°C prior to grinding for RNA extraction and hormone analysis (Supplementary Material 4). Two weeks after storing the material at -80°C, the 36 samples were ground into a fine powder using liquid nitrogen with a mortar and pestle, and the sample was divided into two Eppendorf tubes of 1.5 ml, each with 0.5 g of ground apical internode, for the RNA extraction and hormone analysis.

RNA extraction, cDNA synthesis, and expression of GA-related genes

Using 0.5 g of ground apical internode, total RNA was extracted using the 'Purelink RNA Mini Kit', 'Plant RNA Isolation Aid' and 'PureLink DNAse' (Thermo Scientific Inc, Waltham, USA), following the manufacturer's instructions. RNA extractions were performed in the Karst Lab, University of Alberta. A nanoDrop ND-1000 spectrophotometer (Thermo Scientific Inc, Waltham, USA) was used to check the RNA quality and concentration. The cDNA synthesis was performed using a total of 0.5 μg RNA from each sample, with the Superscript III reverse transcriptase (Thermo Scientific Inc, Waltham, USA). The *PgGA3ox1, PgGA20ox1, PgDELLA1* and *PgGID1* genes were selected based on previous studies (Kayal et al., 2011; Galeano and Thomas, 2020). Also, we used gene-specific primers for the target and reference genes based on previous work (Galeano and Thomas, 2020) (Supplementary Material 5) and determined the standard curve with several cDNA dilutions and the melting curve (Supplementary Material 6). Quantitative reverse transcription PCR (qRT-PCR) was performed for the four target GA-related genes of white spruce (*PgGA3ox1, PgGA20ox1, PgDELLA1, PgGID1*) using the 36 samples from the apical internode. PCR reactions were performed in 10 mL, containing SYBR Green master mix [0.2 mM dNTPs, 0.3 U Platinum Taq Polymerase (Invitrogen, Waltham, USA), 0.25X SYBR Green, and 0.1X ROX], 50 ng of cDNA and 300 nM of each primer. Three technical replicates for each reaction were analyzed on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Waltham, USA) where the first step was 95°C for 2 min followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. The qRT-PCR experiments were done at the Molecular Biology Service Unit, University of Alberta. Melting curves were generated

using the following program: 95°C for 15 s, 60°C for 15 s, and 95°C for 15 s. Translation initiation factor 5A (*PgTIF5A*, GenBank DR448953) was used as a reference gene, following previous work (Kayal et al., 2011; Galeano and Thomas, 2020). Two variables were used from the gene expression: absolute and relative expression. Absolute transcript levels were quantified using standard curves for the target genes (*PgGA3ox1, PgGA20ox1, PgDELLA1, PgGID1*) and reference gene (*PgTIF5A*) for each tissue per family per treatment (Kayal et al., 2011; Galeano and Thomas, 2020), and these values were used for the fixed-effect mixed model, interaction plots and correlations. Relative expression was calculated following the double delta C_T method (Rao et al., 2013), using the control plants as a calibrator to normalize the values between different plates, and the reference gene (*PgTIF5A*) as the control gene.

Hormone analysis

A total of 0.5 g of ground apical internode from the 36 samples were sent to the Aquatic and Crop Resource Development Research Center, National Research Council of Canada (NRCC), Saskatoon, Saskatchewan, Canada, for the quantification of abscisic acid (ABA), abscisic acid glucose ester (ABA-GE), 7'-hydroxy-abscisic acid (7'OH-ABA), phaseic acid (PA), indole-3-acetic acid (IAA), zeatin-O-glucoside-trans (ZOG-t), zeatin-O-glucoside-cis (ZOG-c), zeatin riboside-trans (ZR-t), dihydrozeatin riboside (dhZR), isopentenyladenosine (iPR), and gibberellin 3 (GA3). For the calibration curves, ABA-GE, PA and 7'- OH-ABA were synthesized and prepared at NRCC, Saskatoon, Canada. Also, ABA, IAA, ZR-t, iPR, were purchased from Sigma–Aldrich (Burlington, MA, United States), and ZOG-t, ZOG-c, dhZR and GA3 were purchased from OlChemim Ltd. (Olomouc, Czech Republic). Deuterated forms of the hormones were used as internal standards: d4-ABA, d5-ABA-GE, d4-7'-OH-ABA, d3-PA were synthesized and prepared at NRCC, Saskatoon, Canada (Abrams et al., 2003; Zaharia et al., 2005),, d5-IAA was purchased from Cambridge Isotope Laboratories (Andover, MA), d5-ZOG-t, d5-ZOG-c, d3-ZR-t, d3-dhZR, d6-iPR, and d2-GA3 were purchased from OlChemim Ltd. (Olomouc, Czech Republic). Analysis was performed on a UPLC/ESI-MS/MS utilizing a Waters ACQUITY UPLC system, equipped with a binary solvent delivery manager and a sample manager coupled to a Waters Micromass Quattro Premier XE quadrupole tandem mass spectrometer *via* a Z-spray interface. The MassLynx™ and QuanLynx™ (Micromass, Manchester, UK) were used for data acquisition and data analysis. The procedure for quantification of ABA and ABA catabolites, cytokinins, auxins, and gibberellins in plant tissue was performed using a modified procedure described in previous work (Lulsdorf et al., 2013). Briefly, the analyses utilize the Multiple Reaction Monitoring (MRM) function of the MassLynx v4.1 (Waters Inc) control software. The resulting chromatographic traces are quantified off-line by the QuanLynx v4.1 software (Waters Inc.) wherein each trace is integrated and the resulting ratio of signals (non-deuterated/internal standard) is compared with a previously constructed calibration curve to yield the amount of analyte present (ng per sample). Calibration curves were generated from the MRM signals obtained from standard solutions based on the ratio of the chromatographic peak area for each analyte to that of the corresponding internal standard. The QC samples, internal standard blanks and solvent blanks were also prepared and analyzed along with each batch of tissue samples.

Statistical analysis and models

Assumptions of homogeneity of variance and deviation of residuals from the normal distribution were confirmed before proceeding with the ANOVAs for the different models and mean comparisons. First, we ran a family genetic model (Isik et al., 2017) (Eq. 1), to calculate variance components for each parameter, heritabilities and BLUPs per family:

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{Z_1BS}+\boldsymbol{Z_2Fam}+\boldsymbol{e} \quad \text{(Eq. 1)}$$

Where β is the vector of the block fixed effect; *BS* is the vector of breeding strategy fixed effect; *Fam* is the vector of random family effects, ~N(0, $A\sigma^2_F$); e is the vector of random residual effects, ~N(0, $I\sigma^2_F$); A is the pedigree matrix. Narrow sense heritability (h^2) is then estimated using the additive (σ^2_a) and residual (σ^2_e) variances (Fernández-Paz et al., 2021) (Eq. 2):

$$\boldsymbol{h^2}=\sigma^2_a/[\sigma^2_a+\sigma^2_e] \quad \text{(Eq. 2)}$$

Second, we ran a fixed-effect mixed model to assess main and interaction effects of the breeding strategies and growth groups (Eq. 3):

$$\boldsymbol{Y_{ijkl}}=\mu+\boldsymbol{GG_i}+\boldsymbol{BS_j}+\boldsymbol{Bl_k}+(\boldsymbol{GG}*\boldsymbol{BS})_{ij}+\boldsymbol{e_{ijkl}} \quad \text{(Eq. 3)}$$

Where Y_{ijkl} is the measured value for each growth group, breeding strategy and block; μ is the overall mean; GG_i is the fixed growth group effect; BS_j is the fixed breeding strategy effect; $Bl_k$is the random block effect; $(GG*BS)_{ij}$ is the fixed interaction effect between growth group and breeding strategy; e_{ijkl} is the residual error. For Equation 1, a total of 180 plants were analyzed (10 blocks, 18 families). For Equation 2, a total of 36 plants were analyzed (3 blocks, 12 families) to test the extremes only. The models were assessed with $\alpha \leq 0.05$ and $\alpha \leq 0.01$ (as a more stringent option). Significant differences between means of all traits (growth, biomass, gas exchange, transcript and hormone levels) were determined by Tukey's HSD test with overall $\alpha \leq 0.05$. Pearson's correlation (r) coefficients between growth, transcript and hormone levels were estimated with $\alpha \leq 0.05$ and $\alpha \leq 0.01$. All statistical analyses were carried out using the R environment, using the ASReml package (Gilmour et al., 2009). Graphics were generated using the ggplot2 package (Wickham, 2016) in the R environment and Excel.

Results

Genetic variation of growth and physiology of white spruce seedlings among different breeding strategies

A total of 18 families were obtained from the three breeding strategies studied: open pollination (OP), polymix pollination (PM) and controlled cross pollination (CC). Means of each of the two

different breeding values and three different breeding strategies are shown in Table 2. The highest h^2 was found for height (0.21), followed by above ground dry biomass (0.16), and root:shoot ratio (0.14) (Table 3). All dry biomass, volume and root length showed lower heritabilities of 0.12, 0.11, 0.10, respectively (Table 3). Diameter, apical internode length, root dry biomass and gas exchange parameters showed a heritability below 0.08 (Table 3). We obtained the Best Linear Unbiased Predictions (BLUPs) using the A matrix for each of the 18 families with overlapping 'females' using Model 1 (see Materials and Methods). We expected that families in the High BV ranking would perform better than those in the Mid BV ranking, but in general, families from both rankings showed no clear tendency in growth, biomass or gas exchange parameters for any breeding strategy (Figure 1). Families 129 and 138 (High BV) coming from the PM and CC showed smaller ABLUPS for height, diameter and volume compared to families 122 and 754 coming from OP (Figures 1A–C). Family 927 (High BV) showed the highest ABLUPs for apical internode length among all families and breeding strategies, but family 138 (High BV), coming from CC, showed similar ABLUP values compared to families 122 and 754 (Mid BV), coming from CC, PM and OP (Figure 1D). Families 129, 927 and 138 (High BV) showed the lowest ABLUPs for above ground dry biomass, root dry biomass and total dry biomass compared to family 754 (Mid BV) (Figures 1E–G). On the other hand, CC and PM performed better than OP for family 754 for height, diameter, volume, apical internode length, above ground biomass, and total dry biomass (Figures 1A–G). For root:shoot ratio, OP showed a higher ABLUP for family 754 than the rest of the families (Figure 1H). Regarding gas exchange parameters, ABLUPs showed smaller differences between families compared to growth and biomass traits. For photosynthesis, transpiration and stomatal conductance, family 138 (High BV) showed lower values compared to the other families, and families 754 (Mid BV) and 927 (High BV) showed OP and CC with almost the same ABLUPs (Figures 1I–K).

TABLE 2 Means (± SE) for traits including growth, biomass, gas exchange parameters (n=180, 10 blocks, 18 families), $\delta^{13}C$, hormone levels, absolute gene expression parameters (n=36, 3 blocks, 12 families) in 2-year-old white spruce seedlings from three different breeding strategies (OP, open pollination; PM, polymix pollination; CC, controlled crosses) and two different levels of breeding values (BV) (medium and high).

Traits [1]	Breeding strategy (mean ± SE)					
	OP		PM		CC	
	Mid BV	High BV	Mid BV	High BV	Mid BV	High BV
H	32.6 ± 4.4	37 ± 4.2	33.9 ± 5.3	35.2 ± 4.5	34.7 ± 4.9	34.1 ± 3.7
D	7.2 ± 0.9	7.7 ± 1.0	7.3 ± 1.0	7.2 ± 1.3	7.8 ± 1.3	7.4 ± 1.3
Vol	13.7 ± 4.6	17.3 ± 5.1	14.5 ± 4.5	15.2 ± 6.0	17.3 ± 6.5	15.2 ± 5.7
Ap.Int.L	9.6 ± 2.2	11.2 ± 2.5	10.2 ± 2.7	10.9 ± 2.1	10.4 ± 2.6	11.5 ± 3.2
Root L	29.4 ± 7.	27.4 ± 3.5	25.3 ± 4.8	29.8 ± 8.5	27.2 ± 5.2	27.5 ± 4.3
Ab. g. dm	7.1 ± 3.0	8.6 ± 2.4	7.8 ± 1.9	8.5 ± 2.9	8 ± 2.6	7.5 ± 1.8
Root dm	3.8 ± 1.3	3.9 ± 0.8	3.7 ± 1.1	3.7 ± 1.2	3.8 ± 1.4	3.8 ± 1.0
Total dm	11 ± 4.1	12.4 ± 2.9	11.5 ± 2.7	12.1 ± 3.6	11.8 ± 3.7	11.2 ± 2.4
R:S ratio	0.58 ± 0.2	0.48 ± 0.1	0.48 ± 0.1	0.47 ± 0.2	0.48 ± 0.1	0.52 ± 0.1
A	3.7 ± 2.0	4.5 ± 2.2	3.2 ± 1.9	4.7 ± 2.4	4.6 ± 2.7	3.9 ± 2.1
E	0.17 ± 0.1	0.21 ± 0.1	0.17 ± 0.1	0.23 ± 0.1	0.23 ± 0.1	0.2 ± 0.1
gs	0.06 ± 0.02	0.07 ± 0.04	0.06 ± 0.03	0.08 ± 0.04	0.08 ± 0.04	0.07 ± 0.03
iWUE	59.4 ± 34.1	64.5 ± 13.6	54.3 ± 32.2	58.7 ± 17.1	59.7 ± 16.0	55.3 ± 19.6
$\delta^{13}C$	-29.6 ± 1.3	-29 ± 0.9	-28.8 ± 0.9	-28.7 ± 0.9	-28.8 ± 1.0	-28.6 ± 0.7
ABA	7.7 ± 5.5	6.7 ± 2.5	5.7 ± 2.4	6.1 ± 1.9	6 ± 2.4	6.4 ± 3.4
ABA-GE	4.6 ± 2.8	3.8 ± 1.9	4.5 ± 2.8	5.1 ± 3.2	4.5 ± 3.4	5.9 ± 4.3
PA	23.9 ± 12.8	18.6 ± 10.7	26.1 ± 19.8	30.3 ± 10.7	22 ± 11	27.3 ± 14.8
7'OH-ABA	8.5 ± 5.4	7.6 ± 2.6	15.1 ± 17.3	11.4 ± 10.5	10.4 ± 6.1	15.2 ± 19.2
ZOG-t	4.7 ± 2.0	6.1 ± 3.0	7.6 ± 3.8	5.6 ± 3.1	5.7 ± 3.5	9.4 ± 9.2
ZOG-c	6.7 ± 2.3	8.1 ± 2.9	6.3 ± 2.4	7.8 ± 3.7	8 ± 3.2	12.1 ± 5.8
ZR-t	27.8 ± 14.2	46.1 ± 11.6	44.7 ± 24.2	35.2 ± 28.5	36.2 ± 19.7	41.5 ± 28.2

(Continued)

TABLE 2 Continued

Traits [1]	Breeding strategy (mean ± SE)					
	OP		PM		CC	
	Mid BV	High BV	Mid BV	High BV	Mid BV	High BV
dhZR	7 ± 4.7	8.6 ± 4.1	12.3 ± 8.7	9.1 ± 8.0	8.2 ± 2.5	8.1 ± 5.0
iPR	6.9 ± 2.6	10.6 ± 4.2	10 ± 3.9	7.8 ± 2.3	7.7 ± 3.0	9.3 ± 6.2
IAA	1.1 ± 0.2	1.4 ± 0.5	1.1 ± 0.3	1.2 ± 0.3	1.1 ± 0.2	1.4 ± 0.3
GA3	2.1 ± 2.5	1.4 ± 2.0	2.4 ± 3.3	0.3 ± 0.2	0.5 ± 0.7	1.4 ± 1.6
PgGA3ox1	10.6 ± 1.5	9.9 ± 1.4	10 ± 0.9	9.8 ± 1.1	10.3 ± 1.2	10.8 ± 1.1
PgGA20ox1	8.1 ± 0.5	8.4 ± 0.8	8.1 ± 0.6	8.8 ± 0.4	8.3 ± 0.5	8.2 ± 0.5
PgDELLA1	5.3 ± 0.4	4.8 ± 0.9	5.4 ± 0.3	5.3 ± 0.3	5.3 ± 0.1	5.3 ± 1.2
PgGID1	7.7 ± 0.5	7.4 ± 0.4	7.8 ± 0.4	7.5 ± 0.3	7.8 ± 0.2	7.7 ± 0.3

[1] H, Height (cm); D, Diameter (mm); Vol, Volume (cm^3); Ap. Int. L, Apical Internode Length (cm); Root L, Root length (cm); Ab. g. dm, Above ground dry biomass (g); Root dm, R oot dry biomass (g); Total dm, Total dry biomass (g); R:S ratio, Root:Shoot ratio; A, photosynthesis (μmol CO_2 m^{-2} s^{-1}); E, transpiration (mmol H_20 m^{-2} s^{-1}); gs, stomatal conductance (mol H_20 m^{-2} s^{-1}); iWUE, intrinsic Water Use Efficiency (μmol CO_2 mol^{-1} H_20); $\delta^{13}C$, $\delta^{13}C$ (‰; VPDV); ABA, Abscisic acid (ng/g) (1.0e+2); ABA-GE, Abscisic acid glucose ester (ng/g) (1.0e+3); PA, Phaseic acid (ng/g); 7'OH-ABA, 7'-hydroxy-abscisic acid (ng/g); ZOG-t, Zeatin-O-glucoside-trans (ng/g); ZOG-c, Zeatin-O-glucoside-cis (ng/g); ZR-t, Zeatin riboside-trans (ng/g); dhZR, Dihydrozeatin riboside (ng/g); iPR, Isopentenyladenosine (ng/g); IAA, Indole-3-acetic acid (ng/g) (1.0e+2); GA3, Gibberellin 3 (ng/g) (1.0e+2). PgGA3ox1; PgGA20ox1; PgDELLA1 and PgGID1 are the GA-related genes analyzed in this study using their absolute transcript levels; this unit is the transcript mean of each target gene divided by the transcript mean of the reference gene (PgTIF5A) per reaction.
Last measurements and harvesting were done on Day 92 (28 April 2020).

Families 129 and 138 (High BV) showed the same iWUE for all breeding groups (Figure 1L). In some cases, CC showed the highest ABLUPs for diameter, volume, root dry biomass, and total dry biomass (family 754), and PM showed the highest ABLUPs for transpiration and stomatal conductance (family 138). In other cases, OP showed the highest ABLUPs for height (family 129), apical internode length and photosynthesis (family 138), root:shoot ratio (family 754), iWUE (family 122).

Genetic variation among individuals within families for growth, physiology, hormone levels and gene expression

Significant correlations were found between growth, biomass, gas exchange, $\delta^{13}C$ and hormone traits (Table 4). As expected, the highest correlations were found between height vs. apical internode length (r=0.82, P<0.01), height vs. total dry biomass (r=0.78, P<0.01), volume

TABLE 3 Results from the 'Family Model' for the growth, biomass and gas exchange parameters (n=180, 10 blocks, 18 families) in 2-year-old white spruce seedlings from three different breeding strategies (OP, open pollination; PM, polymix pollination; CC, controlled crosses).

Traits	*P*-value (Breeding Strategy)	Variance Component Family	Variance component Residual	h^2 (± SE)
Height	<0.01	4.68	18.12	**0.21** (± 0.11)
Diameter	<0.01	0.05	1.16	0.04 (± 0.01)
Volume	<0.01	3.17	28.36	**0.11** (± 0.09)
Apical internode length	<0.01	0.46	6.15	0.07 (± 0.02)
Root length	<0.01	4.3	36.84	**0.10** (± 0.09)
Above ground dry biomass	<0.01	0.92	4.89	**0.16** (± 0.03)
Root dry biomass	<0.01	0.02	1.55	0.01 (± 0.01)
Root-Shoot ratio	<0.01	0.002	0.0016	**0.14** (± 0.02)
Total dry biomass	<0.01	1.15	10.09	**0.12** (± 0.03)
Photosynthesis	<0.01	0.01	0.51	0.02 (± 0.01)
Transpiration rate	<0.01	0.0003	0.008	0.04 (± 0.01)
Stomatal conductance	<0.01	0.004	0.12	0.03 (± 0.02)
Intrinsic water use efficiency	<0.01	0.013	0.45	0.03 (± 0.01)

For the genetic model, we used the pedigree matrix (A) to obtain the Best Linear Unbiased Predictions (BLUPs) for the family as random factor (see Materials and Methods, and Figure 1). The table includes the *P*-values for the breeding strategy (fixed factor) from the Wald-F test, the variance component of family and residual, and the narrow-sense heritability (h^2) for each trait. Moderate narrow-sense heritabilities are indicated in bold. Last measurements and harvesting were done on Day 92 (28 April 2020).

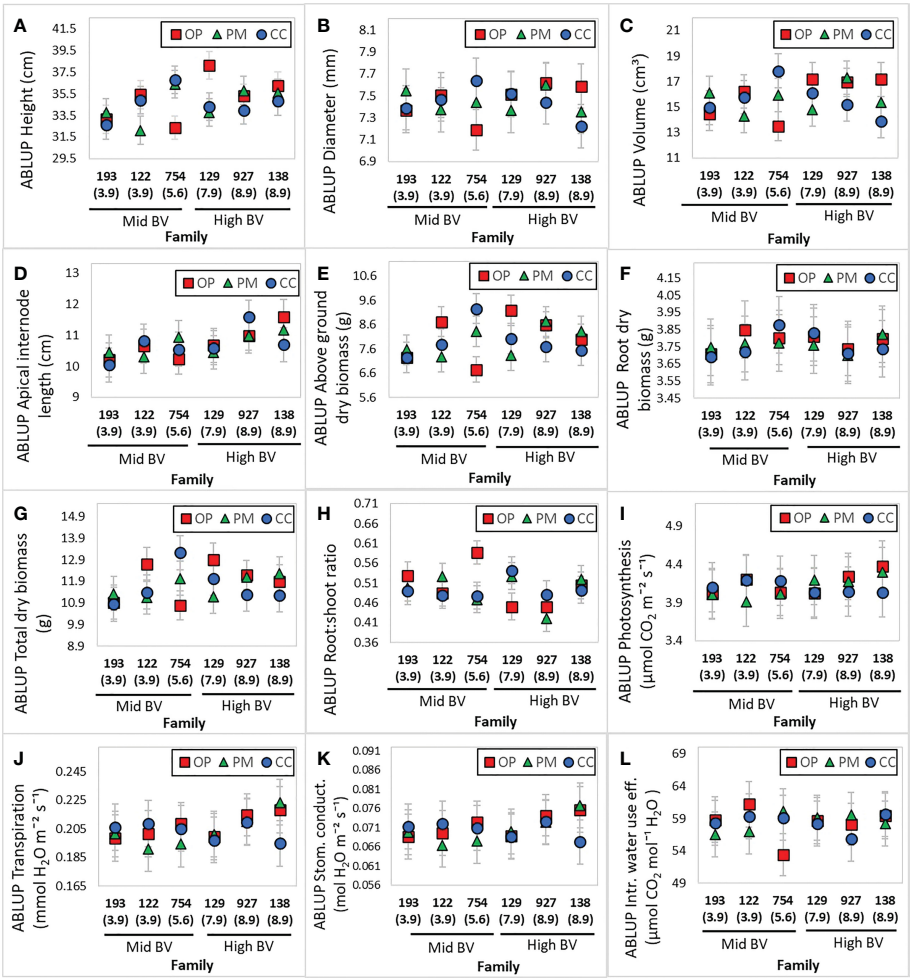


FIGURE 1
Best linear unbiased predictions (BLUPs) of the 'family' random effect from the 'Family Genetic Model' for growth, biomass and gas exchange parameters (n=180, 10 blocks, 18 families) in 2-year-old white spruce seedlings. **(A)** Height, **(B)** Diameter, **(C)** Volume, **(D)** Apical internode length, **(E)** Above ground dry biomass, **(F)** Root dry biomass, **(G)** Total dry biomass, **(H)** Root:shoot ratio, **(I)** Photosynthesis, **(J)** Transpiration, **(K)** Stomatal conductance, **(L)** intrinsic water use efficiency. For the model, we used the pedigree matrix (also called A matrix) with the different relationships between half- and full- sibs among the three different breeding strategies: OP (open pollination), PM (polymix pollination), CC (controlled crosses). Each figure includes the mid and high BV groups. Families were ordered from the lowest to the highest BV (value under each family). Last measurements and harvesting were done on Day 92 (28 April 2020).

vs. total dry biomass (r=0.75, P<0.01), photosynthesis vs. transpiration (r=0.78, P<0.01), stomatal conductance vs. above ground dry biomass (r=0.62, P<0.01), *PgGA3ox1* vs. ABA (r=0.63, P<0.01), ABA vs. PA (r=0.76, P<0.01) (Table 4). We found significant correlations between height vs. photosynthesis (r=0.47, P<0.01), height vs. root:shoot ratio (r=-0.56, P<0.01), height vs. $\delta^{13}C$ (r=0.51, P<0.001), apical internode length vs. *PgGA3ox1* (r=0.35, P<0.05), apical internode length vs. IAA (r=0.48, P<0.01), apical internode length vs. ABA (r=0.42, P<0.05), apical internode length vs. PA (r=0.76, P<0.01) (Table 4). Scatterplots for height vs. total dry biomass (Figure 2A), photosynthesis (Figure 2B), root:shoot ratio (Figure 2C), and $\delta^{13}C$ (Figure 2D), showed considerable genetic variability among all seedlings analyzed in the three breeding strategies. The three replications for family F138 from PM showed similar values for height, but a range between 11 and 16 g for total dry biomass (Figure 2A), while the repetitions for family F138 from OP showed a much greater range for height, varying between 32 and 40 cm (Figures 2A–D). The three plants analyzed for families 122 and 193 from PM showed a similar response among them for photosynthesis, but family 927, from CC, had a range from 0.4 to 6.6 µmol CO_2 m^{-2} s^{-1} (Figure 2B). Trees within families for root:shoot ratio and $\delta^{13}C$ also showed contrasting values, such as family F193 from OP and 927 from CC (Figure 2C), family 927 from PM and F138 from CC (Figure 2D). In general, however, CC and PM plants showed the highest values for total dry biomass, photosynthesis, $\delta^{13}C$, and the lowest root:shoot ratios (Figures 2A–D). Scatterplots for apical internode length vs. *PgGA3ox1* (Figure 3A), IAA (Figure 3B), ABA (Figure 3C), and ABA with PA (Figure 3D), again showed large genetic variability, but with particular individuals outstanding in their performance at the molecular level within a given breeding strategy. For example, two individuals from F927 coming from CC excelled in apical internode growth, as well as gene expression of *PgGA3ox1* and levels of IAA, ABA, PA, and one individual from F927, coming from OP, performed poorly not only for growth but also for hormone levels (Figures 3A–D). In general, individuals from the different growth

TABLE 4 Pearson's correlation matrix between growth, biomass and gas exchange traits for white spruce seedlings (n=36, 3 blocks, 12 families).

Trait [+]	H	D	V	AIL	E	gs	A	iWUE	ADM	RDM	RL	R:S	TDM	$\delta^{13}C$	GA3ox	ABA	PA	ZOG	iPR	IAA
D	**0.35***																			
V	**0.64****	**0.93****																		
AIL	**0.82****	0.29	**0.55****																	
E	**0.63****	0.21	**0.39***	**0.47****																
gs	**0.63****	0.18	**0.36***	**0.46****	**0.99****															
A	**0.47****	0.21	**0.32***	**0.36***	**0.78****	**0.81****														
iWUE	**-0.39***	0.04	-0.12	**-0.33****	**-0.41***	**-0.38***	0.14													
ADM	**0.79****	**0.57****	**0.77****	**0.55****	**0.65****	**0.62****	**0.48****	-0.25												
RDM	**0.49****	**0.35***	**0.46****	0.29	0.28	**0.32***	0.31	-0.02	**0.56****											
RL	**0.33***	0.09	0.25	0.22	0.08	0.04	-0.03	-0.02	**0.41***	0.25										
R:S	**-0.56****	-0.31	**-0.47****	**-0.41****	**-0.53****	**-0.48****	**-0.38***	0.26	**-0.69****	0.12	-0.17									
TDM	**0.78****	**0.56****	**0.75****	**0.52****	**0.61****	**0.61****	**0.47****	-0.21	**0.97****	**0.76****	**0.40***	**-0.51****								
$\delta^{13}C$	**0.51****	0.24	**0.38***	**0.41****	0.21	0.22	0.31	-0.06	**0.43****	**0.48****	0.03	0.26	**0.51****							
GA3ox	0.11	0.16	0.16	**0.35***	-0.06	-0.04	-0.07	-0.04	0.02	0.03	-0.15	0.06	0.03	**0.34***						
ABA	0.31	0.03	0.13	**0.42****	0.08	0.12	0.12	-0.11	0.08	0.02	-0.07	-0.06	0.07	**0.41****	**0.63****					
PA	0.24	-0.01	0.06	**0.34***	0.11	0.12	0.03	-0.21	0.02	-0.04	-0.04	-0.03	0.01	0.21	**0.37***	**0.76****				
ZOG	0.18	-0.11	-0.03	-0.09	0.09	0.12	0.01	-0.17	0.15	0.16	0.01	-0.11	0.17	0.22	0.06	-0.07	-0.14			
iPR	0.09	-0.06	-0.01	0.09	-0.11	-0.07	-0.13	0.05	0.13	0.15	0.09	0.01	0.15	0.02	-0.13	-0.05	-0.16	0.11		
IAA	**0.41***	-0.04	0.11	**0.48****	0.14	0.12	0.06	-0.21	0.26	0.06	0.05	-0.24	0.22	**0.42****	**0.37***	**0.42****	**0.38***	0.21	0.31	
GA3	-0.15	-0.16	-0.14	-0.21	-0.01	0.02	-0.23	-0.24	-0.04	-0.12	-0.12	0.09	-0.07	**-0.35***	-0.28	-0.27	-0.02	0.05	**0.33***	-0.2

[+]H, Height; D, Diameter; V, Volume; AIL, Apical Internode Length; E, transpiration; gs, stomatal conductance; A, photosynthesis; iWUE, intrinsic water use efficiency; ADM, Aboveground Dry Biomass; RDM, Root Dry Biomass; RL, Root Length; R:S, Root:Shoot ratio; TDM, Total Dry Biomass; GA3ox, Gibberellin 3-oxidase gene expression; ABA, Abscisic acid; PA, Phaseic acid; ZOG, Zeatin-O-glucoside (trans); iPR, Isopentenyladenosine; IAA, Indole-3-acetic acid; GA3, Gibberellin 3.
Asterisks and gray shadows indicate significant values at $*P \leq 0.05$ and $**P \leq 0.01$.

10.3389/fpls.2023.1052425

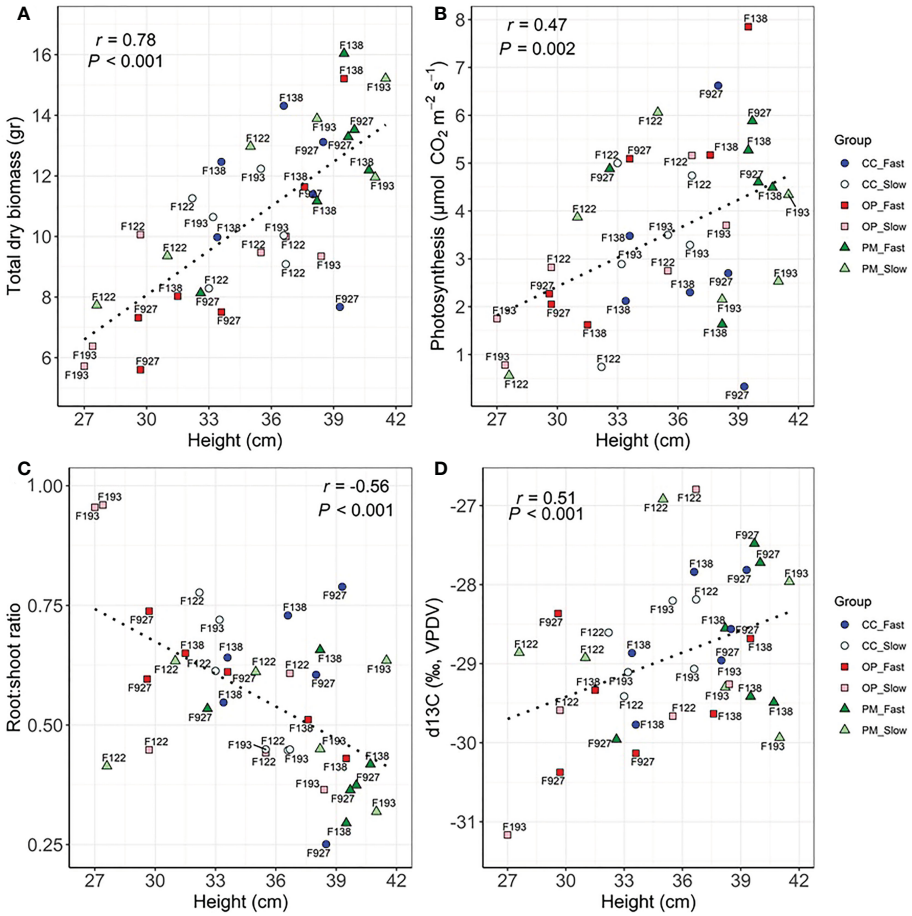


FIGURE 2
Scatterplots between **(A)** height and total dry biomass, **(B)** height and photosynthesis, **(C)** height and Root:shoot ratio, **(D)** height and $\delta^{13}C$ among groups (controlled crosses (CC), polymix (PM) and open pollination (OP) for slow (F193, F122) and fast (F927, F138) growth in 2-year-old white spruce seedlings. Pearson's correlation (r) and p-values (*P*) corresponding to each correlation are shown in each graphic. Each point corresponds to each tree (n=36, 3 blocks, 12 families) (see family number for each point in the graphic). The six groups are detailed for each family with colours and shapes. Last measurements and harvesting were done on Day 92 (28 April 2020).

groups and breeding strategies showed significant overlap in hormone levels (Figures 3A–D).

Interaction effects of breeding strategies and growth groups on growth, biomass, gas exchange, hormones and expression of GA-related genes

Height, diameter, volume, and root length did not show statistically significant main or interaction effects, but apical internode length showed a main effect for growth group (P<0.01) and an interaction effect (P=0.03) (Table 5). Also, above ground dry biomass, root:shoot ratio, iWUE and $\delta^{13}C$ showed statistically significant main effects for growth groups and breeding strategies with no interaction effects (Table 5). Regarding hormone levels and gene expression, abscisic acid glucose ester, phaseic acid, zeatin-O-glucoside-trans, zeatin riboside-trans, indole-3-acetic acid, gibberellin 3, and *PgGA3ox1* showed statistically significant main effects for breeding strategy and interaction effects (Table 6). Controlled crosses from the fast-growth group showed the largest apical internode length, and higher amounts of ABA, PA, IAA, and *PgGA3ox* gene expression, compared to the slow-growth group (Figures 4A–D, G). Controlled cross pollination and PM showed lower levels of zeatin in the fast-growth group compared to the slow-growth group (Figures 4E, F). In general, open pollination showed similar hormone levels of ABA, PA, IAA, zeatin, GA3, *PgGA3ox* expression for the fast and slow growth groups, but presented a higher apical internode length in the fast-growth group compared to the slow-growth group (Figure 4). Also, open pollination showed a statistically significant higher root:shoot ratio, iWUE and $\delta^{13}C$ compared to polymix pollination and controlled crosses (Supplementary Material 7).

Gene expression of GA-related genes in the apical internode of white spruce seedlings from three different breeding strategies

The *PgGA20ox*, *PgDELLA*, and *PgGID* gene expression did not show statistically significant differences between OP, PM and CC, in the fast- and slow-growth groups (Figure 5; Table 6). The only

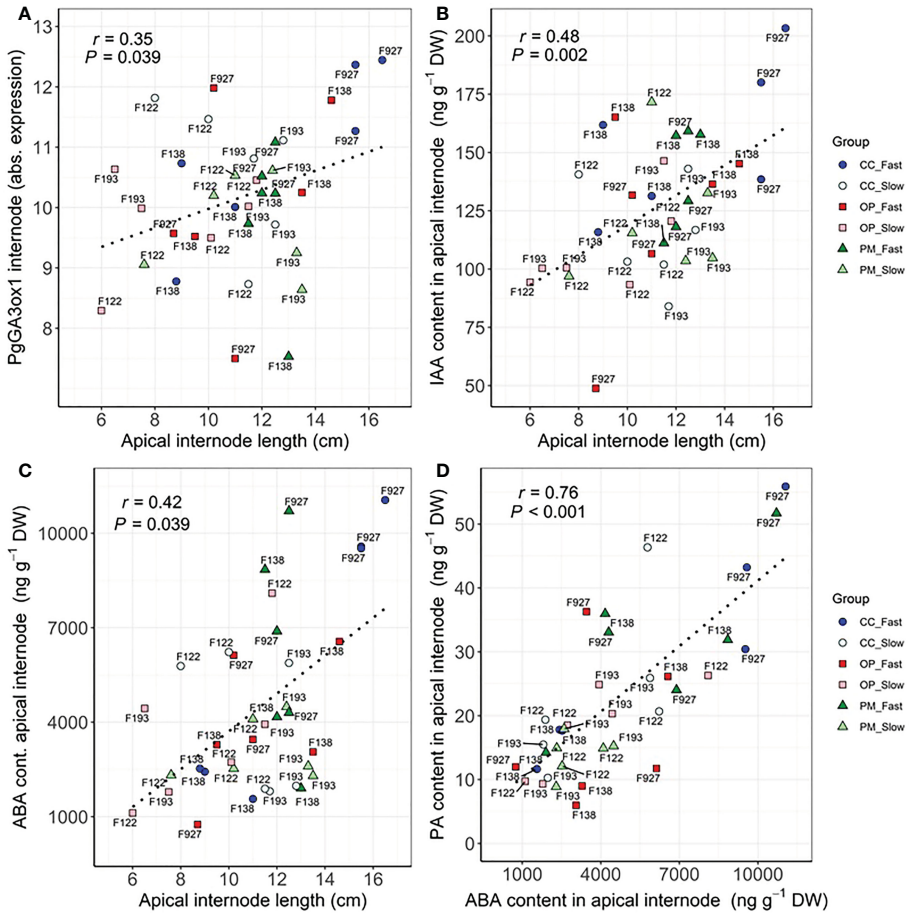


FIGURE 3
Scatterplots between **(A)** apical internode length and *PgGA3ox1* expression in apical internode, **(B)** apical internode length and IAA content in apical internode, **(C)** apical internode length and ABA content in apical internode, **(D)** ABA content and PA content in apical internode among groups (controlled crosses (CC), polymix (PM) and open pollination (OP) for slow (F193, F122) and fast (F927, F138) growth in 2-year-old white spruce seedlings. Pearson's correlation (r) and p-values (*P*) corresponding to each correlation are shown in each graphic. Each point corresponds to each tree (n=36, 3 blocks, 12 families) (see family number for each point in the graphic). The six groups are detailed for each family with colours and shapes. Last measurements and harvesting were done on Day 92 (28 April 2020).

statistically significant difference was found for *PgGA3ox1* gene expression between breeding strategies, for the fast-growth group, with CC exhibiting 4-fold more expression than OP (Figure 5A). Also, the *PgDELLA* and *PgGID* gene expression was 1.5-fold higher in CC than OP and PM for the fast-growth group, which was not statistically significant (Figures 5C, D).

Principal components analysis for growth and molecular traits among the breeding strategies and growth groups

For this analysis, we considered six groups for clustering: OP, PM and CC pollination in the fast- and slow-growth groups (Figure 6). The first principal component associated with growth, biomass and gas exchange parameters explained 44.2% of the variance, whereas the second principal component associated with the molecular traits explained 29.4% of the variance (Figure 6). The position of the six groups in the PCA quadrant shows two clear patterns in the growth, physiological and molecular mechanisms (Figure 6). The first cluster represents the fast-growth for CC and PM, with the highest values for height, diameter, volume, apical internode length, above ground dry biomass, A, E, gs, PA, ABA, IAA, GA3, and *GA3ox1* (green oval, Figure 6). The second cluster represents all the slow-growth (for OP, CC, and PM) and fast-growth families for OP with the highest values of iWUE, $\delta^{13}C$, root length, root dry biomass, zeatin and iPR (purple oval, Figure 6). The OP and CC with slow-growth exhibited the lowest productivity above ground with the highest iWUE and root development, while CC with the fast-growth group showed the best productivity with the lowest iWUE and root length (Figure 6).

Discussion

Genetic variation of trees from different breeding strategies leads to opportunities in selecting new material

The results presented in this study indicate that height, above ground dry biomass and root:shoot ratio are under moderate

TABLE 5 Results (*P*-values) from the ANOVA mixed model analysis of 'Growth group' fixed-effect (fast and slow growth), 'Breeding strategy' fixed-effect (open pollination, polymix pollination, controlled crosses), and 'Growth group' by 'Breeding strategy' interactions on growth, biomass and gas exchange traits (n=120, 10 blocks, 12 families) in 2-year-old white spruce seedlings.

Traits	Growth group	Breeding strategy	Growth group x Breeding strategy
Height	0.109	0.349	0.881
Diameter	0.626	0.583	0.995
Volume	0.396	0.297	0.887
Apical internode length	**<0.01****	0.164	**0.032***
Root length	0.606	0.113	0.388
Above ground dry biomass	**0.041***	0.183	0.782
Root dry biomass	0.714	0.952	0.758
Root-Shoot ratio	**0.013***	**0.031***	0.298
Total dry biomass	0.232	0.339	0.958
Photosynthesis	0.399	0.921	0.552
Transpiration rate	0.361	0.217	0.916
Stomatal conductance	0.366	0.243	0.962
Intrinsic water use efficiency	0.686	**0.034***	0.181
$\delta^{13}C$	0.181	**0.038***	0.824

Families 927 and 138 were used for the 'fast' category, and families 193 and 122 were used for the 'slow' category, each with 30 trees. Asterisks and shaded cells indicate significant values at $*P \leq 0.05$ and $**P \leq 0.01$. Last measurements and harvesting were done on Day 92 (28 April 2020).

TABLE 6 Results (*P*-values) from the ANOVA mixed model analysis of 'Growth group' fixed-effect (fast and slow growth), 'Breeding strategy' fixed-effect (open pollination, polymix pollination, controlled crosses), and 'Growth group' by 'Breeding strategy' interactions on hormone levels and expression of GA genes (n=36, 3 blocks, 12 families) in 2-year-old white spruce seedlings.

Traits	Growth group	Breeding strategy	Growth group x Breeding strategy
Abscisic acid (ABA)	**0.013***	0.518	0.647
Abscisic acid glucose ester (ABA-GE)	**<0.01****	**0.041***	**0.012***
Phaseic acid (PA)	**0.015***	**0.035***	**0.043***
7'-hydroxy-abscisic acid (7'OH-ABA)	0.825	0.631	0.287
Zeatin-O-glucoside-trans (ZOG-t)	**<0.01****	**0.024***	**<0.01****
Zeatin-O-glucoside-cis (ZOG-c)	**0.014***	0.829	0.426
Zeatin riboside-trans (ZR-t)	**0.034***	**0.037***	**0.039***
Dihydrozeatin riboside (dhZR)	0.061	0.364	0.366
Isopentenyladenosine (iPR)	**<0.01****	**0.031***	0.238
Indole-3-acetic acid (IAA)	**0.011***	**<0.01****	**0.032***
Gibberellin 3 (GA3)	0.061	**0.039***	**0.039***
PgGA3ox1	**0.033***	**0.024***	**0.027***
PgGA20ox1	0.508	0.741	0.655
PgDELLA1	0.701	0.502	0.387
PgGID1	0.781	0.626	0.464

The levels of the 11 hormones and gene expression of the four genes were measured in the apical internode (see Materials and Methods). Families 927 and 138 were used for the 'fast' category, and families 193 and 122 were used for the 'slow' category, each with 30 trees. Asterisks and gray shadow indicate significant values at $*P \leq 0.05$ and $**P \leq 0.01$. Last measurements and harvesting were done on Day 92 (28 April 2020).

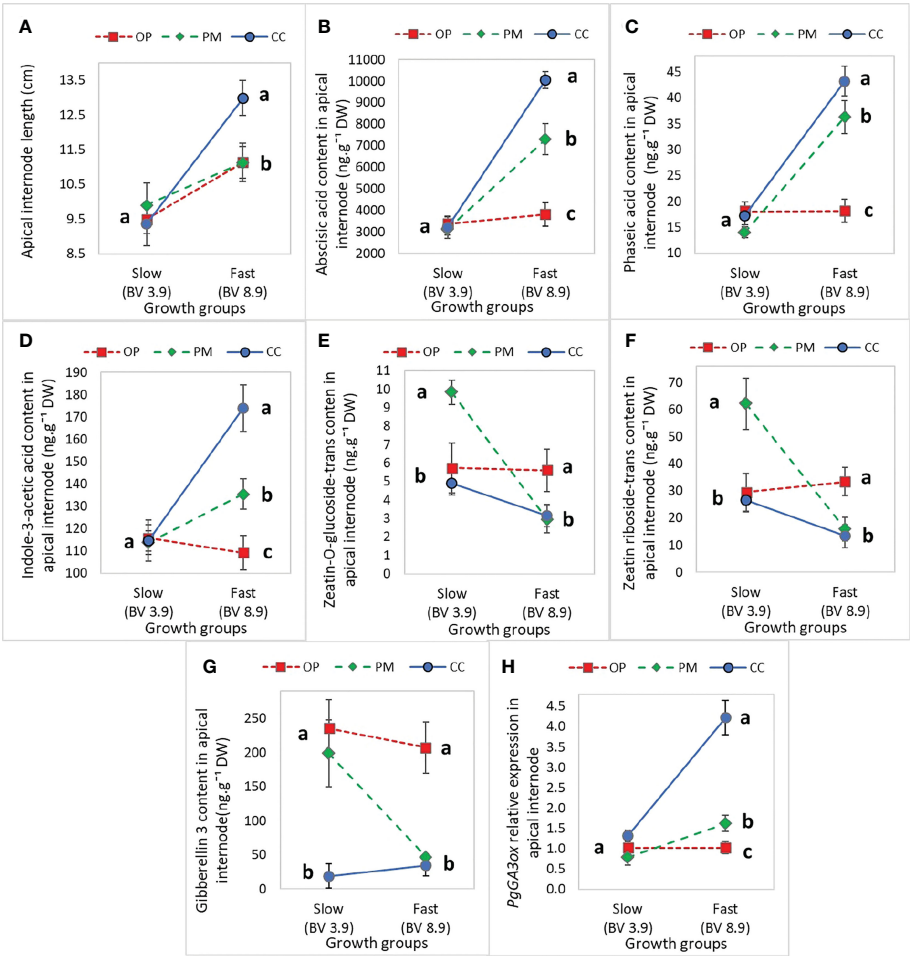


FIGURE 4
Mean (± SE) effect of the interaction between 'Growth group' and 'Breeding strategy' on growth and hormone levels in the apical internode in 2-year-old white spruce seedlings for **(A)** The Apical internode length (n=120, 10 blocks, 12 families), **(B)** Abscisic acid glucose ester in the apical internode, **(C)** Phaseic acid in the apical internode, **(D)** Indole-3-acetic acid, **(E)** Zeatin-O-glucoside-trans, **(F)** Zeatin riboside-trans, **(G)** *PgGA3ox* gene expression, and **(H)** GA3 hormone level (n=36, 3 blocks, 12 families). The 'Growth groups' are slow (F193+F122) and fast (F927+F138), and the 'Breeding strategies' are OP (open pollination), PM (polymix pollination), and CC (controlled crosses) (see Materials and Methods). Mean values are represented by squares (open pollination), triangles (polymix pollination) and circles (controlled crosses). *PgTIF5A* (GenBank accession number DR448953) was used as the control gene. Letters indicate differences (mean) between the breeding strategies using all values from the growth groups (fast, slow) with a Tukey's test, at 95% confidence level. Last measurements and harvesting were done on Day 92 (28 April 2020).

genetic control (heritabilities between 14-21%) for development of 2-year-old white spruce seedlings from three different breeding strategies (Table 3). Furthermore, for total dry biomass, volume and root length, the heritability values were close to 10%, meaning nearly 90% of the variation may be caused by environmental factors, as previously shown (Raj et al., 2006). In trees, narrow-sense heritability for height can range between 0.1 and 0.5 in seedlings at an early stage (Scotti-Saintagne et al., 2004; Sotelo Montes et al., 2007; Tripiana et al., 2007; Bouffier et al., 2008; Callister and Collins, 2008; Ward et al., 2008; Scotti et al., 2010). Heritability of 3-year-old white spruce families from an improvement program in Quebec, Canada ranged between 0.17 to 0.45 for height (Li et al., 1993) and in another Quebec study, heritability was 0.26 for height and 0.14 for diameter (Wahid et al., 2013). In *Pinus rigida*, seedlings at a very young age obtained from trees growing in the Waiden Woods, Massachusetts, showed extensive variation for growth, in contrast to their hypothesis (Raj et al., 2006). Our study also showed heritabilities close to 0 for all gas exchange parameters (Table 3), similar to other studies (Scotti-Saintagne et al., 2004; Brendel et al., 2008; Scotti et al., 2010). Divergence in heritability estimates could be partly explained by differences in seedling age and number of families under investigation, and can only be applied to a particular population growing in a particular environment at a particular point in time (Rweyongeza et al., 2010; Carles et al., 2012). The ABLUP values obtained for growth and physiology traits of the white spruce seedlings in our study showed substantial genetic variability between families from the different breeding strategies (Figure 1). Our experiment found specific faster- and slower-growing families from OP and CC, with different developmental and physiological characteristics. Previous authors have discussed the possible evolutionary genetic meaning of the natural variation when domesticating conifers (Scotti et al., 2010; Gepts, 2014). Although family 754 showed the highest ABLUP for root:shoot ratio from OP among all families (Figure 1H), significant variability

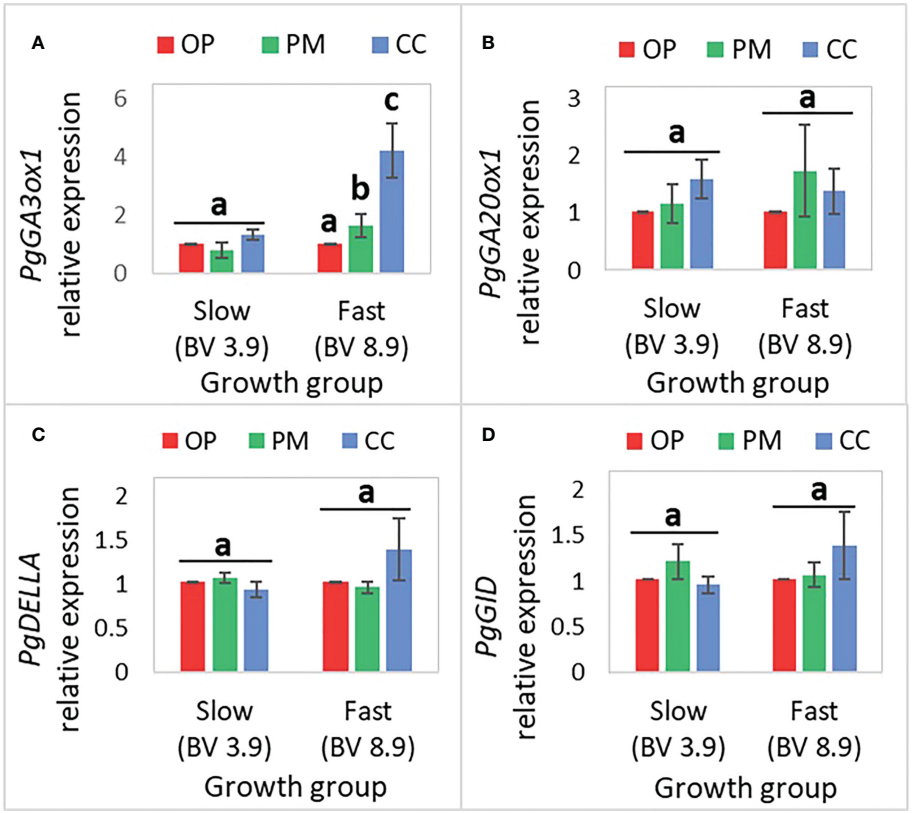


FIGURE 5
Mean (± SE) of the relative expression of GA-related genes in the apical internode (n=36, 3 blocks, 12 families) in 2-year-old white spruce seedlings for **(A)** *PgGA3ox1*, **(B)** *PgGA20ox1*, **(C)** *PgDELLA*, **(D)** *PgGID*. The 'Growth groups' are slow (F193+F122) and fast (F927+F138), and the 'Breeding strategies' are OP (open pollination), PM (polymix pollination), and CC (controlled crosses) (see Materials and Methods). *PgTIF5A* (GenBank accession number DR448953) was used as the control gene. Letters above bars indicate differences (mean) between the breeding strategies in each of the growth groups (fast, slow) with a Tukey's test, at 95% confidence level. Last measurements and harvesting were done on Day 92 (28 April 2020).

in dry root biomass between families and breeding methods is clear (Figures 1F, H), showing large genetic variation in radicular biomass allocation and potential nutrient uptake among families, as previously described (Theodorou and Bowen, 1993). The vast genetic variability among all the seedlings analyzed within families and breeding strategies in our work (Figure 2) is consistent with previous studies. In trials performed in Quebec, Canada, heights for 2-year-old white spruce clones and seedlings varied from 14.4 to 31.8 cm and from 15.8 to 24.3 cm, respectively (Lamhamedi et al., 2000). In this study, the three seedlings from F138 (polymix pollination) showed a range of 11-16 g for total dry biomass, and the three seedlings from F138 (open pollination) exhibited a range of 32-40 cm for height (Figure 2). Large ranges were also found among seedlings from CC and OP in F927 for gas exchange traits, hormone levels and *PgGA3ox1* gene expression (Figures 2, 3). Significant clonal variation was also observed when studying differences of plants from seedlings and clones: for many variables (height, dry biomass of new roots, needle dry biomass and branch density), and differences among clones were significantly greater than differences among seedlings within a family (Lamhamedi et al., 2000). Variation in growth and physiology reflected genetically determined differences among individuals within a family (Lamhamedi et al., 2000). Tree breeding can result in a bigger pool of phenotypic variation that can be managed wisely if the variation is mainly due to genetic rather than environmental effects (Gepts, 2014). These findings suggest there is minimal risk applying any particular breeding strategy in white spruce, as our ability to in fact manipulate the genetics in the first generation is minimal due to the enormous natural variability in these trees. Perhaps our concern with reducing genetic variability, at the expense of gain, is overstated in the early generation breeding stages in conifers, and several generations will be needed, unless a trait is highly heritable (e.g. stem straightness) or has limited variability.

Trade-offs between growth, carbon allocation and gas exchange of plants obtained from different breeding strategies

Correlation analyses showed a general structure with taller white spruce stems associate with larger biomass, photosynthesis, transpiration, water use efficiency ($\delta^{13}C$) and higher levels of hormones (Table 4). This seems to hold for both phenotypic and genetic correlations, thus suggesting that these traits share a genetic bases, as previously found in *Sextonia rubra* (Scotti et al., 2010). In

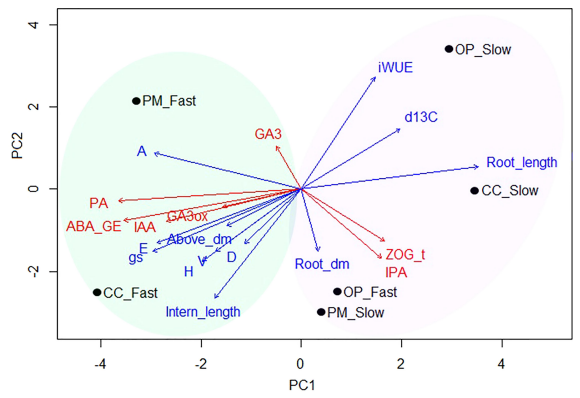


FIGURE 6
Principal component analysis (PCA) biplot of combined data sets with post-hoc fit of growth and gas exchange traits (PC1 = 44.2%, blue arrows), and molecular traits (PC2 = 29.4%, red arrows) over the six groups (black dots) (n=36, 3 blocks, 12 families). The vector lengths represent the importance of each variable along the two plotted axes in relation to all possible principal components, while distances between vectors and families indicate how much variation is explained by each vector. Blue arrows: H, height; D, diameter; V, volume; Intern_length, apical internode length; A, photosynthesis; E, transpiration; gs, stomatal conductance; iWUE, intrinsic water use efficiency; Above_dm, above ground biomass; Root_dm, root biomass; Root_length, root length; d13C, $\delta^{13}C$ isotopes. Red arrows: PA, Phaseic acid; ABA_GE, Abscisic acid glucose ester; IAA, Indole-3-acetic acid; ZOG_t, Zeatin-O-glucoside-trans; IPA, Isopentenyladenosine; GA3, Gibberellin 3; GA3ox, *PgGA3ox1* gene expression. Groups: CC_Fast, controlled crosses with fast growth (F927+F138); CC_Slow, controlled crosses with slow growth (F193; F122); PM_Fast, polymix pollination with fast growth (F927+F138); PM_Slow, polymix pollination with slow growth (F193; F122); OP_Fast, open pollination with fast growth (F927+F138); OP_Slow, open pollination with slow growth (F193; F122).

general, we found a trade-off in developmental traits between breeding strategies, with controlled crosses from the fast-growth group showing the best apical growth, but in some cases, open pollination from the different growth groups showing the best root development (Figures 2–4). In our study, some individuals showed a preference to allocate photoassimilates to roots rather than shoots (e.g. three plants from F122, PM and OP, slow growth), but other seedlings showed the best height with a lower root:shoot ratio (e.g. plants from F927 and F138, PM, fast-growth group), which are not suitable for selection and could be at risk of not developing properly when older (Figure 2C). On the other hand, some individuals showed greater height, root:shoot ratios close to 1.0, greater stomatal opening, and higher water use efficiency (e.g. one plant from F927 and another from F138, CC, fast growth) (Figures 2, 4, 5); identifying themselves as the most suitable for selection suggesting both growth and drought resistance could be selected for simultaneously. Differences in developmental and growth partitioning have been reported previously among clones and families of white spruce during initiation, maturation and germination (Park et al., 1993; Park et al., 1994; Lamhamedi et al., 2000). Previous studies indicated that white spruce families exhibiting inferior height growth showed higher net photosynthesis under drought stress than families exhibiting intermediate and superior height growth (Bigras, 2005). Gas exchange and growth usually show a positive correlation at the intraspecific level, but trade-offs between light and carbon acquisition for individuals within a family have also been reported (Scotti et al., 2010). In our study, negative correlations between photosynthesis and height were shown suggesting a tradeoff in allocation of resources (Figure 2B). Trees that are genetically predisposed to slower growth, may require less water to maintain adequate evapotranspiration, and could therefore be more tolerant to drought events (Moran et al., 2017; Six et al., 2021). As in all tree species, the rate of photosynthesis and stomatal conductance in spruce trees is influenced by light, temperature, atmospheric humidity, CO_2 concentration, soil water availability and phenology (Lamhamedi and Bernier, 1994; Hébert et al., 2011). In our study, white spruce seedlings showed stomatal conductance values from 0.065 to 0.077 mol H_20 m^{-2} s^{-1}, higher than that reported in 2-year-old black spruce seedlings (0.03 mol H_20 m^{-2} s^{-1}) grown in outside sand beds (Lamhamedi and Bernier, 1994). Stomatal conductance influences net photosynthesis by controlling the amount of CO_2 that can enter the mesophyll, with stomatal limitation to net photosynthesis becoming important only at low values of stomatal conductance (Lamhamedi and Bernier, 1994; Hébert et al., 2011). Furthermore, root growth in spruce trees has been shown to decline during the period of shoot growth, as shoot growth itself uses most of the stored and current photosynthates, but at other times of the year, soil temperature is the major regulator of root growth (Lamhamedi and Bernier, 1994). Trees with higher root development generally show higher photosynthesis rates and such individuals can re-establish soil–root contact more rapidly than poor-rooting trees following transplanting (Lamhamedi et al., 2000). It has been described that *Picea* species overcome root deformation by forming adventitious roots after planting, which are critical for water and nutrient uptake, whereas deeper roots are more essential for stability and coping during drought episodes (Lamhamedi et al., 2000).

Trade-offs between hormones levels and gene expression of plants obtained from different breeding strategies

This study showed that *PgGA3ox1* expression was higher in fast vs. slow groups, as reported in previous experiments (Galeano and Thomas, 2020). *PgGA3ox1* showed the highest expression among the other GA-related genes (i.e. *PgGA20ox1*, *PgDELLA*, *PgGID*) as observed in one of our previous studies (Galeano and Thomas, 2020). *PgGA3ox1* is, again, showing a strong influence in growth compared to other GA-related genes, particularly in fast-growing individuals. In general, our study found strong correlations between apical internode length, ABA, PA levels (one of ABA's catabolites), IAA, *PgGA3ox* gene expression, and significant correlations of GA_3 with $\delta^{13}C$ and iPR (Table 4). In particular, we found a pattern of greater accumulation of *PgGA3ox*, IAA, ABA, PA in controlled crosses in the fast-growth group compared to the other breeding strategies (Figures 3, 4, 6). In addition, longer root length, root dry biomass, ZOG_t, and iPR levels were higher in OP material, and the slow-growth group (Figure 6). Our study suggests a combined effect of PA (terpenoid), IAA (auxin), ABA, GA_3 in stem elongation of

white spruce seedlings, and high amounts of IAA could potentially be inhibiting Zeatin synthesis, as previously described (Lulsdorf et al., 2013; Dilworth et al., 2017; Lorrai et al., 2018; Shu et al., 2018; Akhtar et al., 2020). Previous studies in white spruce seedlings also showed different profiles for ABA, cytokinins, auxin and expression of GA-related genes related to bud development (Kayal et al., 2011). In this study, GA_3 hormone content could potentially experience feedback regulation of *PgGA3*, and be regulated by ABA in improved material (e.g. controlled crosses), since ABA and GA antagonistically regulate stem elongation (Cline and Harrington, 2007; Lorrai et al., 2018; Shu et al., 2018; Akhtar et al., 2020). In unimproved material (open pollination), it is possible that regulation is not as strong as in improved seedlings. In Arabidopsis, a disruption of four genes encoding iPR by T-DNA insertion showed failure to form cambium and reduced thickening of the root and stem, demonstrating the relevance of cytokinins for normal development of both the root and shoot (Matsumoto-Kitano et al., 2008). Previous studies have explained that family differences in expression of gibberellin-related genes and endogenous hormone levels may explain much of the natural variation in tree stem growth capacity (Stoehr et al., 1998; Park et al., 2014). *PgGA3ox*, IAA, ABA, PA profiling combined with the measurement of growth and physiological traits could have the potential to accelerate the selection of spruce families at an early stage for rapid growth, as previously suggested (Park et al., 2014; Galeano and Thomas, 2020). Consequently, the identification and monitoring of hormone levels and expression of GA-related genes that control traits relevant to tree domestication is a powerful tool, especially as our knowledge of tree-specific processes is still insufficient and in its infancy (Singh et al., 2021). In any event, other plant hormones not studied here could show correlations between their concentration levels, gene expression, and growth parameters in spruce seedlings.

Implications of this study

The term "Physiological breeding strategy" was recently introduced, which was successfully adapted in legume crop improvement to narrow the gap between breeders and physiologists through collaborative approaches to understand complex traits, potential of physiology based approaches and how they impact yield gains and abiotic stress tolerance in these crops (Shunmugam et al., 2018). Certainly, tree breeders and physiologists have the challenge of studying and understanding the phenotypic plasticity of simple and complex developmental traits, the relationship among them, and their importance as determinants of growth of mature forest trees (Raj et al., 2006; Singh et al., 2021). Therefore, early selection of superior clones based on both physiological and morphological characteristics is feasible (Stoehr et al., 1998; Lamhamedi et al., 2000). In Quebec, Canada, the identification of strong positive genetic correlations allowed diameter to be used as an effective method for indirect selection of white spruce seedlings with heavier root systems (Carles et al., 2012). In Ontario, Canada, root establishment and survival in 2-year-old *Pinus resinosa* seedlings were positively correlated with the length of needles (Paterson and Faylel, 1984). In Georgia, USA, loblolly pine seedlings from a fast-growing provenance showed higher rates of net photosynthesis than seedlings from slow-growing provenances indicating the potential of this physiological trait for selection (Boltz et al., 1986). In our study, several seedlings had both greater root growth and high stem elongation rates (Figure 1), indicating that selection of white spruce trees based on root growth would not exclude genotypes with greater above ground growth, as described in previous studies (Lamhamedi et al., 2000). "Tree diversity breeding" is another new approach that seeks to amplify biodiversity in production systems to address global challenges and maximize linkages between existing tree breeding methods (Graudal et al., 2022). Therefore, significant investments in tree breeding offers the prospect of good economic returns, frequently greater than those from alternative forestry investments, along with good silviculture, management and product processing (Kanowski and Borralho, 2004). Future research should evaluate the response of trees, produced from different breeding strategies, to stressful conditions. It is also very likely that studies in biotechnology, phenomics, genomics, genome-wide associations and epigenetics will become essential methods in tree improvement programs, forest management and conservation practices since they allow for the identification of causal variants underlying phenotypes of interest, the evolutionary trajectory of populations to be studied, and help us understand the phenotypic plasticity and adaptive capacity of trees (Faino and Thomma, 2014; Plomion et al., 2016; Lind et al., 2018; Sow et al., 2018; Ebrahimi et al., 2020; Mahony et al., 2020; Rellstab, 2021; Singh et al., 2021). Compared with agricultural crops, timely domestication of trees is almost unachievable through traditional genetic improvement methods alone, due to the long breeding cycles and rotation times (Pearsall, 2008; Harfouche et al., 2012).

Conclusions

The genetic architecture of parameters such as biomass and gas exchange in white spruce have not been widely studied, particularly in families obtained from controlled crosses. We determined that height, above ground dry biomass, root length and root:shoot ratio traits are heritable, whereas diameter and all gas exchange traits studied appear to be under considerably less genetic control in the 2-year-old white spruce seedlings we studied. Based on the ABLUP values, certain families performed consistently better for both growth and physiology traits generated from polymix pollination and controlled crosses, however, the specific performance of individual families was unpredictable. Interestingly, within each genotype, some growth and physiological traits were similar across breeding strategies, indicating strong genetic control. We found substantial genetic variability in growth, gas exchange, hormone levels and gene expression traits between seedlings within families and breeding strategies when plotting the individual values. Furthermore, we observed trade-offs between growth, carbon allocation, photosynthesis, hormone levels and gene expression within seedlings. Some individuals showed greater height, more biomass in roots than shoots, higher stomatal conductance (stomata were more open), and higher water use efficiency, a series of traits desirable for selection. Controlled crosses from the fast growth group showed the best apical growth, but in some cases, open pollination from the

fast and slow growth groups showed the best root development and higher water use efficiency, based on iWUE and $\delta^{13}C$. We also found a pattern of greater accumulation of *PgGA3ox*, IAA, ABA, and PA in controlled cross seedlings with greater growth compared to seedlings from the other breeding strategies and found higher root length, root dry biomass, ZOG_t level, iPR level in open-pollinated seedlings with slower growth. ABA, PA, IAA and *PgGA3ox* actively regulated and promoted apical growth in the controlled cross seedlings and their levels potentially inhibited Zeatin synthesis. Cytokinins could be synergistically working with GA_3 balancing the root:shoot ratios, while promoting more root development in open-pollinated trees. In this study, we confirmed that hormone levels and expression of GA-related genes have a strong influence on many traits and differences were found between families coming from different breeding strategies. Our findings also show the existence of specific faster- and slower-growing families from open pollination, polymix pollination and controlled-crosses, with different developmental and physiological characteristics. Advancing tree domestication through the application of more advanced breeding approaches is supported by our results, encouraging the use of this phenotypic variation to advance tree improvement programs. For management purposes, it appears that selection based on controlled-crossed families would be reasonably efficient for most growth traits, which showed relatively high ABLUP values and moderate heritability values, however, some individuals from open-pollinated families in this first-generation cycle, also showed outstanding performance, particularly for root development. These OP derived individuals should be selected and used for future breeding as we continue to gain information on parental material through both phenotypic, genetic and physiological study.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

EG and BT contributed to the conception and design of the study. EG and BT organized the database. EG performed the statistical analysis. EG wrote the first draft of the manuscript. EG and BT wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. All authors contributed to the article and approved the submitted version.

Funding

Funding for this manuscript has been provided through the Industrial Research Chair in Tree Improvement held by BT (IRC461040-13) and supported by the Natural Sciences and Engineering Research Council (NSERC), Alberta-Pacific Forest Industries Inc., Alberta Newsprint Company Timber Ltd., Canadian Forest Products Ltd., Millar Western Forest Products Ltd., Huallen Seed Orchard Company Ltd., West Fraser Mills Ltd. (including: Alberta Plywood, Blue Ridge Lumber Inc., Hinton Wood Products (HWP), Sundre Forest Products Inc.), and Weyerhaeuser Company Ltd., (Pembina and Grande Prairie Timberlands). The funders were not involved in the study design, collection, analysis, interpretation of data, writing of this article, or the decision to submit it for publication.

Acknowledgments

We thank Alberta-Pacific Forest Industries Inc., Alberta Newsprint Company Timber Ltd., Canadian Forest Products Ltd., Millar Western Forest Products Ltd., Huallen Seed Orchard Company Ltd., West Fraser Mills Ltd. (including: Alberta Plywood, Blue Ridge Lumber Inc., Hinton Wood Products (HWP), Sundre Forest Products Inc.), and Weyerhaeuser Company Ltd., (Pembina and Grande Prairie Timberlands) for their financial support. We thank Kayla Frankiw for assisting in the plant collection in the greenhouse and Dr. Justine Karst for using her molecular biology lab. We acknowledge Dr. Irina Zaharia from the Aquatic and Crop Resource Development Research Center, National Research Council of Canada (NRCC), Saskatoon, Saskatchewan, Canada, for the quantification of hormones. Finally, we thank the reviewers for their valuable suggestions and comments to improve this article.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1052425/full#supplementary-material

References

Abbo, S., and Gopher, A. (2020). Plant domestication in the neolithic near East: The humans-plants liaison. *Quat. Sci. Rev.* 242, 106412. doi: 10.1016/j.quascirev.2020.106412

Abrams, S. R., Nelson, K., and Ambrose, S. J. (2003). Deuterated abscisic acid analogs for mass spectrometry and metabolism studies. *J. Label. Compd. Radiopharm.* 46, 273–283. doi: 10.1002/jlcr.670

Akhtar, S. S., Mekureyaw, M. F., and Pandey Roitsch, C. T. (2020). Role of cytokinins for interactions of plants with microbial pathogens and pest insects. *Front. Plant Sci.* 10, 1–12. doi: 10.3389/fpls.2019.01777

Arney, S. E., and Mitchell, D. L. (1969). The effect of abscisic acid on stem elongation and correlative inhibition. *New Phytol.* 68, 1001–1015. doi: 10.1111/j.1469-8137.1969.tb06500.x

Bigras, F. J. (2005). Photosynthetic response of white spruce families to drought stress. *New For.* 29, 135–148. doi: 10.1007/s11056-005-0245-9

Boltz, B. A., Bongarten, B. C., and Teskey, R. O. (1986). Seasonal patterns of net photosynthesis of loblolly pine from diverse origins. *Can. J. For. Res.* 16, 1063–1068. doi: 10.1139/x86-184

Bonner, F. T., Karrfalt, R. P., Landis, T. D., Lantz, C. W., and Mangold, R. D. (2009). *The woody plant seed manual* (Minnesota, USA: United States Department of Agriculture, Forest Service).

Bouffier, L., Raffin, A., and Kremer, A. (2008). Evolution of genetic variation for selected traits in successive breeding populations of maritime pine. *Heredity (Edinb).* 101, 156–165. doi: 10.1038/hdy.2008.41

Brendel, O., Le Thiec, D., Scotti-Saintagne, C., Bodénès, C., Kremer, A., Guehl, J. M., et al. (2008). Quantitative trait loci controlling water use efficiency and related traits in quercus robur l. *Tree Genet. Genomes* 4, 263–278. doi: 10.1007/s11295-007-0107-z

Callister, A. N., and Collins, S. L. (2008). Genetic parameter estimates in a clonally replicated progeny test of teak (Tectona grandis linn. f.). *Tree Genet. Genomes* 4, 237–245. doi: 10.1007/s11295-007-0104-2

Carles, S., Lamhamedi, M. S., Beaulieu, J., Stowe, D. C., and Margolis, H. A. (2012). Genetic parameters of morphological and physiological characteristics of containerized white spruce (Picea glauca [Moench.] Voss) seedlings. *Tree Genet. Genomes* 8, 39–51. doi: 10.1007/s11295-011-0418-y

Cline, M. G., and Harrington, C. A. (2007). Apical dominance and apical control in multiple flushing of temperate woody species. *Can. J. For. Res.* 37, 74–83. doi: 10.1139/x06-218

Dilworth, L. L., Riley, C. K., and Stennett, D. K. (2017). *Plant constituents: Carbohydrates, oils, resins, balsams, and plant hormones* (Cambridge, MA, USA: Elsevier Inc).

Ebrahimi, A., Lawson, S. S., McKenna, J. R., and Jacobs, D. F. (2020). Morpho-physiological and genomic evaluation of juglans species reveals regional maladaptation to cold stress. *Front. Plant Sci.* 11, 1–13. doi: 10.3389/fpls.2020.00229

Faino, L., and Thomma, B. P. H. J. (2014). Get your high-quality low-cost genome sequence. *Trends Plant Sci.* 19, 288–291. doi: 10.1016/j.tplants.2014.02.003

Fernández-Paz, J., Cortés, A. J., Hernández-Varela, C. A., Mejía-de-Tafur, M. S., Rodriguez-Medina, C., Baligar, V. C., et al. (2021). Rootstock-mediated genetic variance in cadmium uptake by juvenile cacao (Theobroma cacao l.) genotypes, and its effect on growth and physiology. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.777842

FGRMS (2016). *Alberta Forest genetic resource management and conservation standards volume 1: Stream 1 and stream 2.* (Edmonton, Alberta: Government of Alberta)

Flewelling, J. W. (2008). *Proposal for genetic gain estimation for three Alberta tree improvement programs.* (Edmonton, Alberta: Government of Alberta)

Galeano, E., Bousquet, J., and Thomas, B. R. (2021). SNP–based analysis reveals unexpected features of genetic diversity, parental contributions and pollen contamination in a white spruce breeding program. *Sci. Rep.* 11, 1–14. doi: 10.1038/s41598-021-84566-2

Galeano, E., and Thomas, B. R. (2020). Effect of elevated gibberellic acid application on growth and gene expression patterns in white spruce families from a tree improvement program in Alberta, Canada. *Tree Physiol.* 41, 472–490. doi: 10.1093/treephys/tpaa133

Gärtner, S. M., Lieffers, V. J., and Macdonald, S. E. (2011). Ecology and management of natural regeneration of white spruce in the boreal forest. *Environ. Rev.* 19, 461–478. doi: 10.1139/a11-017

Gepts, P. (2014). Domestication of plants. *Encycl. Agric. Food Syst.* 2, 474–486. doi: 10.1016/B978-0-444-52512-3.00231-X

Gilmour, A. R., Gogel, B. J., Cullis, B R., and Thompson, R. (2009). *ASReml user guide release 3.0. 1st edn* (Hemel Hempstead, England: VSN International Ltd).

Government of Alberta (2018). *Ex situ conservation plan for forest genetic resources. 1st ed.* (Edmonton, Alberta, Canada: Agriculture and Forestry, Government of Alberta).

Grattapaglia, D., Plomion, C., Kirst, M., and Sederoff, R. R. (2009). Genomics of growth traits in forest trees. *Curr. Opin. Plant Biol.* 12, 148–156. doi: 10.1016/j.pbi.2008.12.008

Graudal, L., Dawson, I. K., Hale, I., Powell, W., Hendre, P., and Jamnadass, R. (2022). 'Systems approach' plant breeding illustrated by trees. *Trends Plant Sci.* 27, 158–165. doi: 10.1016/j.tplants.2021.09.009

Harfouche, A., Meilan, R., Kirst, M., Morgante, M., Boerjan, W., Sabatti, M., et al. (2012). Accelerating the domestication of forest trees in a changing world. *Trends Plant Sci.* 17, 64–72. doi: 10.1016/j.tplants.2011.11.005

Hébert, F., Thiffault, N., and Munson, A. D. (2011). Field photosynthesis measurements on black spruce (picea mariana): Does needle age matter? *Commun. Soil Sci. Plant Anal.* 42, 2738–2750. doi: 10.1080/00103624.2011.622821

Heine, A. J., Walker, T. D., McKeand, S. E., Jett, J. B., and Isik, F. (2020). Pollination bag type has a significant impact on cone survival in mass production of controlled pollinated seeds in loblolly pine. *For. Sci.* 66, 589–599. doi: 10.1093/forsci/fxaa013

Isik, F., Holland, J., and Maltecca, C. (2017). *Genetic data analysis for plant and animal breeding. 1st edn* (Cham, Switzerland: Springer).

John, S. (2011). *Region G1 white spruce controlled parentage program plan. seed orchard G351.* (Edmonton, Alberta: The Huallen Seed Orchard Company Ltd.)

Kanowski, P. J., and Borralho, N. M. G. (2004). Economic returns from tree breeding. *Encycl. For. Sci.* Edition 1. (Cambridge, MA, USA: Academic Press), 1561–1568.

Kayal, W., Allen, C. C.G., Ju, C. J.T., Adams, E., King-Jones, S., Zaharia, L., et al. (2011). Molecular events of apical bud formation in white spruce, picea glauca. *Plant Cell Environ.* 34, 480–500. doi: 10.1111/j.1365-3040.2010.02257.x

Lamhamedi, M. S., and Bernier, P. Y. (1994). Ecophysiology and field performance of black spruce (Picea mariana): A review. *Ann. Des. Sci. For.* 51, 529–551. doi: 10.1051/forest:19940601

Lamhamedi, M. S., Chamberland, H., Bernier, P. Y., and Tremblay, F. M. (2000). Clonal variation in morphology, growth, physiology, anatomy and ultratructure of container-grown white spruce somatic plants. *Tree Physiol.* 20, 869–880. doi: 10.1093/treephys/20.13.869

Leakey, R. R. B. (2014). Agroforestry: Participatory domestication of trees. *Encycl. Agric. Food Syst.* 1, 253–269. doi: 10.1016/B978-0-444-52512-3.00025-5

Leakey, R. R. B. (2017a). "Domestication of forest trees: A process to secure the productivity and future diversity of tropical ecosystems," in *Multifunctional agriculture, 1st ed.* Ed. N. Maragioglio (Cambridge, MA, USA: Academic Press Inc), 105–110.

Leakey, R. R. B. (2017b). "Tree domestication," in *Multifunctional agriculture, 1st ed.* Ed. R. Leakey (Cambridge, MA, USA: Academic Press Inc), 103.

Li, P., Beaulieu, J., Corriveau, A., and Bousquet, J. (1993). Genetic variation in juvenile growth and phenology in a white spruce provenance-progeny test. *Silvae Genetica* 42 (1), 52–60.

Lind, B. M., Menon, M., Bolte, C. E., Faske, T. M., Eckert, A. J., et al. (2018). The genomics of local adaptation in trees: are we out of the woods yet? *Tree Genet. Genomes* 14, 1–29. doi: 10.1007/s11295-017-1224-y

Lorrai, R., Boccaccini, A., Ruta, V., Possenti, M., Costantino, P., and Vittorioso, P. (2018). Abscisic acid inhibits hypocotyl elongation acting on gibberellins, DELLA proteins and auxin. *AoB Plants* 10, 1–10. doi: 10.1093/aobpla/ply061

Lulsdorf, M. M., Yuan, H. Y., Slater, S. M.H., Vandenberg, A., Han, X., Zaharia, L. I., et al. (2013). Endogenous hormone profiles during early seed development of c. arietinum and c. anatolicum. *Plant Growth Regul.* 71, 191–198. doi: 10.1007/s10725-013-9819-2

Mahony, C. R., MacLachlan, I. R., Lind, B. M., Yoder, J. B., Wang, T., Aitken, S. N., et al. (2020). Evaluating genomic data for management of local adaptation in a changing climate: A lodgepole pine case study. *Evol. Appl.* 13, 116–131. doi: 10.1111/eva.12871

Matsumoto-Kitano, M., Kusumoto, T., Tarkowski, P., Kinoshita-Tsujimura, K., Václavíková, K., Miyawaki, K., et al. (2008). Cytokinins are central regulators of cambial activity. *Proc. Natl. Acad. Sci. U. S. A.* 105, 20027–20031. doi: 10.1073/pnas.0805619105

Moran, E., Lauder, J., Musser, C., Stathos, A., and Shu, M. (2017). The genetics of drought tolerance in conifers. *New Phytol.* 216, 1034–1048. doi: 10.1111/nph.14774

Park, Y. S., Pond, S. E., and Bonga, J. M. (1993). Initiation of somatic embryogenesis in white spruce (Picea glauca): genetic control, culture treatment effects, and implications for tree breeding. *Theor. Appl. Genet.* 86, 427–436. doi: 10.1007/BF00838557

Park, Y. S., Pond, S. E., and Bonga, J. M. (1994). Somatic embryogenesis in white spruce (Picea glauca): genetic control in somatic embryos exposed to storage, maturation treatments, germination, and cryopreservation. *Theor. Appl. Genet.* 89, 742–750. doi: 10.1007/BF00223714

Park, E. J., Lee, W. Y., Kurepin, L. V., Zhang, R., Janzen, L., and Pharis, R. P. (2014). Plant hormone-assisted early family selection in pinus densiflora *via* a retrospective approach. *Tree Physiol.* 35, 86–94. doi: 10.1093/treephys/tpu102

Paterson, J. M., and Faylel, D. C. F. (1984). Early prediction of plantation performance for red pine. *For. Chron.* 60(6), 340–344. doi: 10.5558/tfc60340-6

Pearsall, D. M. (2008). "Plant domestication," in *Encyclopedia of archaeology*, *1st ed*. Ed. D. M. Pearsall (Cambridge, MA, USA: Academic Press Inc), 1822–1842.

Persson, B. C., Esberg, B., Ólafsson, Ó., and Björk, G. R. (1994). Synthesis and function of isopentenyl adenosine derivatives in tRNA. *Biochimie* 76, 1152–1160. doi: 10.1016/0300-9084(94)90044-2

Plomion, C., et al. (2016). Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Ann. For. Sci.* 73, 77–103. doi: 10.1007/s13595-015-0488-3

Prud'Homme, G. O., Lamhamedi, M. S., Benomar, L., Rainville, A., Deblois, J., Bousquet, J., et al. (2018). Ecophysiology and growth of white spruce seedlings from various seed sources along a climatic gradient support the need for assisted migration. *Front. Plant Sci.* 8, 1–17. doi: 10.3389/fpls.2017.02214

Quesada, T., Parisi, L. M., Huber, D. A., Gezan, S. A., Martin, T. A., and Davis, J. M. (2017). Genetic control of growth and shoot phenology in juvenile loblolly pine (Pinus taeda l.) clonal trials. *Tree Genet. Genomes* 13, 1–15. doi: 10.1007/s11295-017-1143-y

Raj, D., Govindaraju, D., and Orians, C. (2006). Genetic variation among pitch pine *Pinus rigida* families from walden woods: heritability and path analysis of developmental variation of phenotypic traits. *Rhodora* 108, 356–369. doi: 10.3119/0035-4902(2006)108[356:GVAPPP]2.0.CO;2

Rao, X., Huang, X., Zhou, Z., and Lin, X. (2013). An improvement of the 2^(-delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. *Biostat. Bioinforma. Biomath.* 3, 71–85.

Rellstab, C. (2021). Genomics helps to predict maladaptation to climate change. *Nat. Clim. Change* 11, 85–86. doi: 10.1038/s41558-020-00964-w

Rweyongeza, D. M., Yeh, F. C., and Dhir, N. K. (2005). Heritability and correlations for biomass production and allocation in white spruce seedlings. *Silvae Genet.* 54, 228–235. doi: 10.1515/sg-2005-0033

Rweyongeza, D. M., Yeh, F. C., and Dhir, N. K. (2010). Genetic parameters for bud flushing and growth characteristics of white spruce seedlings. *Silvae Genet.* 59, 151–158. doi: 10.1515/sg-2010-0018

Scotti, I., Calvo-Vialettes, L., Scotti-Saintagne, C., Citterio, M., Degen, B., and Bonal, D. (2010). Genetic variation for growth, morphological, and physiological traits in a wild population of the Neotropical shade-tolerant rainforest tree sextonia rubra (Mez) van der werff (Lauraceae). *Tree Genet. Genomes* 6, 319–329. doi: 10.1007/s11295-009-0251-8

Scotti-Saintagne, C., Bodénès, C., Barreneche, T., Bertocchi, E., Plomion, C., and Kremer, A. (2004). Detection of quantitative trait loci controlling bud burst and height growth in quercus robur l. *Theor. Appl. Genet.* 109, 1648–1659. doi: 10.1007/s00122-004-1789-3

Shu, K., Zhou, W., Chen, F., Luo, X., and Yang, W. (2018). Abscisic acid and gibberellins antagonistically mediate plant development and abiotic stress responses. *Front. Plant Sci.* 9, 1–8. doi: 10.3389/fpls.2018.00416

Shunmugam, A. S. K., Kannan, U., Jiang, Y., Daba, K. A., and Gorim, L. Y. (2018). Physiology based approaches for breeding of next-generation food legumes. *Plants* 7, 1–21. doi: 10.3390/plants7030072

Singh, D. P., Singh, A. K., and Singh, A. (2021). "Plant breeding: past, present, and future perspectives," in *Plant breeding and cultivar development*, *1st ed.* Ed. N. Maragioglio (Cambridge, MA, USA: Academic Press Inc), 1–24.

Six, D. L., Trowbridge, A., Howe, M., Perkins, D., Berglund, E., Brown, P., et al. (2021). Growth, chemistry, and genetic profiles of whitebark pine forests affected by climate-driven mountain pine beetle outbreaks. *Front. For. Glob. Change* 4, 1–22. doi: 10.3389/ffgc.2021.671510

Sklar, D. A. (2012). *Tree improvement structure in Alberta- recommendation*. (Edmonton, Alberta: Government of Alberta)

Sotelo Montes, C., Beaulieu, J., and Hernández, R. E. (2007). Genetic variation in wood shrinkage and its correlations with tree growth and wood density of calycophyllum spruceanum at an early age in the Peruvian Amazon. *Can. J. For. Res.* 37, 966–976. doi: 10.1139/X06-288

Sow, M. D., et al. (2018). Epigenetics in forest trees: State of the art and potential implications for breeding and management in a context of climate change. *Adv. Bot. Res.* 88, 387–453. doi: 10.1016/bs.abr.2018.09.003

Stoehr, M. U., L'Hirondelle, S. J., Binder, W. D., and Webber, J. E. (1998). Parental environment aftereffects on germination, growth, and adaptive traits in selected white spruce families. *Can. J. For. Res.* 28, 418–426. doi: 10.1139/x98-012

Theodorou, C., and Bowen, G. D. (1993). Root morphology, growth and uptake of phosphorus and nitrogen of pinus radiata families in different soils. *For. Ecol. Manage.* 56, 43–56. doi: 10.1016/0378-1127(93)90102-S

Tripiana, V., Bourgeois, M., Verhaegen, D., Vigneron, P., and Bouvet, J. M. (2007). Combining microsatellites, growth, and adaptive traits for managing *in situ* genetic resources of eucalyptus urophylla. *Can. J. For. Res.* 37, 773–785. doi: 10.1139/X06-260

Verta, J. P., Landry, C. R., and MacKay, J. J. (2013). Are long-lived trees poised for evolutionary change? single locus effects in the evolution of gene expression networks in spruce. *Mol. Ecol.* 22, 2369–2379. doi: 10.1111/mec.12189

Wahid, N., Lamhamedi, M. S., Rainville, A., Beaulieu, J., and Margolis, H. A. (2013). Genetic control and nursery-plantation genotypic correlations for growth characteristics of white spruce somatic clones. *J. Sustain. For.* 32, 576–593. doi: 10.1080/10549811.2013.791231

Ward, S. E., Wightman, K. E., and Santiago, B. R. (2008). Early results from genetic trials on the growth of Spanish cedar and its susceptibility to the shoot borer moth in the Yucatan peninsula, Mexico. *For. Ecol. Manage.* 255, 356–364. doi: 10.1016/j.foreco.2007.09.057

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. *1st ed.* (New York, USA: Springer-Verlag).

Zaharia, L. I., Galka, M. M., Ambrose, S. J., and Abrams, S. R. (2005). Preparation of deuterated abscisic acid metabolites for use in mass spectrometry and feeding studies. *J. Label. Compd. Radiopharm.* 48, 435–445. doi: 10.1002/jlcr.939

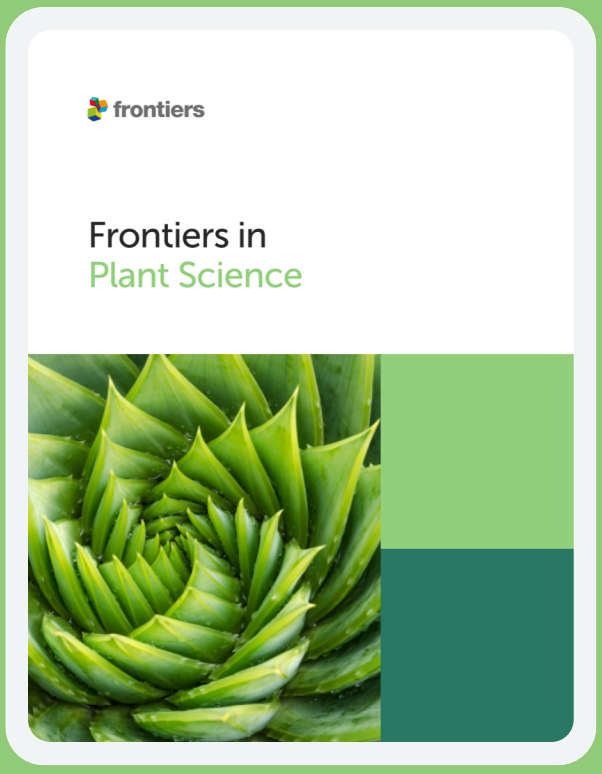
Frontiers in Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact