Molecular Biology and Evolution, 2025, **42**, 1–21
https://doi.org/10.1093/molbev/msaf088
Advance access publication 16 April 2025
Discoveries

MBE OXFORD

Pan-Genome Analysis Reveals Local Adaptation to Climate Driven by Introgression in Oak Species

Yi-Ye Liang [1,2,3,†], Hui Liu [1,2,3,4,†], Qiong-Qiong Lin [1,2,3,†], Yong Shi [1,2,3], Biao-Feng Zhou [1,2,3], Jing-Shu Wang [1,2,3], Xue-Yan Chen [1,2,3], Zhao Shen [1,2,3], Liang-Jing Qiao [1,2,3], Jing-Wei Niu [1,2,3], Shao-Jun Ling [1,2,3], Wen-Ji Luo [1,2,3], Wei Zhao [4], Jian-Feng Liu [5], Yuan-Wen Kuang [2,3], Pär K. Ingvarsson [6], Ya-Long Guo [7], Baosheng Wang [1,2,3,*]

[1]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China
[2]Guangdong Provincial Key Laboratory of Applied Botany and Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China
[3]South China National Botanical Garden, Guangzhou, China
[4]Department of Ecology and Environmental Science, UPSC, Umeå University, Umeå, Sweden
[5]Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China
[6]Department of Plant Biology, Linnean Center for Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden
[7]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: baosheng.wang@scbg.ac.cn.
Associate editor: Tanja Pyhäjärvi

Abstract

The genetic base of local adaptation has been extensively studied in natural populations. However, a comprehensive genome-wide perspective on the contribution of structural variants (SVs) and adaptive introgression to local adaptation remains limited. In this study, we performed de novo assembly and annotation of 22 representative accessions of *Quercus variabilis*, identifying a total of 543,372 SVs. These SVs play crucial roles in shaping genomic structure and influencing gene expression. By analyzing range-wide genomic data, we identified both SNPs and SVs associated with local adaptation in *Q. variabilis* and *Quercus acutissima*. Notably, SV-outliers exhibit selection signals that did not overlap with SNP-outliers, indicating that SNP-based analyses may not detect the same candidate genes associated with SV-outliers. Remarkably, 29%–37% of candidate SNPs were located in a 250 kb region on chromosome 9, referred to as *Chr9-ERF*. This region contains 8 duplicated ethylene-responsive factor (*ERF*) genes, which may have contributed to local adaptation of *Q. variabilis* and *Q. acutissima*. We also found that a considerable number of candidate SNPs were shared between *Q. variabilis* and *Q. acutissima* in the *Chr9-ERF* region, suggesting a pattern of repeated selection. We further demonstrated that advantageous variants in this region were introgressed from western populations of *Q. acutissima* into *Q. variabilis*, providing compelling evidence that introgression facilitates local adaptation. This study offers a valuable genomic resource for future studies on oak species and highlights the importance of pan-genome analysis in understating mechanism driving adaptation and evolution.

Keywords: pan-genome, local adaptation, adaptive introgression, structural variants, oak

Introduction

Local adaptation is a crucial driving force that promotes genetic differentiation and speciation (Savolainen et al. 2013; Wadgymar et al. 2022). Genomic data have been extensively employed to investigate the genetic basis of local adaptation in natural populations (Savolainen et al. 2013; Sork 2018). However, these studies often rely on mapping short reads to a single reference genome, which fails to capture the full spectrum of genetic variation within a species, particularly for large structural variants (SVs) (Bock et al. 2023; Song et al. 2023; Feng et al. 2024). It has been shown that SVs significantly influence the expression patterns of genes involved in environmental stress responses (Hämälä et al. 2021; Kang et al. 2023; Yan et al. 2023; Shi et al. 2024), and contribute to phenotypic variation (Liu et al. 2020; Qin et al. 2021; Chen et al. 2023a, 2023b). Although various methods have been developed to call SVs from short-read sequencing data, they often suffer from low confidence (Cameron et al. 2019; Kosugi et al. 2019). This limitation is particularly pronounced in forest trees with complex genomic structures and high levels of genetic diversity (Feng et al. 2024). With the advent of long-read sequencing technology, graph-based pan-genomes have been developed to identify SVs based on multiple high-quality assemblies and integrate genetic variation of a species in a more intuitive way (Garrison et al. 2018; Della Coletta et al. 2021). Using graph-based pan-genome analyses, recent studies have detected many SVs associated with important adaptive and agronomic traits in model species and crops, such as *Arabidopsis thaliana* (Kang et al. 2023), rice (Qin et al. 2021), soybean (Liu et al. 2020), and tomato (Zhou et al.

Received: November 13, 2024. **Revised:** April 1, 2025. **Accepted:** April 2, 2025

2022c). Until now, pan-genome data remain limited in non-model, long-lived forest tree species (Feng et al. 2024), and a comprehensive genome-wide view regarding the contribution of SVs to genome evolution and adaptation is generally lacking in these species.

The sources of genetic variation underlying local adaptation may originate from de novo mutations, ancestral polymorphisms, or gene flow (Flood and Hancock 2017). For long-lived forest trees, the time required for new mutations to arise can be extensive; therefore, ancestral polymorphisms and gene flow act as crucial drivers of local adaptation, enabling rapid adjustments of species to changing environments (Savolainen et al. 2007; Flood and Hancock 2017; Feng et al. 2024). Hybridization is frequently observed between closely related species with incomplete reproductive isolation, either in areas of sympatric distributions or in secondary contact zones following a period of geographic isolation (Canestrelli et al. 2016; Cannon and Petit 2019). Although most of the foreign genetic variants introduced through hybridization are neutral or deleterious and are rapidly purged from the recipient species' genome, advantageous variants may be preserved through selection, resulting in adaptive introgression (Moran et al. 2021). While adaptive introgression has been well documented in plant species (Savolainen et al. 2007; Kremer and Hipp 2020), few empirical studies have explored how beneficial alleles involved in this process are selected and maintained in both donor and recipient species.

Oaks (genus *Quercus*, Fagaceae) comprise more than 450 species and dominate the temperate and subtropical forests of the Northern Hemisphere (Denk et al. 2017), representing an "evolutionary success story" (Kremer and Hipp 2020). Many oak species are distributed across large environment gradients within their geographic ranges, and exhibit high levels of genetic and phenotypic variation (Cavender-Bares 2019; Kremer and Hipp 2020; Liang et al. 2021). Common garden experiments and population genetics analyses have revealed local adaptation in oaks (Rellstab et al. 2016; Martins et al. 2018; Leroy et al. 2020a; Gao et al. 2021; Zhou et al. 2022a). Given the high frequency of hybridization among oak species (Zeng et al. 2011; Ortego et al. 2017, 2018; Kim et al. 2018; Burge et al. 2019; Kremer and Hipp 2020; Leroy et al. 2020b; Zhou et al. 2022b), adaptive introgression may have effectively facilitated the spread of beneficial alleles across species, enabling them to adapt to new environments. Evidence of adaptive introgression has been observed in oak species. For example, O'Donnell et al. (2021) demonstrated that asymmetrical introgression from an endemic California oak to a nonendemic species occurred during historical climate changes, suggesting the adaptive introgression of climate-associated genes. Similarly, a common garden study documented the introgression of loci associated with coastal environment adaptability from *Q. dentata* into *Q. mongolica* (Nagamitsu et al. 2024). A recent study using whole-genome sequencing data revealed that the introgression from *Q. robur* likely contributes to the adaptive divergence of *Q. petraea* populations (Leroy et al. 2020a). The well documented signatures of adaptive introgression, combined with increasingly accessible genomic resources (e.g. Plomion et al. 2018; Fu et al. 2022; Sork et al. 2022; Wang et al. 2023b), make oak species an interesting system for investigating how introgression contributes to adaptive evolution.

In this study, we focused on 2 sympatric oak species, *Quercus acutissima* and *Quercus variabilis*, both belonging to the section *Cerris* (Denk et al. 2017). These species diverged during the late Neogene (Liang et al. 2021) and are distributed across a broad environmental gradient, ranging from temperate to subtropical regions of East Asia (Huang et al. 1999). Temperature and precipitation act as strong environment selective pressures across their distribution ranges. Populations from eastern and northern China occupy colder and more arid niches, while populations in southwestern China inhabit more humid niches with higher temperature and precipitation (Zhao 2018; Yuan et al. 2023). Population genomic analyses revealed extensive gene flow between these 2 species and suggested that frequent hybridization has facilitated adaptation to local environments (Fu et al. 2022). However, the genetic basis of local adaptation and the extent of adaptive introgression remains poorly understood. Moreover, previous studies relying on a single reference genome and utilizing short-read resequencing data were not able to reveal how SVs contribute to adaptive evolution in these species. To address these issues, we first constructed a graph-based pan-genome based on de novo assemblies of 22 *Q. variabilis* individuals to capture the majority of genetic diversity within this species. Next, we performed range-wide whole-genome resequencing of both *Q. acutissima* and *Q. variabilis* and integrated both SNPs and SVs data to explore the genetic basis of local adaptation. Finally, we focused on a genomic region of ~250 kb with strong signals of selection to investigate how adaptive introgression driving local adaption.

Results

Chromosome-Level Assembly and Annotation of 22 *Q. variabilis* Individuals

We sampled 22 individuals of *Q. variabilis* across the species' distribution range for genomic analyses. One genome (CLM) has previously published (Wang et al. 2023b), while the others were de novo assembled for this study (Fig. 1a and supplementary fig. S1 and table S1, Supplementary Material online). A telomere-to-telomere (T2T) haplotype-resolved genome assembly was constructed for the sample "SDTS" (Material and Methods), and a gap-free haplotype (SDTS-hap1) of this sample was used as a reference for subsequent analyses (supplementary fig. S2, Supplementary Material online). For each of remaining 20 individuals, 21–39 gigabases (Gb) of HiFi long reads (~26–49× coverage) were assembled into contigs and then anchored into the 12 pseudo-chromosomes of SDTS-hap1 using RagTag (Alonge et al. 2022). All assemblies exhibit high completeness and accuracy, with a contig N50 of 11.1–64.8 megabases (Mb), complete Benchmarking Universal Single-Copy Orthologues (BUSCO) scores of 97.5%–98.3%, *K*-mer based estimates of assembly consensus qualities of 58.9–65.3, and a long terminal repeats (LTRs) assembly index of 21.2–23.0 (supplementary table S1, Supplementary Material online). The sizes of the 22 *Q. variabilis* assemblies ranged from 777.3 to 838.1 Mb, which likely can be attributed to differing degrees of expansion and retention of LTRs (supplementary fig. S3 and table S2, Supplementary Material online). The *Q. variabilis* assemblies exhibit a high level of synteny (synteny relationship index, SRI = 0.77–0.87; Fig. 1b and d), close to reports in other plants, such as rice (Qin et al. 2021) and tomato (Zhou et al. 2022c). However, the degree of synteny varies across the genome and was positively correlated with the density of transposable elements (TE) (Spearman's

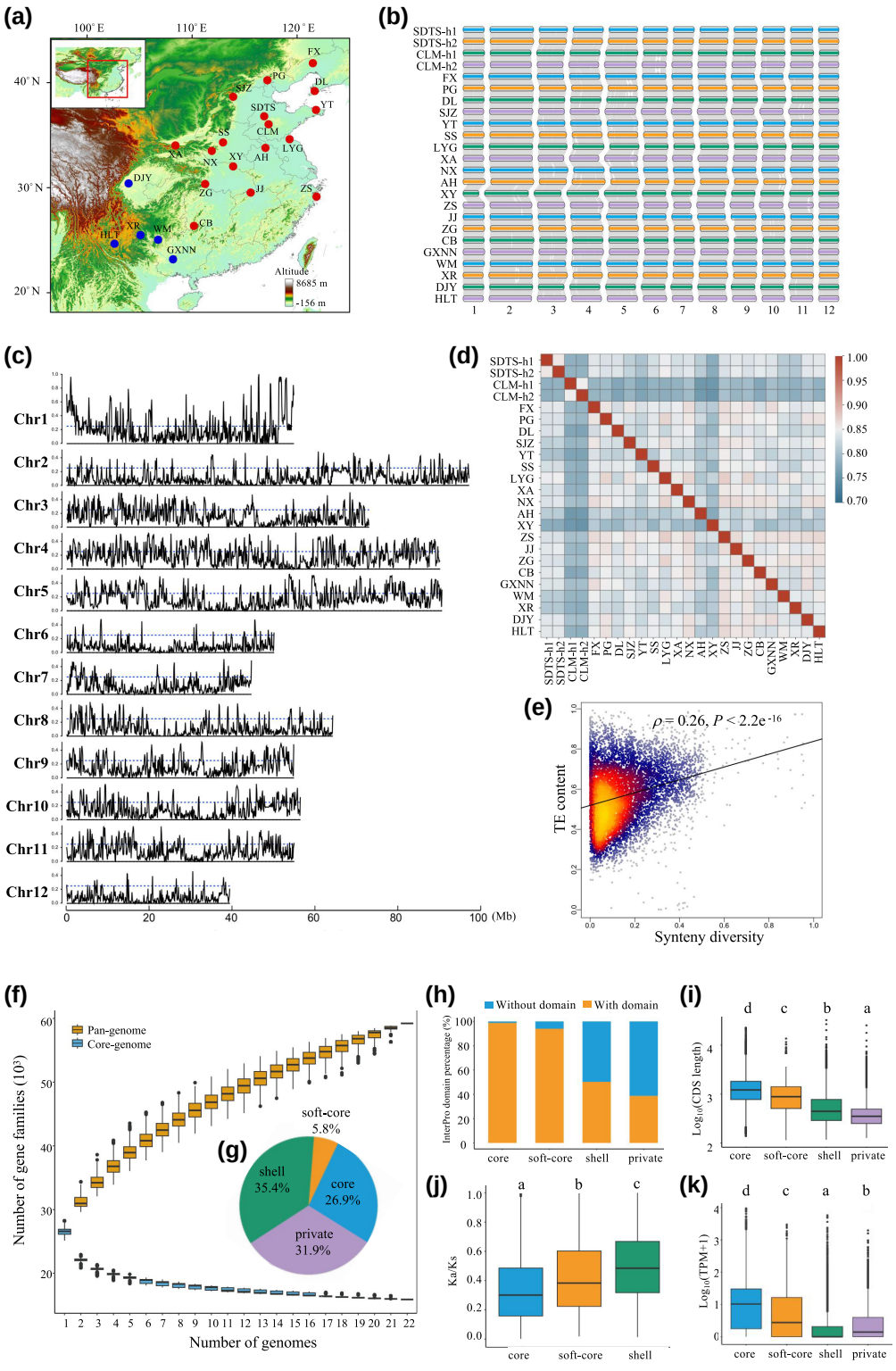

Fig. 1. Genome assembly and pan-genome analyses of 22 *Q. variabilis* accessions. a) Geographic distribution of the 22 *Q. variabilis* accessions sampled for genome assembly. Each dot represents an individual belong to East (red) and West (blue) genetic groups identified in supplementary fig. S1, Supplementary Material online. b) Collinearity among *Q. variabilis* genome assemblies. The chromosomes are represented by colorful boxes, and collinear regions among these genomes are shown as gray blocks. Two individuals SDTS and CLM were assembled to haplotype level. c) Variation in synteny diversity across genome estimated in 100 kb nonoverlapping windows. The dashed blue line indicate threshold for synteny diversity of 0.25. d) Heatmap showing the SRI between each pair of *Q. variabilis* genomes. e) Positive correlation between synteny diversity and TE content across 100 kb nonoverlapping windows. In the plot, the yellow-to-red-to-blue color gradient represents decreasing data density. Spearman's correlation coefficient (ρ) and *P*-value are shown. f) Changes in the number of gene families in the pan-genome and core-genome with increasing sample size. g) Pie chart illustrating the proportion of core, soft-core, shell, and private gene clusters. h) Percentage of genes with InterPro domains in various gene clusters. i–k) Comparisons of the length of coding region (CDS), K_a/K_s ratios, and expression levels among genes in different categories. In each box plot, the center line denotes the median, while the lower and upper bounds of the boxes represent the first and third quartiles, respectively. The whiskers extend to 1.5 times the interquartile range, and outliers are not shown in the plot. Different letters (a–d) above each box indicate significant differences among gene clusters, as determined by Wilcoxon–Mann–Whitney *U*-tests with Bonferroni correction ($P < 0.05$).

$\rho = 0.259$, $P < 2.2e^{-16}$; Fig. 1c and e). It is worth noting that the level of synteny could be overestimated due to the use of RagTag to scaffold contigs according to the same reference genome. Future studies utilizing ultra-long reads and HiC sequencing data would help to assemble each genome into chromosome-level assemblies without relying on a reference genome. More accurate synteny estimates could then be obtained from these nonreference-guided genome assemblies.

To construct the gene repertoire of *Q. variabilis*, we annotated 33,210–37,752 protein-coding genes in each of the 22 assemblies (supplementary table S3, Supplementary Material online), and clustered them into 59,238 pan-gene clusters, which included 15,909 (26.9%) core (present in all 22 individuals), 3,411 (5.8%) soft-core (present in 20-21 individuals), 20,989 (35.4%) shell (present in 2-19 individuals), and 18,929 (31.9%) private (present in only 1 genome) gene categories (Fig. 1f and g). As the sample size increased, the number of core gene families appeared to plateau, although a slight decline was observed (Fig. 1f). In contrast, the pan-gene set did not show saturation, even when all genomes were analyzed (Fig. 1f). These results suggested that our sample size may have captured the core gene set of *Q. variabilis*, but it was not enough to represent the entire array of diverse gene families within this species. We also noticed that core genes tend to exhibit longer coding sequences (CDS), a higher percentage of genes with InterPro domains, a lower ratio of nonsynonymous to synonymous (*Ka*/*Ks*) substitutions, and a higher level of expression than shell and private genes ($P < 0.05$, Wilcoxon–Mann–Whitney *U*-test with Bonferroni correction; Fig. 1h–k), suggesting that core genes are functionally conserved. This is congruent with results of gene enrichment analyses, which indicate that core genes were enriched for essential regulatory and metabolic process, while shell and private genes were enriched for biological processes such as defense response to biotic and abiotic factors (supplementary fig. S4, Supplementary Material online).

Structural Variation Identification and Graph-Based Pan-Genome Construction

To construct a graph-based pan-genome for *Q. variabilis*, we employed 6 complementary approaches to identify SVs among the 22 *Q. variabilis* genomes (Fig. 2a and supplementary fig. S5, Supplementary Material online). After merging redundant SVs, we retained a total of 543,372 nonredundant SVs, covering 625 Mb of sequences absent from the reference genome. This SV set including 294,985 insertions (INS), 246,773 deletions (DEL), and 1,614 inversions (INV) (Fig. 2b). Most of these SVs were short and present in only a few individuals, displaying a more left-skewed allele frequency spectrum compared to synonymous and nonsynonymous SNPs (Fig. 2c), suggesting that these SVs are either newly emerged or more deleterious to fitness than SNPs (Hämälä et al. 2021). The distribution of SVs was uneven across the genome, and the number of SVs was positively correlated with the density of TEs (Spearman's $\rho = 0.59$, $P < 2.2e^{-16}$; Fig. 2d and e), similar to findings in other plants such as *A. thaliana* (Kang et al. 2023), *Camellia sinensis* (Chen et al. 2023b), and rice (Qin et al. 2021). These results suggest a significant role of TEs in driving the formation and evolution of SVs in plant genome.

To assess the impact of SVs on gene expression, we categorized genes into 26,265 SV-genes (genic bodies overlapping with SVs) and 7,601 non-SV-genes (gene bodies not overlapping with SVs). The expression levels of SV-genes were not significantly different from those of non-SV-genes ($W = 98{,}685{,}380$, $P > 0.05$, Wilcoxon–Mann–Whitney *U*-test; Fig. 2f). However, genes with SVs located in exons (SV-exon-genes) exhibited much lower expression levels compared to non-SV-genes ($W = 76{,}966{,}553$, $P = 3.68e^{-7}$, Wilcoxon–Mann–Whitney *U*-test; Fig. 2f). Furthermore, SV-exon-genes had a higher proportion of low-expressed genes (fragments per kilobase of transcript per million, FPKM < 5) and fewer highly expressed genes (FPKM ≥ 15) compared to non-SV-genes (Fig. 2g). These results suggest that SVs may influence gene expression by altering the coding regions in *Q. variabilis*. By combining all nonredundant SVs, indels, and SNPs (Materials and Methods), we constructed a pan-genome graph, which serves as a valuable reference for the identification of SVs in *Q. variabilis* and related species.

Parallel East–West Divergence Between *Q. acutissima* and *Q. variabilis*

To characterize genomic variation across the species' range, we analyzed whole-genome re-sequencing data for *Q. variabilis* (143 individuals) and *Q. acutissima* (93 individuals), as well as a related species *Q. chenii* (21 individuals) (Fig. 3a and b and supplementary table S4, Supplementary Material online). Using the graph-based pan-genome of *Q. variabilis* as a reference, we detected 21.7 million high-quality SNPs and 259,877 SVs that were used in subsequent analyses. Population structure analyses based on SNPs and SVs, conducted using NGSadmix (Skotte et al. 2013), principal component analysis (PCA) and neighbor-joining (NJ) clustering revealed clear divergence among the 3 species, as well as east–west population differentiation within *Q. acutissima* and *Q. variabilis* (Fig. 3a and b and supplementary figs. S6–S8, Supplementary Material online). In each of these 2 species, populations from eastern China and southwestern China clustered into an "East group" and "West group," respectively (Fig. 3a and b and supplementary figs. S6–S8, Supplementary Material online). The estimated genetic differentiation between the East and West groups was generally low across the genome ($F_{ST} = 0.024$–0.031, $d_{XY} = 7.53 \times 10^{-3}$–$7.85 \times 10^{-3}$), and both groups displayed similarly high levels of genetic variation ($\pi = 7.533 \times 10^{-3}$–$7.751 \times 10^{-3}$; Fig. 3c and supplementary table S5, Supplementary Material online). These results suggest that many polymorphisms are shared between the East and West groups in each species, most likely due to incomplete lineage sorting and/or extensive gene flow.

To infer the divergence history of these species, we conducted coalescent simulations using fastsimcoal2 v2.6.0 (Excoffier et al. 2013). The best-fitting model suggested that the 3 species diverged sequentially 16.15–13.84 million years ago (Ma), followed by divergence within *Q. acutissima* and *Q. variabilis* around 5.66–5.50 Ma (Fig. 3d and supplementary table S6, Supplementary Material online). This timing coincides with major tectonic events and paleoclimate oscillations during the Miocene–Pliocene in East Asia (Zhisheng et al. 2001 ; Li et al. 2018). Intraspecific gene flow between genetic groups was generally higher than interspecific gene flow, except that relatively high gene flow (per generation migration rate = 1.34×10^{-5}–1.60×10^{-5}) was observed between the 2 West groups of *Q. variabilis* and *Q. acutissima* (Fig. 3d and supplementary table S6, Supplementary Material online). Overall, we detected parallel

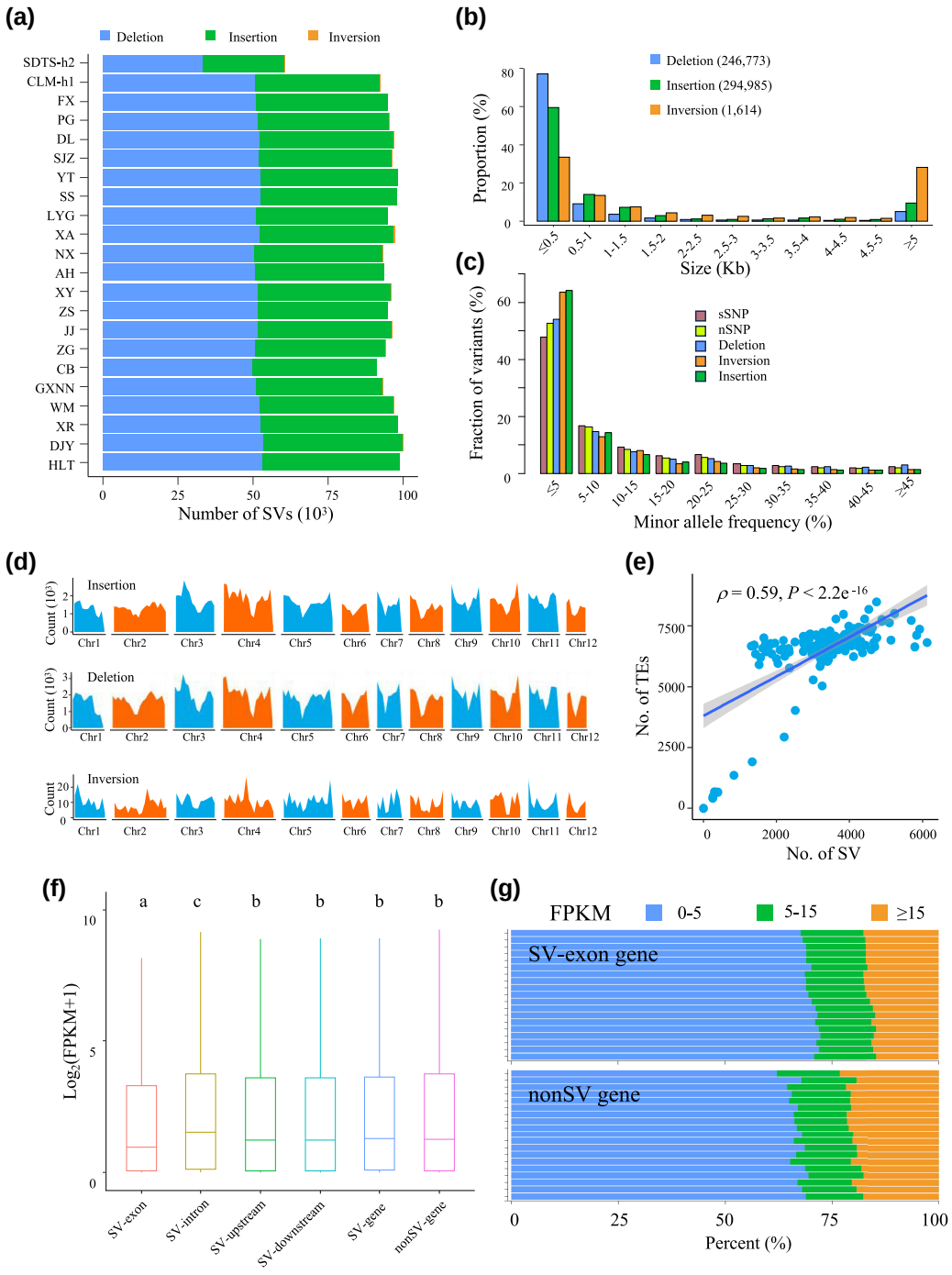

Fig. 2. Characterization of SVs in the 22 *Q. variabilis* genomes. a) Number of SVs (deletion, insertion, and inversion) detected in each individual using SDTS-hap1 as the reference genome. There are fewer SVs in SDTS-h2 compared to the other genomes because it represents a comparison between 2 haplotypes of the same individual (SDTS). b) The size distributions of 3 types of SVs. c) Allele frequency spectrum of SVs, synonymous (sSNP), and nonsynonymous nucleotide variants (nSNP). d) Genome-wide distribution of SVs. The number of insertions (top panel), deletions (middle panel), and inversions (bottom panel) were counted in 5 Mb nonoverlapping windows. e) Spearman's correlation between the number of SVs and the number of TEs counted in 5 Mb nonoverlapping windows. f) Comparisons of expression levels between non-SV-genes and SV-genes in leave tissue. The SV-genes were further categorized into 4 groups based on the position of the SV relative to genic regions. In each box, the center line denotes the median, while the lower and upper bounds of the boxes represent the first and third quartiles, respectively. The whiskers extend to 1.5 times the interquartile range, and outliers are not shown in the plot. Different letters (a–c) above each box indicate significant differences of expression levels among gene categories, as determined by Wilcoxon–Mann–Whitney *U*–tests with Bonferroni correction ($P < 0.05$). g) Comparisons of gene expression between SV-exon-genes (SV located in exon) and non-SV-genes in leave tissue across 19 *Q. variabilis* individuals. Each row represents an accession, with different colors indicating the percentages of genes exhibiting varying expression levels. FPKM, fragments per kilobase of transcript per million mapped reads.

east–west divergence initiated almost simultaneously within *Q. variabilis* and *Q. acutissima*, suggesting that common evolutionary forces may have driven genomic differentiation.

Genomic Basis of Local Adaptation

To investigate the genetic basis of local adaptation in the *Q. acutissima* and *Q. variabilis*, we applied 3 complementary

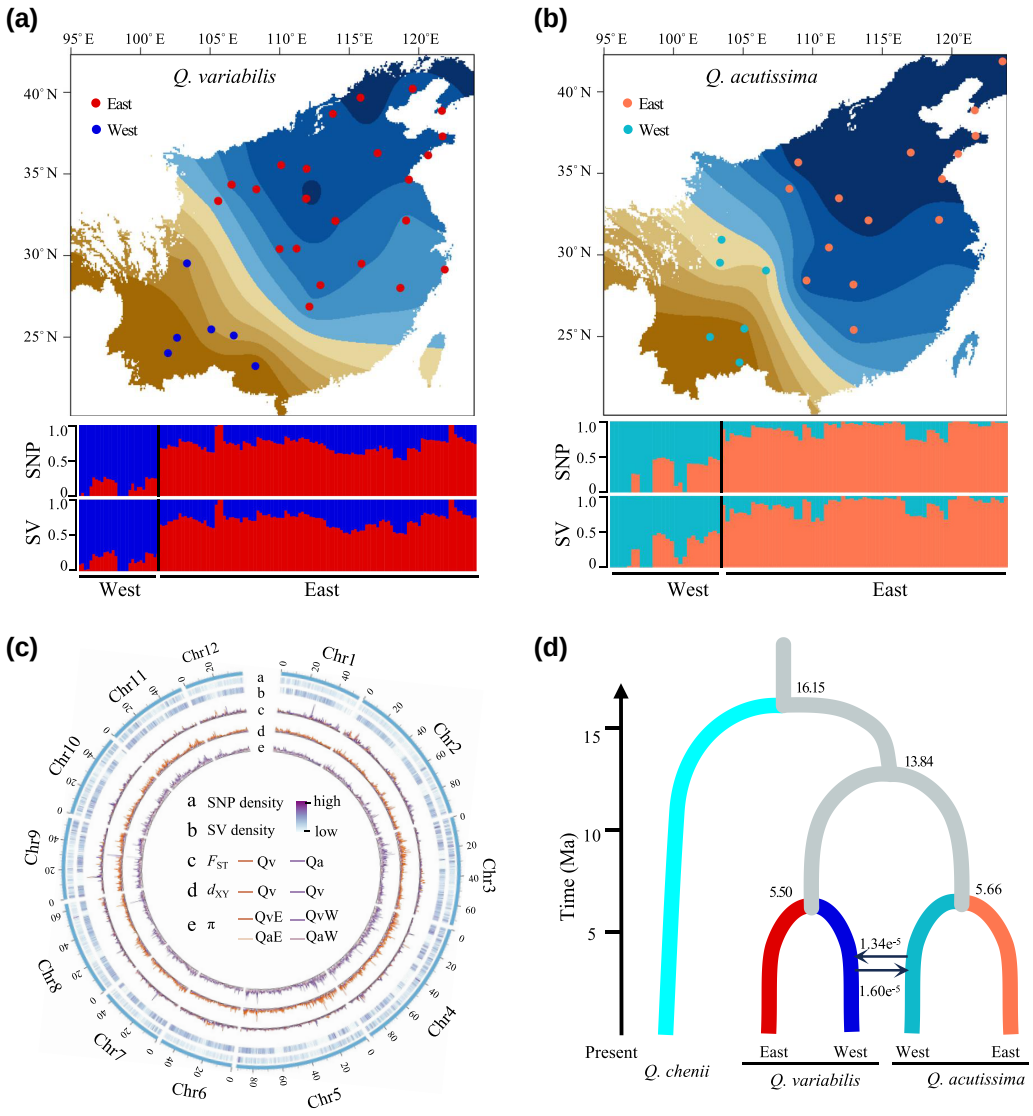

Fig. 3. Genetic divergence and demographic history of *Q. variabilis* and *Q. acutissima*. a, b) Upper panels: Spatial interpolation of 2 genetic groups inferred by NGSadmix in *Q. variabilis* (a) and *Q. acutissima* (b). Each dot represents 1 population that was assigned to an East or West genetic group based on population structure analyses. Colors gradients represent different genetic groups. Lower panels: Ancestral proportion for the populations of *Q. variabilis* (a) and *Q. acutissima* (b) estimated based on SNPs and SVs when $K=2$. Each vertical bar represents the genetic composition of an individual, with different colors corresponding to each identified ancestry. The 2 genetic groups are indicated below the bar plots. c) Circos display of the genome-wide patterns of genetic variation estimated in 100 kb nonoverlapping windows for *Q. acutissima* and *Q. variabilis*. Tracks from outermost to innermost are SNP density (a), SV density (b), F_{ST} between East and West groups of each species (c), d_{XY} between East and West groups of each species (d), and nucleotide diversity (π) of 4 genetic group (e). Qv, *Q. variabilis*; Qa, *Q. acutissima*; QvE and QvW, East and West groups of *Q. variabilis*, respectively; QaE and QaW, East and West groups of *Q. acutissima*, respectively. d) Schematic diagram of the best-fitting demographic model inferred by fastsimcoal2. The strength and direction of gene flow between the 2 West groups of *Q. acutissima* and *Q. variabilis* were shown. All demographic parameters are detailed in supplementary table S6, Supplementary Material online.

methods to identify genetic variations potentially under natural selection. First, we performed genotype-environment association (GEA) analyses using latent factor mixed-effect models (LFMMs) (Frichot et al. 2013), identifying 10,474 and 9,326 SNPs significantly associated with at least 1 climate variable in *Q. acutissima* and *Q. variabilis*, respectively (supplementary fig. S9, Supplementary Material online). Second, we employed a multivariate approach, redundancy analysis (RDA), to evaluate associations between genetic and environmental factors (Capblancq and Forester 2021), detecting 1,620 and 4,393 candidate SNPs in *Q. acutissima* and *Q. variabilis*, respectively (supplementary fig. S9, Supplementary Material online). Finally, we performed a genome scan using PCAdapt (Duforet-Frebourg et al. 2014), identifying 11,913 and 5,767 SNPs strongly associated with population structure in *Q. acutissima* and *Q. variabilis*, respectively (supplementary fig. S9, Supplementary Material online). By integrating the candidate SNP detected through these methods, we obtained a total of 1,149 and 2,916 SNPs common to all 3 approaches, which were hence considered as strong candidates for local adaptation in *Q. acutissima* and *Q. variabilis*, respectively (Fig 4a and b and supplementary fig. S9, Supplementary Material online). Functional enrichment analyses showed that genes harboring candidate SNPs were involved in multiple abiotic stress-related functions (supplementary table S7, Supplementary Material online), suggesting a polygenic basis of adaptation. Using the same approaches, we detected 70 and 38 candidate SVs potentially

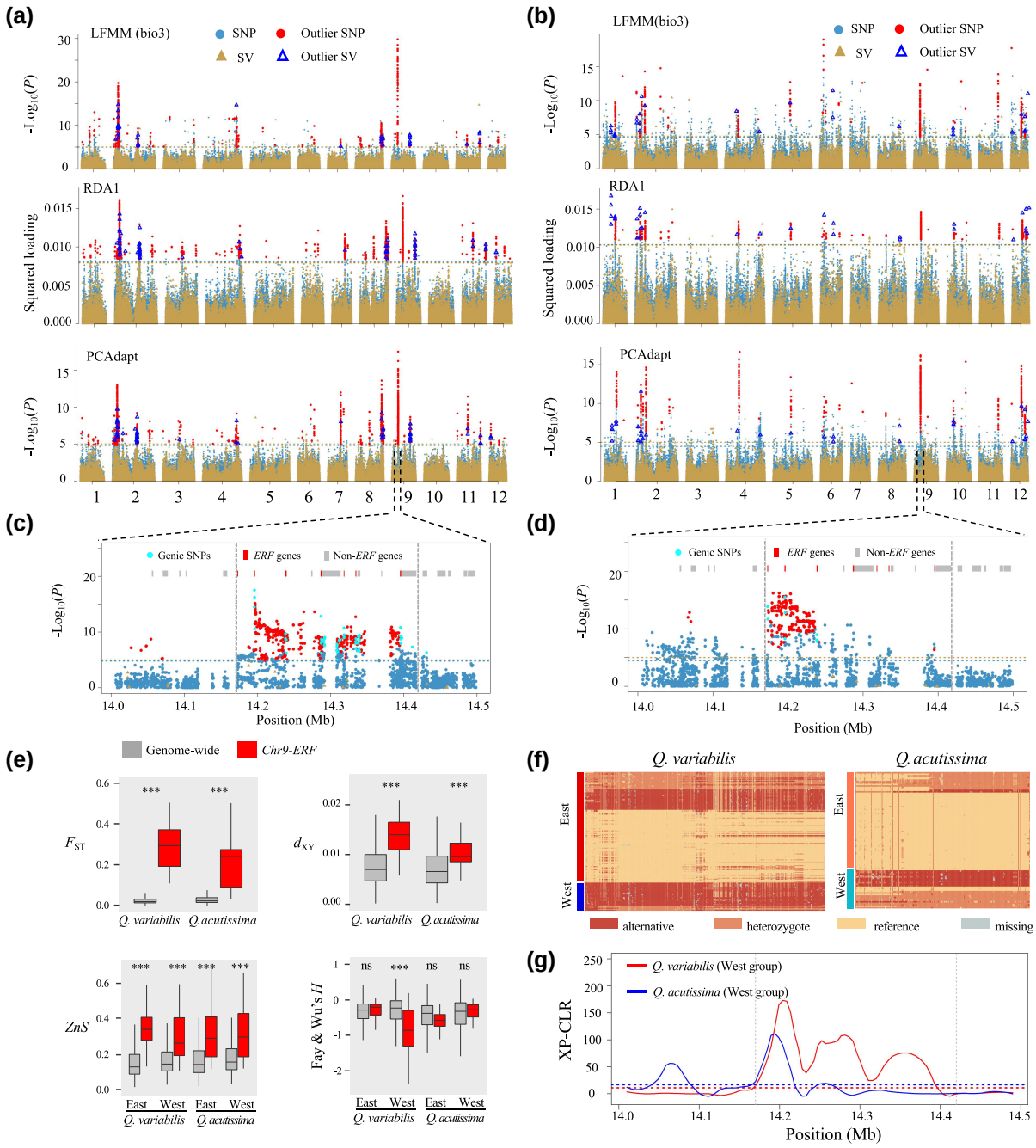

Fig. 4. Genomic signatures of local adaptation in *Q. variabilis* and *Q. acutissima*. a, b) Manhattan plots for candidate SNPs potentially under selection identified by a LFMM (upper panels), RDA (middle panels), and PCAdapt (lower panels) in *Q. variabilis* (a) and *Q. acutissima* (b). LFMM analyses revealed associations between genetic variants and isothermality (bio3), and RDA was used to estimate squared SNP loadings for the first RDA axis after controlling for population structure. Significant SNPs and SVs are indicated by red dots and blue triangles, respectively. Horizontal dashed lines represent the significance thresholds for each test. c, d) Magnification of PCAdapt results around the *Chr9-ERF* region for *Q. variabilis* (c) and *Q. acutissima* (d). The vertical dashed lines indicate the boundaries of the *Chr9-ERF* region (14.17–14.42 Mb on chromosome 9). Eight *ERF* genes annotated in the *Chr9-ERF* region are marked with red boxes, while other genes are indicated by gray boxes. Significant SNPs located in genic regions are represented by light blue dots. e) Comparison of 4 summary statistics between the *Chr9-ERF* region (red) and the genome-background (gray). The significance of each comparison was evaluated by using randomization test with 100,000 permutations (*$P<0.05$; **$P<0.01$; ***$P<0.001$; ^{ns}nonsignificant). f) Genotypes at 847 and 435 candidate SNPs located in the *Chr9-ERF* regions for *Q. variabilis* (left) and *Q. acutissima* (right), respectively. Rows indicate individuals from the East and West genetic groups defined in Fig. 3, while columns represent genotyped loci. g) Magnification of XP-CLR scores surrounding the *Chr9-ERF* region. The lines depicted are LOESS curves generated with a span of 0.2. Horizontal lines indicate the top 1% cutoff for XP-CLR scores. The vertical dashed lines indicate the boundaries of the *Chr9-ERF* region.

associated with adaptation in *Q. variabilis* and *Q. acutissima*, respectively. Genes located within 10 kb of these SVs are associated with stress and pathogen resistance and development functions (supplementary table S8, Supplementary Material online). Both candidate SNPs and SVs exhibited high genetic differentiation between the East and West genetic groups of *Q. variabilis* and *Q. acutissima* ($F_{ST}=0.26$–0.79 for SNPs and 0.21–0.66 for SVs), ~10–30 times greater than the genomic background (supplementary table S5, Supplementary Material online). These findings suggest that both SNPs and

SVs have facilitated adaptive differentiation between populations of *Q. variabilis* and *Q. acutissima*. Additionally, most candidate SVs were located more than 10 kb from adjacent candidate SNPs (supplementary fig. S10 and table S9, Supplementary Material online), indicating that they may be independently selected.

Approximately 29.05% (847) of the *Q. variabilis* candidate SNPs and 37.86% (435) of the *Q. acutissima* candidate SNPs were located in a region ~250 kb in size (~0.03% of the assembled genome) on chromosome 9 (Fig. 4c and d). This region contained 8 members of the ethylene-responsive factor (*ERF*) gene family and is henceforth referred to as the "*Chr9-ERF*" region (Fig. 4c and d and supplementary table S10, Supplementary Material online). GEA analyses conducted by LFMM revealed that candidate SNPs in the *Chr9-ERF* region were significantly associated with temperature (bio3 and bio4) and precipitation variables (bio15 and bio18; Fig. 4 and supplementary fig. S9, Supplementary Material online). Additionally, these candidate SNPs overlapped with 7 and 3 *ERF* genes in *Q. variabilis* and *Q. acutissima*, respectively (Fig. 4c and d and supplementary table S10, Supplementary Material online). The enrichment of climate-associated SNPs related to ecologically relevant *ERF* genes in the *Chr9-ERF* region suggested that this region is likely involved in local adaptation in *Q. variabilis* and *Q. acutissima*.

We found additional evidence for natural selection in the *Chr9-ERF* region. First, this region showed significant elevated relative differentiation (F_{ST}), net divergence (d_{XY}), and relative node depth (*RND*) between genetic groups ($P < 10^{-5}$, randomization test; Fig. 4e and supplementary table S5, Supplementary Material online). Second, the 2 genetic groups of *Q. variabilis* and *Q. acutissima* were each dominated by distinct alleles at the candidate SNPs in the *Chr9-ERF* region (Fig. 4f). Third, this region exhibited an excess of high-frequency derived alleles (more negative Fay and Wu's H) and strong linkage disequilibrium (LD; higher Z_{nS}) in 1 or both groups of *Q. variabilis* and *Q. acutissima* ($P < 10^{-5}$, randomization test; Fig. 4e and supplementary table S5, Supplementary Material online), in accordance with expectations under positive selection. Finally, we scanned the genome for signals of positive selection using the cross-population composite likelihood ratio test (XP-CLR) (Chen et al. 2010). Strong signals of selection were identified in the *Chr9-ERF* region for both genetic groups of *Q. variabilis* and *Q. acutissima* (Fig. 4g and supplementary fig. S11, Supplementary Material online).

Repeated Selection on Genetic Variants Between *Q. acutissima* and *Q. variabilis*

Given the largely sympatric distribution of *Q. acutissima* and *Q. variabilis*, these species are likely subjected to similar environmental selective pressure, which may lead to the recruitment of the same alleles, genes, or pathways for adaptation. We adopted the "molecular parallelism" framework proposed by Lee and Coop (2017, 2019), which suggests that parallel changes in allele frequency driven by natural selection can occur independently across species. According to this framework, beneficial alleles may arise through 3 distinct pathways: de novo mutations, standing genetic variation, or adaptive introgression. To test the hypothesis of molecular parallelism, we examined whether shared beneficial variants between *Q. acutissima* and *Q. variabilis* were more prevalent than expected by chance. In the *Chr9-ERF* region, 323 candidate SNPs were shared by *Q. acutissima* and *Q. variabilis*, accounting for 74.25% and 38.13% of the *Chr9-ERF* outliers found in these 2 species, respectively (Fig. 5a and supplementary table S11, Supplementary Material online). These ratios of overlapping SNPs were significantly higher than expected ($P < 2.2e^{-16}$, Fisher's exact test; supplementary table S12, Supplementary Material online), suggesting molecular parallelism at the SNP level in the *Chr9-ERF* region. Because selection acting on the same SNP does not necessarily favor the same allele, we further examined whether identical beneficial alleles were selected in parallel. Following the methods developed by Wang et al. (2021), we defined a SNP as "codirectional" (i.e. parallelism) if it showed the same allele frequency shift in the 2 West groups of *Q. variabilis* and *Q. acutissima*. Otherwise, a SNP was classified as "antidirectional" (i.e. nonparallelism). All 323 shared candidate SNPs were codirectional, while only 53.5% of randomly sampled SNPs displayed codirectionality (Fig. 5b and supplementary table S11, Supplementary Material online). In addition, similar patterns of geographic clines were detected for all shared candidate SNPs between *Q. acutissima* and *Q. variabilis*, but not for any background SNPs (Fig. 5c and d). These results provide strong evidence that same beneficial alleles were reused for adaptation within the *Chr9-ERF* region in *Q. acutissima* and *Q. variabilis*.

We further examined the extent of molecular parallelism outside the *Chr9-ERF* region. At the SNP level, 16 (0.77%–2.24%) candidate SNPs located outside *Chr9-ERF* were shared between species ($P < 2.2e^{-16}$, Fisher's exact test; Fig. 5e and supplementary table S12, Supplementary Material online), while 698/714 (97.76%) and 2,053/2,069 (99.23%) candidate SNPs were specific to *Q. acutissima* and *Q. variabilis*, respectively (supplementary table S12, Supplementary Material online). To examine the extent of molecular parallelism at the genic level, we considered a gene as a candidate for selection if any candidate SNPs were found within 10 kb upstream and downstream. In total, 137 and 412 candidate genes were identified in *Q. acutissima* and *Q. variabilis*, respectively, and only 5 (1.21%–3.65%) were shared between species ($P < 2.2e^{-16}$, Fisher's exact test; Fig. 5e and supplementary table S12, Supplementary Material online). We further examined the degree of parallelism at the level of gene function. In particular, we annotated candidate genes using Gene Ontology (GO) terms and quantified the overlap in functional enrichment between species. We found significant overlap in gene functions (868 GO terms; $P < 2.2e^{-16}$, Fisher's exact test), accounting for 35.69% and 71.74% of candidate functions in *Q. variabilis* and *Q. acutissima*, respectively (Fisher's exact test; Fig. 5e and supplementary table S12, Supplementary Material online). These results suggest that the extent of molecular parallelism is greater at the level of gene function than at SNP or gene levels.

Introgression Between *Q. acutissima* and *Q. variabilis*

Adaptive introgression is expected to generate clusters of locally adapted loci (Yeaman and Whitlock 2011) and may thus explain the patterns of molecular parallelism in the *Chr9-ERF* region. To test for gene flow between *Q. acutissima* and *Q. variabilis*, we phased the whole-genome re-sequencing

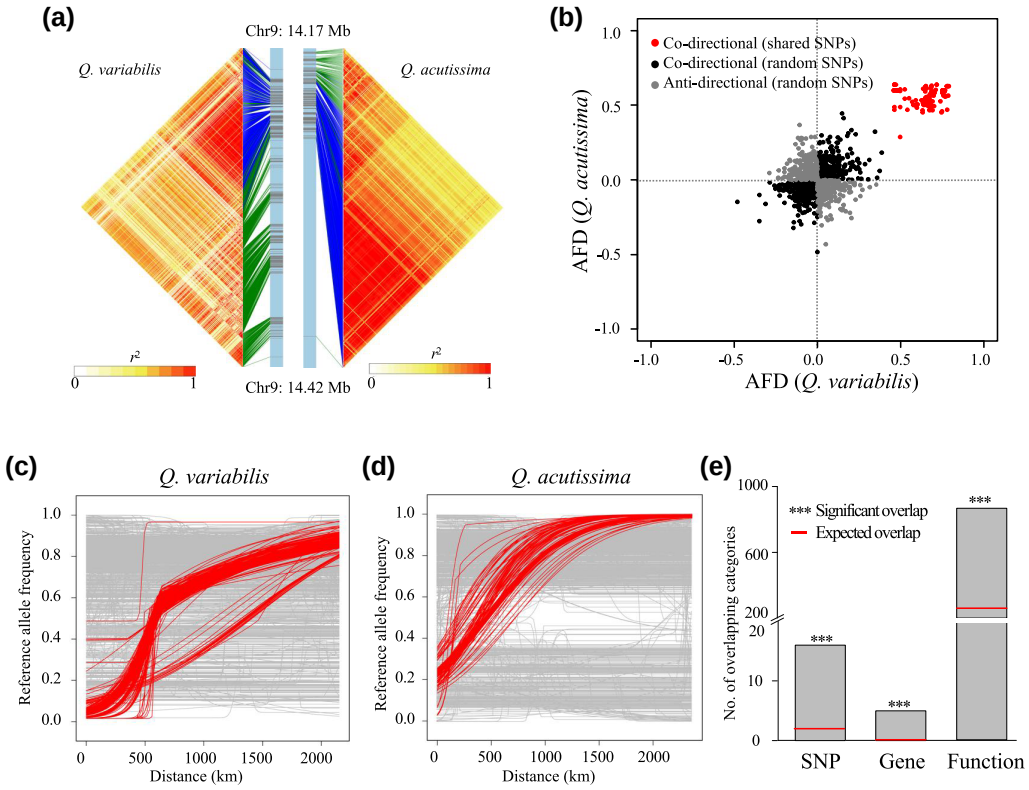

Fig. 5. Patterns of molecular parallelism within and outside the *Chr9-ERF* region in *Q. variabilis* and *Q. acutissima*. a) LD block plot showing pairwise r^2 values among all candidate SNPs in the *Chr9-ERF* region for *Q. variabilis* (left) and *Q. acutissima* (right). The blue and green lines indicate shared and species-specific candidate SNPs, respectively. The genomic positions of candidate SNPs are shown in gray at the middle bar. b) Alternative allele frequency difference (AFD) between the West and East groups in *Q. variabilis* (*x* axis) and *Q. acutissima* (*y* axis). Red dots indicate the 323 shared candidate SNPs located in the *Chr9-ERF* region that were codirectional. Black and gray dots represent 1,000 randomly sampled SNPs of which 535 were codirectional (black) and 459 were antidirectional (gray). c, d) Geographic clines of 323 shared candidate SNPs (red lines) and the 1,000 randomly sampled SNPs (gray lines) in *Q. variabilis* (c) and *Q. acutissima* (d). e) Number of overlapping candidate SNPs, genes and functions (GO terms) for local adaptation outside the *Chr9-ERF* region between *Q. variabilis* and *Q. acutissima*. Categories indicated by asterisks exhibited higher overlap of the candidates than expected by chance (***$P < 2.2e^{-16}$, Fisher's exact test).

data and inferred the phylogenetic relationships among haplotypes of the 2 species and *Q. chenii*. Phylogenetic analyses based on the *Chr9-ERF* region grouped haplotypes into 3 clades (Fig. 6a). The first clade consisted of haplotypes mainly from the East genetic group of *Q. variabilis* (henceforth "Qv-E"), and the second clade included haplotypes mainly from the East group of *Q. acutissima* (henceforth "Qa-E"). The third clade contained haplotypes from all 3 species and was further subdivided into 3 haplotype groups ("Qa-W," "Qv-W," and "Qc") according to the origins of these haplotypes. The Qa-W group contained haplotypes mainly from the West group of *Q. acutissima*, the Qv-W group dominated the West group of *Q. variabilis*, and the Qc group included all haplotypes from *Q. chenii* (Fig. 6a and b and supplementary fig. S13, Supplementary Material online). Notably, Qa-W and Qv-W haplotype groups were not monophyletic, while Qc haplotypes formed a distinct group nested within Qa-W and Qv-W haplotypes (Fig. 6a). This grouping pattern based on the *Chr9-ERF* region was different from that based on genome-wide SNPs, where each species was recovered as monophyletic (Fig. 6a). Therefore, we infer that the close relationship between Qa-W and Qv-W haplotypes on the *Chr9-ERF* gene tree was most likely due to interspecific gene flow between the West groups of *Q. acutissima* and *Q. variabilis*.

Introgression between *Q. acutissima* and *Q. variabilis* in the *Chr9-EFR* region was further supported by the observed patterns of genetic variation. First, efficient local ancestry inference (ELAI) analyses (Guan 2014) revealed that the predominant ancestry of Qv-W haplotypes in the *Chr9-ERF* region was Qa-W haplotypes (allele dosages = ca. 0.8), which is significantly different from the genome-wide averages (Fig. 6c and supplementary fig. S14, Supplementary Material online). Second, the Qa-W and Qv-W haplotype groups had an excess of shared derived variants in the *Chr9-ERF* region, showing significant introgression signals (f_d and d_f statistics; false discovery rate (FDR) < 0.01; Fig. 6d and supplementary fig. S15, Supplementary Material online). Third, the Qv-W and Qa-W haplotypes showed the lowest genetic divergence (F_{ST} and d_{XY}) among all comparisons of haplotype groups in the *Chr9-ERF* region (Fig. 6e, f, and i and supplementary fig. S16 and table S13, Supplementary Material online). Notably, both Qv-W and Qa-W haplotypes showed elevated divergence from Qv-E haplotypes across the whole *Chr9-ERF* region relative to genomic background ($P < 10^{-5}$, randomization test; Fig. 6g and h and supplementary fig. S16 and table S13, Supplementary Material online). By contrast, the elevated divergence between Qv-W and Qa-W vs. Qa-E was solely detected across the first part of the *Chr9-ERF* region (Fig. 6g and h and supplementary fig. S16 and table S13, Supplementary Material online). These results were compatible with a scenario in which Qa-W diverged from Qa-E haplotypes across the first part of the *Chr9-ERF* region in *Q. acutissima* and then introgressed into *Q. variabilis*,

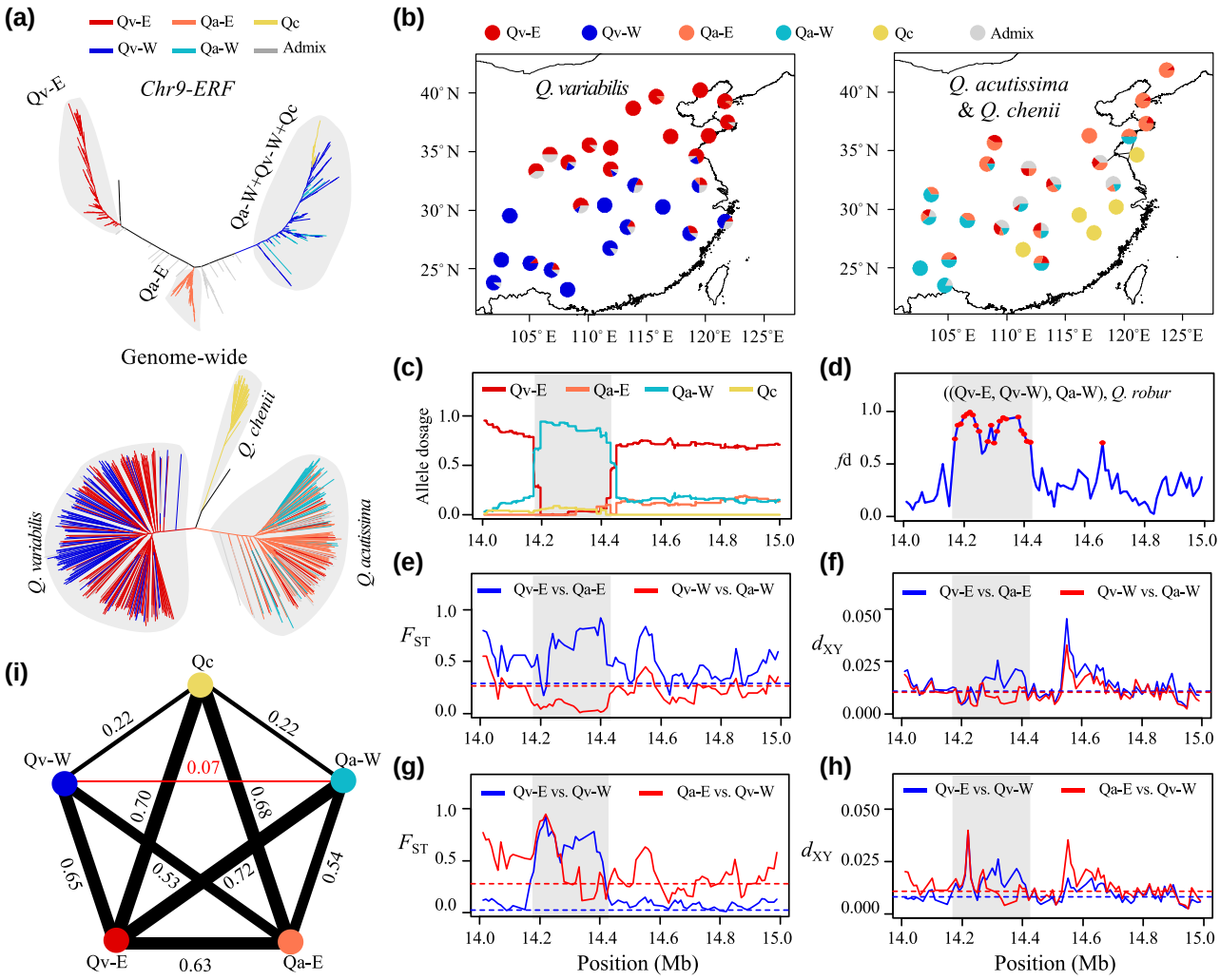

Fig. 6. Adaptive introgression of beneficial alleles in the *Chr9-ERF* region between *Q. variabilis* and *Q. acutissima*. a) Phylogenetic relationships of haplotypes based on the *Chr9-ERF* region (upper) and genome-wide (lower) SNP data. Five haplotype groups were identified based on the *Chr9-ERF* topology and origins of these haplotypes. b) Geographic distribution of haplotype groups in *Q. variabilis* (left) and in *Q. acutissima* and *Q. chenii* (right). c) Allele dosage of Qv-W haplotypes inferred using ELAI. Four haplotype groups (Qv-E, Qa-E, Qa-W, and Qc) were used as sources. d) Gene flow (f_d) between Qv-W and Qa-W estimated using the ABBA–BABA test. Red dots indicate windows with a significant introgression signal (FDR < 0.01). e–h) Genetic divergence (F_{ST} and d_{XY}) between haplotype groups. Horizontal dashed lines represent the genome-wide averages. In (c–h), all statistics were estimated in 10 kb nonoverlapping windows, and the shaded area indicates the *Chr9-ERF* region. i) Genetic differentiation (F_{ST}) between all haplotype groups. The numbers above the lines indicate the average F_{ST} between groups in the *Chr9-ERF* region, and a thicker line reflects higher values. The lowest F_{ST} value (Qv-W vs. Qa-W) is highlighted in red. For c-i, Qv-W, and Qa-W represent haplotypes exclusively from the West genetic groups of *Q. variabilis* and *Q. acutissima*, respectively. Similar results were obtained for Qv-W and Qa-W haplotypes derived from the East genetic groups of these 2 species (see supplementary figs. S14–16 and table S13, Supplementary Material online).

generating the Qv-W haplotypes. The alternative hypothesis of gene flow from *Q. variabilis* to *Q. acutissima* was deemed unlikely, as it would result in high divergence between Qv-W and Qa-E haplotypes across the entire length of the *Chr9-ERF* region. Altogether, our results indicate that beneficial variants in the *Chr9-EFR* region originated in the West groups of *Q. acutissima*, followed by introgression into *Q. variabilis*. These results point to gene flow as an important source of advantageous variants for adaptive evolution in oak species.

Discussion

Structural Variation Contributes to Adaptive Evolution

In this study, we performed de novo assembly and annotation of 22 representative accessions of *Q. variabilis*. Our pan-genome analyses revealed that up to 73% of gene families are dispensable (i.e. absent in some accessions), a significantly higher proportion compared to other plants studied with similar sample sizes, such as *A. thaliana* (31%) (Kang et al. 2023), *Panicum miliaceum* (58%) (Chen et al. 2023a), and soybean (64%) (Liu et al. 2020). The high proportion of dispensable gene families in *Q. variabilis* aligns with the observed high genetic diversity in this species. Gene enrichment analyses indicate that these dispensable genes are enriched in functions related to abiotic and biotic stress responses, consistent with findings in other plant species (Liu et al. 2020; Qin et al. 2021; Chen et al. 2023a; Kang et al. 2023). Previous studies have demonstrated that dispensable genes contribute to local adaptation and reproduction isolation (Hu et al. 2022; Li and Lee 2023; Wang et al. 2023a; Liu et al. 2024; Shi et al. 2024). Based on these findings, we propose that dispensable gene families may have significant functional implications

and play important roles in driving adaptive divergence in *Q. variabilis*.

Our pan-genome analyses also identified 543,372 non-redundant SVs among the 22 *Q. variabilis* assemblies, covering 625 Mb sequences absent from the reference genome. Genes with SV located in exon regions (SV-exon-genes) exhibited lower expression levels compared to non-SV-genes. These findings suggest that SVs may play a crucial role in shaping genomic architecture and influencing gene expression in *Q. variabilis*, similar to observations in other species such as cacao (Hämälä et al. 2021) and *C. sinensis* (Chen et al. 2023b). Alternatively, the low expression of SV-exon-genes could be due to these genes being less likely to be under purifying selection, making them more likely to accumulate SVs as a result of genetic drift. We also found that SV-outliers exhibit signals of selection that do not overlap with SNP-outliers, suggesting that SNP-based analyses may not detect the same candidate genes associated with SV-outliers. Previous studies have demonstrated that SVs can generate significant phenotypic variation affecting important agronomic and adaptive traits in plants (Liu et al. 2020; Qin et al. 2021; Chen et al. 2023a, 2023b; Kang et al. 2023; Yan et al. 2023). Therefore, SVs likely contribute to adaptive evolution in *Q. variabilis* and *Q. acutissima*.

While our study provides valuable insights into the role of SVs in adaptive evolution, several limitations should be noted. First, the exclusion of SVs not present in the 22 *Q. variabilis* assemblies may have led to an underestimated contribution of SVs to the evolution of *Q. variabilis* and *Q. acutissima*. To ensure high-quality SV calls, we focused exclusively on SVs identified in the 22 *Q. variabilis* assemblies. While this conservative approach can minimize false positive SV calls, it may have overlooked SVs absent from the pan-genome panel, including those specific to *Q. acutissima* or other *Q. variabilis* individuals not sampled. However, if the adaptive SVs that originated from *Q. acutissima* are present at a high frequency in *Q. variabilis*, they may be polymorphic and detectable by our pan-genome panel. Second, the reliance on short-read sequencing data limited our ability to genotype a significant proportion of SVs in population level. Approximate 52% of the SVs in the panel were genotyped in fewer than half of the *Q. acutissima* or *Q. variabilis* individuals using short-read sequencing data, and were therefore excluded from population genetic analyses. A similar challenge has been observed in rice, where only 16% of the SVs identified in the pan-genome could be genotyped in other individuals based on short-read sequencing data (Qin et al. 2021). Future studies that integrate long-read sequencing data and more advanced analytical approaches will provide deeper insights into the role of SVs in genome evolution and adaptation.

Selection in the Chr9-ERF Region Drives Genetic Divergence in *Q. variabilis* and *Q. acutissima*

Population genomic analyses based on SNPs and SVs revealed a similar east–west divergence pattern between *Q. variabilis* and *Q. acutissima* and identified a genomic region (*Chr9-ERF*) enriched for candidate variants involved in local adaptation in both species. This region exhibited elevated genetic divergence and alleles frequency differentiation between genetic groups, along with an excess of high-frequency derived alleles and increased LD compared to the genomic background, consistent with expectations under natural selection. Additionally, several candidate SNPs were located within or adjacent to *ERF* genes in the *Chr9-ERF* region. The *ERF* genes belong to the *AP2/ERF* transcription factors family, which plays an essential role in plant responses to environmental stresses, including heat, drought and cold (Agarwal et al. 2017; Huang et al. 2021; Han et al. 2022). The East and West groups of *Q. variabilis* and *Q. acutissima* occupy distinct ecological niches, with varying temperatures and water availability. The East group is distributed in eastern and northern China, where the climate is arid and highly seasonal, while the West group is found in southwestern China, characterized by a more humid climate with lower seasonality (Zhao 2018). Therefore, the *ERF* genes may have contributed to the adaptation of *Q. variabilis* and *Q. acutissima* populations to their local habitats. Feature studies focused on validating the functions of candidate SNPs by dissecting the complex climate-adaptive traits and their underlying genes would provide valuable insights.

A considerable number of candidate SNPs were shared between *Q. variabilis* and *Q. acutissima* in the *Chr9-ERF* region, suggesting a pattern of repeated selection. This result is consistent with previous studies, which have shown that closely related species often use same genes or alleles to adapt to similar environmental conditions (Bohutínská et al. 2021 ; Wang et al. 2021; Montejo-Kovacevich et al. 2022; Wang et al. 2023c). The *Q. variabilis* and *Q. acutissima* are recently diverged species that occupy largely overlapping regions. Therefore, the observed repeated selection in the *Chr9-ERF* region is likely the result of similar genomic backgrounds and consistent environmental selective pressure acting on both species. Additionally, the *Chr9-ERF* region contains multiple duplicated *ERF* genes, consistent with previous findings that duplicated genes are repeatedly implicated in adaptation likely because increased redundancy makes them more likely targets of positive selection (Yeaman et al. 2016; Nocchi et al. 2024). Outside of the *Chr9-ERF* region, we observed molecular parallelism primarily at the level of gene function rather than at the level of SNPs or individual genes. Our findings align with theoretical predictions and empirical evidence, suggesting that repeated selection tends to occur at higher levels of organization during polygenic adaptation (Chevin et al. 2010; Yeaman 2015; Bohutínská et al. 2021; Cheng et al. 2021b; Wilkinson et al. 2021). Overall, our study demonstrated that the same alleles were repeatedly selected in the *Chr9-ERF* region but not in other regions, a pattern likely influenced by multiple factors, including the genomic basis of adaptive traits and the evolutionary history of functional genes (Arendt and Reznick 2008; Bolnick et al. 2018; James et al. 2023).

Adaptive Introgression Provides Advantageous Variants for Local Adaptation

The role of adaptive introgression in facilitating the spread of beneficial alleles across species boundaries is increasingly recognized as a key driver of local adaptation, particularly in species with overlapping ranges and frequent hybridization events (Savolainen et al. 2007; Kremer and Hipp 2020). In this study, we provide compelling evidence that the beneficial haplotypes in the *Chr9-ERF* region likely originated in the West group of *Q. acutissima* and were subsequently introgressed into *Q. variabilis*. This hypothesis is supported by phylogenetic analyses, local ancestry inference, and signals of introgression (e.g. f_d and d_f statistics). Notably, 26% and 62% of outliers in the *Chr9-ERF* region are species-specific

to *Q. acutissima* and *Q. variabilis*, respectively. This distinct mutational spectrum between species contradicts expectations from recent gene flow and instead supports a scenario of independent selective sweeps acting on this genomic region in each species. The *Chr9-ERF* region (~250 kb) exhibits an exceptional length compared to other introgressed segments between West groups of *Q. acutissima* and *Q. variabilis* (10 – 120 kb; supplementary fig. S15B, Supplementary Material online). Considering the high recombination rate in *Q. variabilis*, the maintenance of *Chr9-ERF* haplotype block strongly implies strong selection suppressing recombination after introgression. Therefore, the *Chr9-ERF* region in *Q. variabilis* is most likely a consequence of adaptive introgression. The strong signals of adaptation and introgression in the *Chr9-ERF* region aligns with theoretical expectation that adaptive introgression often generates large genomic blocks containing tightly linked adaptive alleles (Yeaman and Whitlock 2011). Similarly, a recent study demonstrated that adaptive introgression between *Q. acutissima* and *Q. variabilis* resulted in clusters of introgressed SNPs in specific chromosome regions (Fu et al. 2022). We further propose that extensive intraspecific gene flow may have facilitated the regional spread of *Chr9-ERF* haplotypes across southwestern populations of *Q. variabilis*, highlighting the interplay between selection and gene flow in shaping adaptive landscapes. Furthermore, phylogenetic analyses revealed that the haplotypes of the West group of *Q. variabilis* did not form a monophyletic group but were intermixed with haplotypes from the West group of *Q. acutissima* (Fig. 6a). This pattern suggests that multiple haplotypes with different genetic backgrounds were transferred from *Q. acutissima* to *Q. variabilis* through repeated hybridization events, supporting a scenario of recurrent introgression. Repeated introgression can increase the initial frequency of beneficial haplotypes in the recipient species, allowing selection to act more efficiently on the introgressed alleles and thereby promoting adaptation. This mechanism resembles the selection dynamics observed in standing genetic variation (Barrett and Schluter 2008).

While advantageous variants in the *Chr9-ERF* region could theoretically arise from de novo mutations or standing genetic variation, these scenarios are very unlikely. First, given the recent divergence between *Q. acutissima* and *Q. variabilis*, the probability that all 323 shared mutations in the *Chr9-ERF* region accumulated independently after species divergence is extremely low (Probability = $2.75e^{-240}$; see Materials and Methods). Second, although recently diverged lineages may share polymorphisms inherited from their most recent common ancestor, these variants are not expected to cluster in a short genomic region (Yeaman 2013). Chromosome inversions can maintain multiple adaptive alleles in LD (Wellenreuther and Bernatchez 2018). However, we observed high collinearity across genome assemblies around the *Chr9-ERF* region (Fig. 1b and supplementary fig. S12, Supplementary Material online), suggesting that the shared advantageous polymorphisms in *Q. acutissima* and *Q. variabilis* within this region is not due to an inversion.

Altogether, our results revealed the introgression of the *Chr9-ERF* region from *Q. acutissima* to *Q. variabilis*, providing a clear example of how gene flow can drive local adaptation in response to similar environmental pressures. While previous studies in oaks have identified adaptive introgression events involving SNPs in oaks (e.g., Leroy et al. 2020a; Fu et al. 2022), our findings highlight the role of large genomic blocks, such as the ~250 kb *Chr9-ERF* region, in facilitating the transfer of multiple linked adaptive alleles. Considering the extensive gene flow observed among oak species (Kremer and Hipp 2020), we propose that adaptive introgression may have played a pivotal role in adaptation and speciation in oaks and could also contribute to their response to future climate change. Future research should focus on functional validation of the introgressed alleles and their role in climate adaptation, as well as exploring the broader genomic landscape of introgression.

Conclusion

Our study applied pan-genome analyses to capture and describe most of the diversity in *Q. variabilis* genomes and provided a comprehensive view of how SVs contribute to local adaptation. We identified a genomic region (*Chr9-ERF*) under strong selection in both *Q. acutissima* into *Q. variabilis*, which may have played a role in the genetic divergence within each species. We further demonstrated that advantageous variants in the *Chr9-ERF* region were introgressed from western populations of *Q. acutissima* into *Q. variabilis*, suggesting that hybridization facilitated local adaptation in these species. In summary, this study provides a valuable genetic resource by assembling multiple *Q. variabilis* genomes and highlights the importance of pan-genome analysis in understanding patterns of genetic variation and the underlying mechanisms of adaptive evolution in forest trees.

Materials and Methods

Genome Sequencing

Genomic DNA was extracted from fresh leaves using a Plant DNA Kit (Bioteke, Beijing, China). For short-read sequencing, a paired-end library with insert sizes of ~350 bp was constructed and sequenced on an Illumima NovaSeq 6000 Sequencing System to generate 150 bp paired-end reads. For PacBio long-read sequencing, a 15 kb HiFi library was prepared and sequenced on PacBio Sequel II platform. To improve the assembly quality, ultra-long reads (N50 > 100 kb) was generated using Oxford Nanopore Technologies (ONT). The ONT library was prepared using an Oxford Nanopore SQK-LSK109 kit and sequenced on the PromethION platform. For Hi-C sequencing, a Hi-C library was constructed and sequenced on an Illumina HiSeq 2500 platform to obtain 150 bp paired-end reads. For RNA sequencing, a total of 19 individuals were included (supplementary table S1, Supplementary Material online). Among them, the "SDTS" individual was sequenced for its leaf, root, and flower tissues. For the remaining 18 individuals included in pan-genome analysis, transcriptome sequencing was performed on their leaf tissues. However, due to insufficient tissue availability, transcriptome data was missing for 3 individuals: "CLM," "GXNN," and "SS." Total RNA was extracted using a plant RNA extraction kit (Omega Biotek, Norcross, GA, USA). The mRNA was purified using poly-T oligo-attached magnetic beads bound to complementary DNA and sequenced to generate 150 bp paired-end reads on the NovaSeq 6000 platform.

De Novo Genome Assembly

To generate a chromosome-level haplotype-resolved reference genome assembly for the "SDTS" individual, we first integrated HiFi long reads, ONT ultra-long reads, and Hi-C reads

to produce 2 haplotype assemblies at the contig-level using hifiasm v0.18.6-r513 (Cheng et al. 2021a). Next, the Hi-C reads were mapped onto the assembled haplotype contigs using Juicer v2.0 (Durand et al. 2016b), and chromosomes were assembled using the 3D-DNA (v201008) pipeline (Dudchenko et al. 2017), with the parameters "-m haploid -i 150000 -r 0 –editor-repeat-coverage 5." The chromosome segmentation boundaries and assembly errors were manually checked and adjusted using Juicebox v1.11.08 (Durand et al. 2016a). The ultra-long ONT reads were used to close remained gaps with LR_Gapcloser (Xu et al. 2019). To correct the assembly errors, Nextpolish2 v0.1.01 (Hu et al. 2024) was used to polish the assembly based on HiFi reads and short reads. For the other 20 individuals (supplementary table S1, Supplementary Material online), HiFi long reads were assembled into contigs and then anchored to the 12 pseudo-chromosomes of SDTS-hap1 using RadTag (Alonge et al. 2022).

To evaluate the quality of the genome assembly, we first estimated the BUSCO score using BUSCO v5.4.3 (Simão et al. 2015). Second, we assessed the quality of repetitive genomic regions using the LAI vbeta3.2 program (Ou et al. 2019). Finally, we estimated the per-base consensus accuracy (QV) using Merqury v1.3 (Rhie et al. 2020), based on PacBio HiFi long reads.

Genome Annotation

To annotate the repeat regions in the *Q. variabilis* genome, we used TRF 4.09 (Benson 1999) to detect tandem duplications and employed extensive de novo TE annotator v1.9.6 (Ou et al. 2019) for annotating TEs. Additionally, we utilized RepeatMasker v4.1.0 (Tarailo-Graovac and Chen 2009) to identify the repeat regions based on Repbase (http://www.girinst.org/repbase).

To predict protein-coding genes, we employed 3 approaches, homology-based prediction, ab initio prediction and RNA-seq-assisted prediction. For homolog-based prediction, proteins sequences of related species were download from the Swiss-Prot database (https://www.uniprot.org/downloads), *A. thaliana* Araport11 (Cheng et al. 2017) and *Populus trichocarpa* v4.1 (Tuskan et al. 2006). For ab initio prediction, coding regions of genes were predicted by using Augustus v3.4.0 (Stanke et al. 2008; Keller et al. 2011), GeneMark-ES v4.69_lic (Lomsadze et al. 2005), and SNAP v2006-07-28 (Korf 2004). For RNA-seq-assisted prediction, 21 *Q. variabilis* transcriptomes (see above) were assembled using StringTie v2.2.1 (Pertea et al. 2015) based on RNA sequencing data and mapped to the genome assembly using Hisat2 v2.2.1 (Kim et al. 2015). The gene models were then predicted using PASAv2.5.2 (Haas et al. 2003) based on the alignment. Finally, gene models predicted by the 3 methods were integrated into consensus gene models using MAKER v3.01.04 (Cantarel et al. 2008) and EVidenceModeler v1.1.1 (Haas et al. 2008), and further improved using PASA. The longest transcript of each predicated gene model was retained to obtain a nonredundant gene set. The completeness of gene models was assessed by searching the gene content of the embryophyta_odb10 database using BUSCO (Simão et al. 2015).

To infer gene functions, the protein sequences of *Q. variabilis* were searched against SwissProt (https://www.uniprot.org/) and NR (ftp://ftp.ncbi.nih.gov/blast/db/) databases using BLASTP v2.13.0+ (Altschul et al. 1990). The protein domains were predicted based on 6 publicly available databases (ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE) using InterProScan v5.56-89.0 (Blum et al. 2021). The GO terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs were inferred by using eggNOG-mapper v2.1.9 (Cantalapiedra et al. 2021) and KofamKOALA v1.3.0 (Aramaki et al. 2020), respectively.

To annotate noncoding RNAs (ncRNAs), we predicted tRNAs using tRNAscan-SE v1.4 (Chan and Lowe 2019), detected rRNAs by aligning rRNA sequences from related species to the genome using BLASTN, and identified other ncRNAs (e.g. miRNAs and snRNAs) using infernal (Nawrocki et al. 2009) based on the Rfam database.

Collinearity Analyses

To quantify the collinearity among the 22 *Q. variabilis* assemblies, we identified and visualized syntenic blocks across the 12 pseudo-chromosomes using JCVI (Tang et al. 2024). To further assess the synteny relationships between each pair of assemblies, we calculated the SRI parameter developed by Chen et al. (2023b). Finally, we extracted nonsyntenic regions between each of the 20 assemblies and SDTS-hap1 using the SyRI tool (Goel et al. 2019) and calculated synteny diversity using SynDiv (Du et al. 2024) in 100 kb nonoverlapping windows across the genome.

We also investigate genome collinearity around *Chr9-ERF* region between *Q. variabilis* and *Q. acutissima*. To do that, we sequenced 2 *Q. acutissima* individuals on the PacBio Sequel II platform (supplementary table S1, Supplementary Material online). These 2 individuals (Qa-LN and Qa-KM) were chosen to represent the East and West groups of *Q. acutissima*, respectively. A total of 21.44–42.05 Gb of PacBio long reads was obtained for each individual (supplementary table S1, Supplementary Material online). Draft genomes of these 2 individuals were assembled using Hifiasm (Cheng et al. 2021a) following the same procedure applied for *Q. variabilis*. These newly generated assemblies were aligned against the reference genome of *Q. variabilis*, and the collinearity around the *Chr9-ERF* region was investigated using SyRI (Goel et al. 2019) and Plotsr (Goel and Schneeberger 2022).

Pan-Gene Analyses

To build the gene repertoire of *Q. variabilis*, we clustered the protein-coding genes from 21 assemblies and 1 previously published assembly (CLM) (Wang et al. 2023b) using OrthoFinder v2.5.4 (Emms and Kelly 2019). For the 2 haplotype-resolved genomes (SDTS and CLM), we included only the genes annotated in the longer haplotype (SDTS-hap1 and CLM-hap1) to avoid redundancy. The genes clusters were classified into 4 categories: core gene clusters, conserved in all the 22 individuals; soft-core gene clusters, present in 20–21 individuals; shell gene clusters, found in 2–19 individuals, and private gene clusters, containing genes from only 1 individual. Nonsynonymous/synonymous substitution ratios (Ka/Ks) within the core, soft-core and shell gene clusters were computed using KaKs_Calculator v3.0 (Zhang 2022). Functional enrichment analyses for each gene categories were performed using clusterProfiler v4.6.0 (Wu et al. 2021).

Graph-Based Pan-Genome Construction

To construct a graph-based pan-genome of *Q. variabilis*, we called genomic variants (SNP, indels and SVs) across 22 *Q. variabilis* assemblies (supplementary table S1,

Supplementary Material online), using SDTS-hap1 as the reference. To detect SVs, we aligned HiFi long reads to the reference genome and identified SVs using 4 methods: PBSV v2.9.0 (https://github.com/PacificBiosciences/pbsv), SVIM v2.0.0 (Heller and Vingron 2019), Sniffles v2.2 (Sedlazeck et al. 2018), and cuteSV v2.0.3 (Jiang et al. 2020). In addition, we called SVs using 2 assemblies-based pipelines. First, assemblies were aligned to the reference genome using minimap2 v2.24-r1122 (Li 2018) with parameters "-x asm5 -a --eqx –cs" and SVs were called using SVIM-asm v1.0.3 (Heller and Vingron 2021). Second, assemblies were aligned to the reference genome using nucmer in MUMmer v4.0.0rc1 (Kurtz et al. 2004) with parameters "--maxmatch -c 500 -b 500 -l 100." Alignments with <90% identity and shorter than 100 bp were discarded. SVs (>50 bp) were then identified using Assemblytics v1.2.1 (Nattestad and Schatz 2016). Finally, SVs detected by the 6 methods were merged using SURVIVOR v1.0.3 (Jeffares et al. 2017), allowing a maximum distance of 50 bp and reporting only calls supported by at least 2 callers.

To call SNPs and indels (<50 bp), we mapped HiFi long reads of each individual to the reference genome using Winnowmap2 v2.03 (Jain et al. 2022) with the parameters "x map-pb -a -Y -L --eqx --cs --MD --secondary = no." Variants were called using DeepVariant v1.5.0 (Poplin et al. 2018) with the parameters "--model_type = PACBIO." The graph-based pan-genome was constructed by integrating the linear reference genomes (SDTS-hap1) and the identified variations using the graph toolkit (vg) v1.50.1 (Garrison et al. 2018).

Whole-Genome Re-Sequencing and SNP and SV Calling

To facilitate population genomic analyses, 143 individuals were sampled from 29 *Q. variabilis* populations, 93 individuals from 21 *Q. acutissima* populations, and 21 individuals from 5 *Q. chenii* populations (supplementary table S4, Supplementary Material online). In each population, 3–5 adult trees separated by at least 50 m were selected. Among these samples, 107 individuals were sequenced for this study, and the remaining 150 individuals were sequenced in previous studies (Liang et al. 2021; Yuan et al. 2023; supplementary table S4, Supplementary Material online). For each individual, genomic DNA was extracted from silica gel-dried leaves using a Plant DNA Kit (Bioteke) and sequenced on the Illumina NovaSeq 6000 Sequencing System to generate 150 bp paired-end reads. Sequencing for each sample was carried out to achieve coverage of 28.3×–50.8× (supplementary table S4, Supplementary Material online).

Raw sequencing reads were filtered using Trimmomatic v0.39 (Bolger et al. 2014) to remove adapter sequences and low-quality base pairs (base quality <20). Trimmed reads with length ≥ 36 bp were subsequently aligned to reference genome (SDTS-hap1) using BWA-MEM with default parameters in BWA v0.7.15 (Li and Durbin 2010). Aligned reads were sorted using SAMtools v1.8 (Li et al. 2009). PCR duplicates were marked using MarkDuplicate and AddOrReplaceReadGroup implemented in GATK v4.1 (DePristo et al. 2011) and filtered out using Picard (https://github.com/broadinstitute/picard). Genotypes of each individual were called using HaplotypeCaller implemented in GATK and then merged into a VCF file. Sites with base quality less than 30, with mapping quality <30 or located in repeat regions were excluded from SNP calling.

To obtain high-quality SNPs, we further applied strict filtering steps. (i) Homozygous genotypes were set as missing if supported by fewer than 4 reads. (ii) Heterozygous genotypes were set as missing if the minor allele was supported by <2 reads or if the read ratio (number of reads supporting the minor allele divided by the number of reads supporting the major allele) was <0.1. (iii) Sites with more than 2 alleles, mean sequencing depth lower than 5 or higher than 100, proportion of called heterozygous genotypes >50% or genotyped in fewer than 70% individuals were removed. (iv) All indels were removed. A total of 21,695,563 reliable SNPs called in 257 individuals were retained for subsequent analyses, unless specified otherwise.

To genotype the SVs, Illumina short reads were mapped to the graph-based pan-genome using vg giraffe (Sirén et al. 2021) and the SV genotyping was performed using vg call (Hickey et al. 2020). To ensure the acquisition of high-quality SVs, we retained only SVs with quality scores exceeding 30 and genotyped in more than 50% individuals, obtaining 259,877 high-quality SVs.

Population Structure Analyses

To infer population structure within and between species, we conducted admixture analyses using NGSadmix (Skotte et al. 2013), applied a PCA using PCAngsd v1.0.1 (Meisner and Albrechtsen 2018), and constructed NJ trees using MEGA-X (Kumar et al. 2018). NGSadmix was run with predefined clusters (K) ranging from 1 to 10, each repeated 10 times with different random seeds. The optimal K (i.e. the number of putative genetic groups) was selected by the ΔK method (Evanno et al. 2005). Population structure analyses were performed based on common SNPs and SVs with minor allele frequency (MAF) > 0.1, missing rate < 0.1, and LD (r^2) < 0.2. PLINK v1.07 (Purcell et al. 2007) was used to estimate LD between SNPs. To account for sequencing errors and uncertainty in genotype calling, NGSadmix and PCAngsd were applied based on genotype likelihoods.

Estimation of Population Summary Statistics

To investigate the landscape of genomic variation, ANGSD v0.920 (Korneliussen et al. 2014) was used to calculate population summary statistics, including π (Tajima 1989), Tajima's D (Tajima 1989), F_{ST} (Weir and Cockerham 1984), and d_{XY} (Nei 1987), based on the folded site frequency spectrum (SFS). We also calculated Fay and Wu (2000) H based on the unfolded SFS using ANGSD. To control for variation in mutation rates, we estimated RND (Feder et al. 2005) between populations and species. All summary statistics were calculated across 10 kb nonoverlapping sliding windows. To quantify the recombination rate across the genome, we calculated the ZnS statistic (Kelly 1997) by averaging the LD (r^2) values between all SNP pairs within a 10 kb sliding window.

To polarize the ancestral state of each SNP, we used *Q. robur* (Plomion et al. 2018) and *Q. suber* (Ramos et al. 2018) as 2 outgroup species because they are closely related to the 3 oak species investigated in this study. Briefly, we downloaded genome sequences of *Q. robur* from the European Nucleotide Archive (accession number: PRJEB19898) and *Q. suber* from GenBank (accession number: PKMF0000000) and then constructed a multiple genome alignment with the *Q. variabilis* reference genome. Because *Q. robur* was more distantly related to *Q. variabilis* (Zhou et al. 2022b), the ancestral state was

determined as follows: (i) the allele shared with *Q. robur* was the ancestral allele. Otherwise, (ii) the allele shared with *Q. suber* was recognized as ancestral. If none of the alleles were shared with *Q. robur* or *Q. suber*, the ancestral state was defined as "unknown" for the SNP. The ancestry state was determined for 98.34% of all SNPs.

Demographic History Analyses

To infer the demographic history of the 3 species, we conducted coalescent simulations in fastsimcoal2 v2.6.0.3 (Excoffier et al. 2013). The 2D joint unfolded SFS was calculated using ANGSD v0.920 (Korneliussen et al. 2014) based on intergenic sites, which are least affected by selection. Eight isolation-with-continuous-gene flow models were evaluated (supplementary fig. S17 and table S14, Supplementary Material online). Each model contained 5 current populations, representing the 2 genetic groups of *Q. variabilis*, 2 groups of *Q. acutissima* and *Q. chenii*. These models considered different scenarios of changes in effective population size and divergence processes within *Q. variabilis* and *Q. acutissima* (supplementary fig. S17, Supplementary Material online). According to Liang et al. (2021), all models treated *Q. variabilis* and *Q. acutissima* as sister species, with continuous gene flow between populations along the divergence processes. For each model, the global maximum likelihood (ML) parameter estimates were obtained from 60 independent runs with a minimum of 100,000 coalescent simulations as well as 40 iterations of the likelihood maximization algorithm. The best model was chosen based on Akaike's weight of evidence following Excoffier et al. (2013).

To estimate the confidence intervals of the parameters of the best-fit model, we performed 100 parametric bootstrap replicates and optimized model parameters with 50 independent runs for each simulated dataset. To assess the goodness-of-fit of the best-fit model, we calculated the expected SFS and summary statistics (π, F_{ST}, and *ZnS*) based on 100,000 coalescent simulations. The estimated summary statistics and SFSs based on simulated data were consistent with those based on observed data, suggesting that the best-fit model accurately captured aspects of demographic history (supplementary fig. S18, Supplementary Material online).

To convert demographic parameter estimates into absolute values, we estimated the mutation rate in oak species by using the formula $\mu = d/2T$ (Nei 1987), where d and T represent the genetic divergence and divergence time between *Q. variabilis* (representing the subgenus *Cerris*) and *Q. robur* (representing the subgenus *Quercus*), respectively. The genomes of *Q. robur* (Plomion et al. 2018) and *Q. variabilis* were aligned using lastZ v1.04.15 (https://lastz.github.io/lastz/), and genetic divergence (d) was estimated as 0.07 based on intergenic sites. By using a divergence time of 52 million years ago (Ma) between the subgenera *Quercus* and *Cerris* (Zhou et al. 2022b) and a generation time of 100 years for oak species (Cavender-Bares et al. 2011), the mutation rate was estimated as 6.75×10^{-8} per site per generation. Notably, the mutation rate and generation time of forest trees are difficult to estimate in their nature habitat; thus inferred population demographic parameters may require adjustments as these estimates change.

Detection of Genome-Wide Signatures of Local Adaptation

We applied 3 complementary methods to identify candidate SNPs and SVs associated with local adaptation in *Q. variabilis* and *Q. acutissima*, separately. First, we performed GEA analyses to identify SNPs and SVs that were significantly associated with climate variables by using latent factor mixed models (LFMMs) (Frichot et al. 2013), implemented in the R package lfmm2 (Caye et al. 2019). Because there were 2 genetic groups detected in each species (see Results), we set $K = 2$ to account for the confounding effects of population structure in LFMM analyses. We downloaded 19 environmental variables at 30 s resolution from worldclim1 (http://worldclim.com/version1). The importance ranking of the 19 variables was estimated by using a gradient forest analysis implemented in the R package gradientForest (Ellis et al. 2012). After comprehensive consideration of the importance ranking and correlations between the 19 environmental variables, the 6 top-ranked environmental variables (Bio2, Bio3, Bio4, Bio5, Bio15, and Bio18) with Spearman correlation coefficients $|\rho| < 0.65$ were retained for GEA analyses and RDA (supplementary figs. S19 and S20, Supplementary Material online).

Second, we used a multivariate approach, RDA, to evaluate associations between genetic and environmental factors by using the R package vegan (Van den Wollenberg 1977; Capblancq and Forester 2021). Only the first 3 RDA axes were considered, which jointly explained 60.91% and 61.73% of environmental variation in *Q. variabilis* and *Q. acutissima*, respectively. Outlier SNPs and SVs were identified as those loaded in the tail of the distribution (mean ± 4 standard deviations).

Third, we scanned the genome for loci that were strongly related to population structure using PCAdapt (Duforet-Frebourg et al. 2014). This method does not need any prior categorization of individuals into populations. Because the first principal component (PC1) clearly divided each species into 2 distinct genetic groups (supplementary fig. S8, Supplementary Material online), we looked for SNPs and SVs strongly associated with PC1. For both LFMM and PCAdapt, the Benjamini–Hochberg FDR method was used to correct for multiple testing (Benjamini and Hochberg 1995). SNPs and SVs with FDR < 0.01 were recognized as outliers.

We also scanned the genome for signatures of positive selection in each genetic groups of *Q. variabilis* and *Q. acutissima* using the cross-population composite likelihood ratio test (XP-CLR) (Chen et al. 2010). This method estimates the likelihood of allele frequency differentiation across multiple loci between 2 populations. The program XP-CLR (Chen et al. 2010) was run from each genetic group, using the other group as reference. For example, signals of selection in West group of *Q. variabilis* were tested using East group of this species, and vice versa. We calculated XP-CLR scores in 10 kb nonoverlapping windows containing at least 10 common SNPs (MAF > 0.05). To minimize the influence of dependency on the composite likelihood score, we used phased data and applied a down weighting strategy for pairs of SNPs exhibiting high LD ($r^2 > 0.95$). Windows with the top 1% of XP-CLR scores were considered as candidate regions under selection.

Functional Enrichment Analyses

To investigate whether any predicted functions were over represented for candidate genes, we performed enrichment analyses for GO items, Pfam entries, and KEGG pathways. Candidate genes were defined as those with candidate SNPs or SVs located within 10 kb up- and down-stream. To account

for potential sampling bias caused by different gene lengths, GO analyses were conducted using GOWINDA v1.12 (Kofler and Schlötterer 2012). Pfams and KEGG enrichment analyses were performed using clusterProfiler v4.6.0 (Wu et al. 2021). The significance of enrichment was tested by using 100,000 permutations, and multiple comparisons were corrected by using the Benjamini–Hochberg FDR method (Benjamini and Hochberg 1995). Functional categories with an FDR < 0.05 were considered significantly enriched. Additionally, we discarded any GO, Pfam, or KEGG terms supported by fewer than 5 genes to prevent misleading patterns caused by rare terms.

Molecular Parallelism Analyses

To investigate the molecular parallelism in *Chr9-ERF* region, we tested whether the same beneficial alleles were selected in parallel between *Q. variabilis* and *Q. acutissima*. To do that, SNPs were classified into "codirectional" and "antidirectional" categories following Wang et al. (2021). A SNP was defined as codirectional when the same allele showed a high or low frequency in the West groups of both *Q. variabilis* and *Q. acutissima*, while a SNP was defined as an antidirectional SNP when the same allele showed different trends in comparisons between species. We categorized 323 shared candidate SNPs in the *Chr9-ERF* region and 1,000 randomly selected SNPs across the genome. The randomly selected SNPs match the MAF of shared candidate SNPs (i.e., MAF = 0.07–0.36 and 0.10–0.49 in *Q. acutissima* and *Q. variabilis*, respectively). We further examined geographic clines in allele frequencies for the 323 shared-SNPs and 1,000 randomly selected SNPs using the R package HZAR v0.2.5 (Derryberry et al. 2014). For both *Q. variabilis* and *Q. acutissima*, the geographic axis started from the most southwest population (Kunming, KM), and other populations were ranked according the geographic distance from the KM population. Based on the AIC criterion, the best cline model was chosen from 15 models for each SNP (Derryberry et al. 2014).

We also examined the extent of molecular parallelism at the genic and functional levels. A gene was considered as a candidate for selection if any candidate SNPs were found within 10 kb upstream or downstream. To assess the degree of parallelism at the level of gene function, we annotate these candidate genes using GO terms and quantified the overlap in functional enrichment between species. In total, 142 and 424 candidate genes were identified in *Q. acutissima* and *Q. variabilis*, respectively. These genes were further annotated to 1,366 and 2,616 GO terms in *Q. acutissima* and *Q. variabilis*, respectively.

Estimating the Probability of Repeat De Novo Mutation

The probability of a repeat de novo mutation occurring in 2 species is estimated as: $\text{Prob} = (\mu \times T)^2$, where T is the divergence time and μ is the mutation rate. Given a divergence time of 13.83 million years between *Q. acutissima* and *Q. variabilis*, and a mutation rate of 6.75×10^{-10} per site per year for these 2 species, the probability of a repeat de novo mutation at a given site is $0.00934 \times 0.00934 = 8.72 \times 10^{-5}$. The probability of 323 de novo mutations occurring in a 250 kb region can be estimated as: $C\ (2.5 \times 10^5,\ 323) \times (8.72 \times 10^{-5})^{323} = 2.75 \times 10^{-240}$.

Testing for Introgression in the Chr9-ERF Region

To investigate gene flow between species, phylogenetic analyses, local ancestry inferences, and ABBA–BABA tests were performed. The phylogenetic analyses were conducted using RAxML v8.2.12 (Stamatakis 2014) under the GTRGAMMA model. All genotypes were phased using Beagle v4.1 (Browning and Browning 2013) with default settings, and a ML tree was constructed for phased haplotypes based on common SNPs (MAF ≥ 0.05). The best-scoring ML tree was selected from 1,000 ML trees, and the topological confidence of the best tree was evaluated based on 1,000 nonparametric bootstrap replicates. *Q. robur* was used as an outgroup species for phylogenetic analyses. The topology of the ML tree based on the *Chr9-ERF* region was compared with that of the tree based on genome-wide SNPs, and the conflict between gene trees was considered a signal of gene flow.

Introgression in the *Chr9-ERF* region was further validated by estimating the local ancestry of 2 haplotype groups ("Qv-W" and Qa-W) using the ELAI method (Guan 2014). Following Guan (2014), 4-way admixture analyses were conducted using 20 lower-layer clusters, 20 expectation maximization (EM) iterations and 50 admixture generations for each optimization run. Three independent runs were carried out with random seeds and results were averaged across runs. The local allele dosages of Qv-W haplotypes were estimated by using Qa-E, Qa-W, Qc, and Qv-E haplotypes as 4 sources.

ABBA–BABA tests were conducted to estimate 2 statistics, f_d (Malinsky et al. 2015) and d_f (Pfeifer and Kapan 2019) in 10 kb nonoverlapping windows. By integrating information for both f_d and genetic distance, the d_f statistic is robust to the time of gene flow (Pfeifer and Kapan 2019). In addition, both positive f_d and d_f values indicate gene flow between P3 and P2, whereas negative d_f points toward gene flow between P3 and P1 and negative f_d is meaningless. The f_d statistic was estimated using the pipeline developed by Simon Martin (https://github.com/simonhmartin/genomics_general) and the d_f statistic was estimated using PopGenome v2.7.1 (Pfeifer and Kapan 2019). To test the significance of gene flow, Z-scores and P-values were estimated for windows with at least 50 SNPs and corrections for multiple comparisons were performed using the Benjamini–Hochberg FDR method (Benjamini and Hochberg 1995). Because Qv-W and Qa-W haplotypes were each found in 2 different genetic groups of *Q. acutissima* and *Q. variabilis* (Fig. 6 and supplementary S13, Supplementary Material online), they were each divided into 2 subgroups. The Qv-W haplotypes from the West and East genetic groups of *Q. variabilis* were divided into Qv-W1 and Qv-W2 subgroups, respectively; the Qa-W haplotypes from West and East genetic groups of *Q. acutissima* were divided into Qa-W1 and Qa-W2 subgroups, respectively. ELAI and ABBA–BABA analyses were conducted based on 4 different combinations of haplotype subgroups. The results based on the combination of Qv-W1 and Qa-W1 are presented in the main text and Fig. 6, while the results for other combinations are presented in supplementary figs. S14–S16 and table S13, Supplementary Material online.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank T. Mitchell-Olds, P. Tiffin, V. Sork, the anonymous reviewer, and the editor for their insightful comments, which have significantly improved this manuscript. We thank S. Yuan, Z.-C. Jin, Q.-Q. An, Y.-R. Fan, X.-F. Yang, J. Yang, L. Zhao, and N. Wang for their contributions to sample collections and preparations. This work was supported by the Guangdong Flagship Project of Basic and Applied Basic Research (2023B0303050001), National Natural Science Foundation of China (NSFC 32301617 and 32161123003), and China Postdoctoral Science Foundation (2023000082).

Author Contributions

Y.-Y.L. and B.W. designed the project. Y.-Y.L., B.-F.Z., S.Y., X.-Y.C., Z.S., J.-S.W., S.-J. L., W.-J. L., L.-J.Q., J.-W.N., J.-F.L., and Y.-W.K. collected data. Y.-Y.L., H.L., Q.-Q.L., S.Y., W.Z., P.K.I., Y.-L.G., and B.W. analyzed and interpreted the data. Y.-Y.L. and B.W. wrote the paper. All authors read and approved the paper.

Conflict of Interest

The authors declare no competing interests.

Data Availability

The pan-genome data used in our analyses have been deposited at the National Genomics Data Center under the project PRJCA027131. All re-sequencing data have been submitted to NCBI under the project PRJNA1033777. The SNPs and SVs called in this study have been deposited in DRYAD (https://doi.org/10.5061/dryad.kd51c5bgv). Custom scripts used for data analyses have been deposited in DRYAD (https://doi.org/10.5061/dryad.kd51c5bgv).

References

Agarwal PK, Gupta K, Lopato S, Agarwal P. Dehydration responsive element binding transcription factors and their applications for the engineering of stress tolerance. *J Exp Bot*. 2017:68(9):2135–2148. doi:10.1093/jxb/erx118.

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. 2022:23(1):258. doi:10.1186/s13059-022-02823-7.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990:215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 2020:36(7):2251–2252. doi:10.1093/bioinformatics/btz859.

Arendt J, Reznick D. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol Evol*. 2008:23(1):26–32. doi:10.1016/j.tree.2007.09.011.

Barrett RD, Schluter D. Adaptation from standing genetic variation. *Trends Ecol Evol*. 2008:23(1):38–44. doi:10.1016/j.tree.2007.09.008.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995:57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999:27(2):573–580. doi:10.1093/nar/27.2.573.

Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. 2021:49(D1):D344–D354. doi:10.1093/nar/gkaa977.

Bock DG, Cai Z, Elphinstone C, González-Segovia E, Hirabayashi K, Huang K, Keais GL, Kim A, Owens GL, Rieseberg LH. Genomics of plant speciation. *Plant Commun*. 2023:4(5):100599. doi:10.1016/j.xplc.2023.100599.

Bohutínská M, Vlček J, Yair S, Laenen B, Konečná V, Fracassetti M, Slotte T, Kolář F. Genomic basis of parallel adaptation varies with divergence in Arabidopsis and its relatives. *Proc Natl Acad Sci U S A*. 2021:118(21):e2022713118. doi:10.1073/pnas.2022713118.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014:30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE. (Non)parallel evolution. *Annu Rev Ecol Evol Syst*. 2018:49(1):303–330. doi:10.1146/annurev-ecolsys-110617-062240.

Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013:194(2):459–471. doi:10.1534/genetics.113.150029.

Burge DO, Parker VT, Mulligan M, Sork VL. Influence of a climatic gradient on genetic exchange between two oak species. *Am J Bot*. 2019:106(6):864–878. doi:10.1002/ajb2.1315.

Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*. 2019:10(1):3240. doi:10.1038/s41467-019-11146-4.

Canestrelli D, Porretta D, Lowe WH, Bisconti R, Carere C, Nascetti G. The tangled evolutionary legacies of range expansion and hybridization. *Trends Ecol Evol*. 2016:31(9):677–688. doi:10.1016/j.tree.2016.06.010.

Cannon CH, Petit RJ. The oak syngameon: more than the sum of its parts. *New Phytol*. 2019:226(4):978–983. doi:10.1111/nph.16091.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol*. 2021:38(12):5825–5829. doi:10.1093/molbev/msab293.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 2008:18(1):188–196. doi:10.1101/gr.6743907.

Capblancq T, Forester BR. Redundancy analysis: a Swiss army knife for landscape genomics. *Methods Ecol Evol*. 2021:12(12):2298–2309. doi:10.1111/2041-210X.13722.

Cavender-Bares J. Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. *New Phytol*. 2019:221(2):669–692. doi:10.1111/nph.15450.

Cavender-Bares J, Gonzalez-Rodriguez A, Pahlich A, Koehler K, Deacon N. Phylogeography and climatic niche evolution in live oaks (*Quercus series* Virentes) from the tropics to the temperate zone. *J Biogeogr*. 2011:38(5):962–981. doi:10.1111/j.1365-2699.2010.02451.x.

Caye K, Jumentier B, Lepeule J, François O. LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Mol Biol Evol*. 2019:36(4):852–860. doi:10.1093/molbev/msz008.

Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol*. 2019:1962:1–14. doi:10.1007/978-1-4939-9173-0_1.

Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome Res*. 2010:20(3):393–402. doi:10.1101/gr.100545.109.

Chen J, Liu Y, Liu M, Guo W, Wang Y, He Q, Chen W, Liao Y, Zhang W, Gao Y, *et al*. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat Genet*. 2023a:55(12):2243–2254. doi:10.1038/s41588-023-01571-z.

Chen S, Wang P, Kong W, Chai K, Zhang S, Yu J, Wang Y, Jiang M, Lei W, Chen X, *et al*. Gene mining and genomics-assisted breeding

empowered by the pangenome of tea plant *Camellia sinensis*. *Nat Plants*. 2023b:9(12):1986–1999. doi:10.1038/s41477-023-01565-z.
Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. 2017:89(4):789–804. doi:10.1111/tpj.13415.
Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021a:18(2):170–175. doi:10.1038/s41592-020-01056-5.
Cheng Y, Miller MJ, Zhang D, Xiong Y, Hao Y, Jia C, Cai T, Li S-H, Johansson US, Liu Y, *et al*. Parallel genomic responses to historical climate change and high elevation in East Asian songbirds. *Proc Natl Acad Sci U S A*. 2021b:118(50):e2023918118. doi:10.1073/pnas.2023918118.
Chevin LM, Martin G, Lenormand T. Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. *Evolution*. 2010:64(11):3213–3231. doi:10.1111/j.1558-5646.2010.01058.x.
Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN. How the pan-genome is changing crop genomics and improvement. *Genome Biol*. 2021:22(1):3. doi:10.1186/s13059-020-02224-8.
Denk T, Grimm GW, Manos PS, Deng M, Hipp AL. An updated infrageneric classification of the oaks: review of previous taxonomic schemes and synthesis of evolutionary patterns. In: Gil-Pelegrín E, Peguero-Pina JJ, Sancho-Knapik D, editors. *Oaks physiological ecology. Exploring the functional diversity of genus Quercus L*. Cham: Springer International Publishing; 2017. p. 13–38.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011:43(5):491–498. doi:10.1038/ng.806.
Derryberry EP, Derryberry GE, Maley JM, Brumfield RT. HZAR: hybrid zone analysis using an R software package. *Mol Ecol Resour*. 2014:14(3):652–663. doi:10.1111/1755-0998.12209.
Du ZZ, He JB, Jiao WB. SynDiv: an efficient tool for chromosome collinearity-based population genomics analyses. *Plant Commun*. 2024:5(12):101071. doi:10.1016/j.xplc.2024.101071.
Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017:356(6333):92–95. doi:10.1126/science.aal3327.
Duforet-Frebourg N, Bazin E, Blum MGB. Genome scans for detecting footprints of local adaptation using a Bayesian Factor Model. *Mol Biol Evol*. 2014:31(9):2483–2495. doi:10.1093/molbev/msu182.
Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. 2016a:3(1):99–101. doi:10.1016/j.cels.2015.07.012.
Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016b:3(1):95–98. doi:10.1016/j.cels.2016.07.002.
Ellis N, Smith SJ, Roland Pitcher C. Gradient forests: calculating importance gradients on physical predictors. *Ecology*. 2012:93(1):156–168. doi:10.1890/11-0252.1.
Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019:20(1):238. doi:10.1186/s13059-019-1832-y.
Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 2005:14(8):2611–2620. doi:10.1111/j.1365-294X.2005.02553.x.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013:9(10):e1003905. doi:10.1371/journal.pgen.1003905.
Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*. 2000:155(3):1405–1413. doi:10.1093/genetics/155.3.1405.
Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE, Aluja M. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in Rhagoletis. *Proc Natl Acad Sci U S A*. 2005:102(suppl_1):6573–6580. doi:10.1073/pnas.0502099102.
Feng J, Dan X, Cui Y, Gong Y, Peng M, Sang Y, Ingvarsson PK, Wang J. Integrating evolutionary genomics of forest trees to inform future tree breeding amid rapid climate change. *Plant Commun*. 2024:7(10):101044. doi:10.1016/j.xplc.2024.101044.
Flood PJ, Hancock AM. The genomic basis of adaptation in plants. *Curr Opin Plant Biol*. 2017:36:88–94. doi:10.1016/j.pbi.2017.02.003.
Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol*. 2013:30(7):1687–1699. doi:10.1093/molbev/mst063.
Fu R, Zhu Y, Liu Y, Feng Y, Lu RS, Li Y, Li P, Kremer A, Lascoux M, Chen J. Genome-wide analyses of introgression between two sympatric Asian oak species. *Nat Ecol Evol*. 2022:6(7):924–935. doi:10.1038/s41559-022-01754-7.
Gao J, Liu Z-L, Zhao W, Tomlinson KW, Xia S-W, Zeng Q-Y, Wang X-R, Chen J. Combined genotype and phenotype analyses reveal patterns of genomic adaptation to local environments in the subtropical oak *Quercus acutissima*. *J Syst Evol*. 2021:59(3):541–556. doi:10.1111/jse.12568.
Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, *et al*. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018:36(9):875–879. doi:10.1038/nbt.4227.
Goel M, Schneeberger K. Plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*. 2022:38(10):2922–2926. doi:10.1093/bioinformatics/btac196.
Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019:20(1):277. doi:10.1186/s13059-019-1911-0.
Guan YT. Detecting structure of haplotypes and local ancestry. *Genetics*. 2014:196(3):625–642. doi:10.1534/genetics.113.160697.
Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 2003:31(19):5654–5666. doi:10.1093/nar/gkg770.
Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol*. 2008:9(1):R7. doi:10.1186/gb-2008-9-1-r7.
Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P. Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proc Natl Acad Sci U S A*. 2021:118(35):e2102914118. doi:10.1073/pnas.2102914118.
Han J, Xie X, Zhang Y, Yu X, He G, Li Y, Yang G. Evolution of the DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN subfamily in green plants. *Plant Physiol*. 2022:190(1):421–440. doi:10.1093/plphys/kiac286.
Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. *Bioinformatics*. 2019:35(17):2907–2915. doi:10.1093/bioinformatics/btz041.
Heller D, Vingron M. SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*. 2021:36(22-23):5519–5521. doi:10.1093/bioinformatics/btaa1034.
Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol*. 2020:21(1):35. doi:10.1186/s13059-020-1941-7.
Hu H, Scheben A, Verpaalen B, Tirnaz S, Bayer PE, Hodel RGJ, Batley J, Soltis DE, Soltis PS, Edwards D. Amborella gene presence/absence variation is associated with abiotic stress responses that may

contribute to environmental adaptation. *New Phytol*. 2022:233(4):1548–1555. doi:10.1111/nph.17658.

Hu J, Wang Z, Liang F, Liu S-L, Ye K, Wang D-P. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. *Genomics Proteomics Bioinformatics*. 2024:22(1):qzad009. doi:10.1093/gpbjnl/qzad009.

Huang C, Zhang Y, Bartholomew B. Fagaceae. In: Wu ZY, Raven P, editors. *Flora of China*. Beijing and St Louis: Sciense Press and Missouri Botanical Garden Press; 1999. p. 314–400.

Huang J, Zhao X, Bürger M, Wang Y, Chory J. Two interacting ethylene response factors regulate heat stress response. *Plant Cell*. 2021:33(2):338–357. doi:10.1093/plcell/koaa026.

Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods*. 2022:19(6):705–710. doi:10.1038/s41592-022-01457-8.

James ME, Brodribb T, Wright IJ, Rieseberg LH, Ortiz-Barrientos D. Replicated evolution in plants. *Annu Rev Plant Biol*. 2023:74(1):697–725. doi:10.1146/annurev-arplant-071221-090809.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017:8:14061. doi:10.1038/ncomms14061.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020:21(1):189. doi:10.1186/s13059-020-02107-y.

Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, Liu W, Chen C, Song Y, Tan L, *et al*. The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat Commun*. 2023:14(1):6259. doi:10.1038/s41467-023-42029-4.

Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 2011:27(6):757–763. doi:10.1093/bioinformatics/btr010.

Kelly JK. A test of neutrality based on interlocus associations. *Genetics*. 1997:146(3):1197–1206. doi:10.1093/genetics/146.3.1197.

Kim BY, Wei X, Fitz-Gibbon S, Lohmueller KE, Ortego J, Gugger PF, Sork VL. RADseq data reveal ancient, but not pervasive, introgression between Californian tree and scrub oak species (*Quercus* sect. *Quercus*: Fagaceae). *Mol Ecol*. 2018:27(22):4556–4571. doi:10.1111/mec.14869.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015:12(4):357–360. doi:10.1038/nmeth.3317.

Kofler R, Schlötterer C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*. 2012:28(15):2084–2085. doi:10.1093/bioinformatics/bts315.

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

Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generationsequencing data. *BMC Bioinforma*. 2014:15(1):356. doi:10.1186/s12859-014-0356-4.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019:20(1):117. doi:10.1186/s13059-019-1720-5.

Kremer A, Hipp AL. Oaks: an evolutionary success story. *New Phytol*. 2020:226(4):987–1011. doi:10.1111/nph.16274.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018:35(6):1547–1549. doi:10.1093/molbev/msy096.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol*. 2004:5(2):R12. doi:10.1186/gb-2004-5-2-r12.

Lee KM, Coop G. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics*. 2017:207(4):1591–1619. doi:10.1534/genetics.117.300417.

Lee KM, Coop G. Population genomics perspectives on convergent adaptation. *Philos Trans R Soc Lond B Biol Sci*. 2019:374(1777):20180236. doi:10.1098/rstb.2018.0236.

Leroy T, Louvet JM, Lalanne C, Le Provost G, Labadie K, Aury JM, Delzon S, Plomion C, Kremer A. Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytol*. 2020a:226(4):1171–1182. doi:10.1111/nph.16095.

Leroy T, Rougemont Q, Dupouey JL, Bodénès C, Lalanne C, Belser C, Labadie K, Le Provost G, Aury JM, Kremer A, *et al*. Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytol*. 2020b:226(4):1183–1197. doi:10.1111/nph.16039.

Li C, Jiang X, Gong W, Li D, Li C. Surface uplift of the Central Yunnan Plateau since the Pliocene. *Geol J*. 2018:53(1):386–396. doi:10.1002/gj.2904.

Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018:34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010:26(5):589–595. doi:10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009:25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Li J, Lee CR. The role of gene presence-absence variations on genetic incompatibility in Asian rice. *New Phytol*. 2023:239(2):778–791. doi:10.1111/nph.18969.

Liang YY, Shi Y, Yuan S, Zhou BF, Chen XY, An QQ, Ingvarsson PK, Plomion C, Wang B. Linked selection shapes the landscape of genomic variation in three oak species. *New Phytol*. 2021:233(1):555–568. doi:10.1111/nph.17793.

Liu JN, Yan L, Chai Z, Liang Q, Dong Y, Wang C, Li X, Li C, Mu Y, Gong A, *et al*. Pan-genome analyses of 11 *Fraxinus* species provide insights into salt adaptation in ash trees. *Plant Commun*. 2024:6(1):101137. doi:10.1016/j.xplc.2024.101137.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, *et al*. Pan-genome of wild and cultivated soybeans. *Cell*. 2020:182(1):162–176.e13. doi:10.1016/j.cell.2020.05.023.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 2005:33(20):6494–6506. doi:10.1093/nar/gki937.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. Genomic islands of speciation separate cichlid ecomorphs in an east African crater lake. *Science*. 2015:350(6267):1493–1498. doi:10.1126/science.aac9927.

Martins K, Gugger PF, Llanderal-Mendoza J, González-Rodríguez A, Fitz-Gibbon ST, Zhao JL, Rodríguez-Correa H, Oyama K, Sork VL. Landscape genomics provides evidence of climate-associated genetic variation in Mexican populations of *Quercus rugosa*. *Evol Appl*. 2018:11(10):1842–1858. doi:10.1111/eva.12684.

Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*. 2018:210(2):719–731. doi:10.1534/genetics.118.301336.

Montejo-Kovacevich G, Meier JI, Bacquet CN, Warren IA, Chan YF, Kucka M, Salazar C, Rueda-M N, Montgomery SH, McMillan WO, *et al*. Repeated genetic adaptation to altitude in two tropical butterflies. *Nat Commun*. 2022:13(1):4676. doi:10.1038/s41467-022-32316-x.

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. The genomic consequences of hybridization. *Elife*. 2021:10:e69016. doi:10.7554/eLife.69016.

Nagamitsu T, Uchiyama K, Izuno A, Shimizu H. Common-garden study of introgression at loci associated with traits adaptive to coastal environment from *Quercus dentata* into *Quercus mongolica* var. *crispula*. *Plant Species Biol*. 2024:39(5):231–248. doi:10.1111/1442-1984.12472.

Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 2016:32(19):3021–3023. doi:10.1093/bioinformatics/btw369.

Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009:25(10):1335–1337. doi:10.1093/bioinformatics/btp157.

Nei M. *Molecular evolutionary genetics*. New York: Columbia University Press; 1987.

Nocchi G, Whiting JR, Yeaman S. Repeated global adaptation across plant species. *Proc Natl Acad Sci U S A*. 2024:121(52):e2406832121. doi:10.1073/pnas.2406832121.

O'Donnell ST, Fitz-Gibbon ST, Sork VL. Ancient introgression between distantly related white oaks (*Quercus* sect. *Quercus*) shows evidence of climate-associated asymmetric gene exchange. *J Hered*. 2021:112(7):663–670. doi:10.1093/jhered/esab053.

Ortego J, Gugger PF, Sork VL. Impacts of human-induced environmental disturbances on hybridization between two ecologically differentiated Californian oak species. *New Phytol*. 2017:213(2):942–955. doi:10.1111/nph.14182.

Ortego J, Gugger PF, Sork VL. Genomic data reveal cryptic lineage diversification and introgression in Californian golden cup oaks (section *Protobalanus*). *New Phytol*. 2018:218(2):804–818. doi:10.1111/nph.14951.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019:20(1):275. doi:10.1186/s13059-019-1905-y.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015:33(3):290–295. doi:10.1038/nbt.3122.

Pfeifer B, Kapan DD. Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics*. 2019:20(1):207. doi:10.1186/s12859-019-2747-z.

Plomion C, Aury JM, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, Le Provost G, *et al*. Oak genome reveals facets of long lifespan. *Nat Plants*. 2018:4(7):440–452. doi:10.1038/s41477-018-0172-3.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, *et al*. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018:36(10):983–987. doi:10.1038/nbt.4235.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007:81(3):559–575. doi:10.1086/519795.

Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, He Q, Ou S, Zhang H, Li X, *et al*. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*. 2021:184(13):3542–3558.e16. doi:10.1016/j.cell.2021.04.046.

Ramos AM, Usié A, Barbosa P, Barros PM, Capote T, Chaves I, Simões F, Abreu I, Carrasquinho I, Faro C. The draft genome sequence of cork oak. *Sci Data*. 2018:5:180069. doi:10.1038/sdata.2018.69.

Rellstab C, Zoller S, Walthert L, Lesur I, Pluess AR, Graf R, Bodénès C, Sperisen C, Kremer A, Gugerli F. Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present and future climatic conditions. *Mol Ecol*. 2016:25(23):5907–5924. doi:10.1111/mec.13889.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020:21(1):245. doi:10.1186/s13059-020-02134-9.

Savolainen O, Lascoux M, Merilä J. Ecological genomics of local adaptation. *Nat Rev Genet*. 2013:14(11):807–820. doi:10.1038/nrg3522.

Savolainen O, Pyhäjärvi T, Knürr T. Gene flow and local adaptation in trees. *Annu Rev Ecol Evol Syst*. 2007:38(1):595–619. doi:10.1146/annurev.ecolsys.38.091206.095646.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*. 2018:15(6):461–468. doi:10.1038/s41592-018-0001-7.

Shi T, Zhang X, Hou Y, Jia C, Dan X, Zhang Y, Jiang Y, Lai Q, Feng J, Feng J, *et al*. The super-pangenome of Populus unveils genomic facets for its adaptation and diversification in widespread forest trees. *Mol Plant*. 2024:17(5):725–746. doi:10.1016/j.molp.2024.03.009.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015:31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang P-C, Carroll A, *et al*. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*. 2021:374(6574):abg8871. doi:10.1126/science.abg8871.

Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013:195(3):693–702. doi:10.1534/genetics.113.154138.

Song B, Ning W, Wei D, Jiang M, Zhu K, Wang X, Edwards D, Odeny DA, Cheng S. Plant genome resequencing and population genomics: current status and future prospects. *Mol Plant*. 2023:16(8):1252–1268. doi:10.1016/j.molp.2023.07.009.

Sork VL. Genomic studies of local adaptation in natural plant populations. *J Hered*. 2018:109(1):3–15. doi:10.1093/jhered/esx091.

Sork VL, Cokus SJ, Fitz-Gibbon ST, Zimin AV, Puiu D, Garcia JA, Gugger PF, Henriquez CL, Zhen Y, Lohmueller KE, *et al*. High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. *Nat Commun*. 2022:13(1):2047. doi:10.1038/s41467-022-29584-y.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014:30(9):1312–1313. doi:10.1093/bioinformatics/btu033.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*. 2008:24(5):637–644. doi:10.1093/bioinformatics/btn013.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989:123(3):585–595. doi:10.1093/genetics/123.3.585.

Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS, Zhang Y, Huang Y, Wang Y, Yim WC, *et al*. JCVI: a versatile toolkit for comparative genomics analysis. *Imeta*. 2024:3(4):e211. doi:10.1002/imt2.211.

Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2009:Chapter 4:4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s25.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*. 2006:313(5793):1596–1604. doi:10.1126/science.1128691.

van den Wollenberg AL. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*. 1977:42(2):207–219. doi:10.1007/BF02294050.

Wadgymar SM, DeMarche ML, Josephs EB, Sheth SN, Anderson JT. Local adaptation: causal agents of selection and adaptive trait divergence. *Annu Rev Ecol Evol Syst*. 2022:53(1):87–111. doi:10.1146/annurev-ecolsys-012722-035231.

Wang D, Wang H, Xu X, Wang M, Wang Y, Chen H, Ping F, Zhong H, Mu Z, Xie W, *et al*. Two complementary genes in a presence-absence variation contribute to *indica-japonica* reproductive isolation in rice. *Nat Commun*. 2023a:14(1):4531. doi:10.1038/s41467-023-40189-x.

Wang L, Josephs EB, Lee KM, Roberts LM, Rellán-Álvarez R, Ross-Ibarra J, Hufford MB. Molecular parallelism underlies convergent highland adaptation of maize landraces. *Mol Biol Evol*. 2021:38(9):3567–3580. doi:10.1093/molbev/msab119.

Wang L, Li L-L, Chen L, Zhang R-G, Zhao S-W, Yan H, Gao J, Chen X, Si Y-J, Chen Z, *et al*. Telomere-to-telomere and haplotype-resolved

genome assembly of the Chinese cork oak (*Quercus variabilis*). *Front Plant Sci*. 2023b:14:1290913. doi:10.3389/fpls.2023.1290913.

Wang Y, Wang Y, Cheng X, Ding Y, Wang C, Merilä J, Guo B. Prevalent Introgression underlies convergent evolution in the diversification of *Pungitius Sticklebacks*. *Mol Biol Evol*. 2023c:40(2):msad026. doi:10.1093/molbev/msad026.

Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984:38(6):1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x.

Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol*. 2018:33(6):427–440. doi:10.1016/j.tree.2018.04.002.

Wilkinson MJ, Roda F, Walter GM, James ME, Nipper R, Walsh J, Allen SL, North HL, Beveridge CA, Ortiz-Barrientos D. Adaptive divergence in shoot gravitropism creates hybrid sterility in an Australian wildflower. *Proc Natl Acad Sci U S A*. 2021:118(47):e2004901118. doi:10.1073/pnas.2004901118.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*. 2021:2(3):100141. doi:10.1016/j.xinn.2021.100141.

Xu G-C, Xu T-J, Zhu R, Zhang Y, Li S-Q, Wang H-W, Li J-T. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience*. 2019:8(1):giy157. doi:10.1093/gigascience/giy157.

Yan H, Sun M, Zhang Z, Jin Y, Zhang A, Lin C, Wu B, He M, Xu B, Wang J, *et al.* Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nat Genet*. 2023:55(3):507–518. doi:10.1038/s41588-023-01302-4.

Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A*. 2013:110(19):E1743–E1751. doi:10.1073/pnas.1219381110.

Yeaman S. Local adaptation by alleles of small effect. *Am Nat*. 2015:186(S1):S74–S89. doi:10.1086/682405.

Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski KA, Smets P, Wang T, Gray LK, *et al.* Convergent local adaptation to climate in distantly related conifers. *Science*. 2016:353(6306):1431–1433. doi:10.1126/science.aaf7812.

Yeaman S, Whitlock MC. The genetic architecture of adaptation under migration-selection balance. *Evolution*. 2011:65(7):1897–1911. doi:10.1111/j.1558-5646.2011.01269.x.

Yuan S, Shi Y, Zhou BF, Liang YY, Chen XY, An QQ, Fan YR, Shen Z, Ingvarsson PK, Wang B. Genomic vulnerability to climate change in *Quercus acutissima*, a dominant tree species in East Asian deciduous forests. *Mol Ecol*. 2023:32(7):1639–1655. doi:10.1111/mec.16843.

Zeng Y-F, Liao W-J, Petit RJ, Zhang D-Y. Geographic variation in the structure of oak hybrid zones provides insights into the dynamics of speciation. *Mol Ecol*. 2011:20(23):4995–5011. doi:10.1111/j.1365-294X.2011.05354.x.

Zhang Z. KaKs_Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinformatics*. 2022:20(3):536–540. doi:10.1016/j.gpb.2021.12.002.

Zhao Y. Vegetation and climate reconstructions on different time scales in China: a review of Chinese palynological research. *Veg Hist Archaeobot*. 2018:27(2):381–392. doi:10.1007/s00334-017-0655-6.

Zhisheng A, Kutzbach JE, Prell WL, Porter SC. Evolution of Asian monsoons and phased uplift of the Himalaya–Tibetan plateau since Late Miocene times. *Nature*. 2001:411(6833):62–66. doi:10.1038/35075035.

Zhou BF, Shi Y, Chen XY, Yuan S, Liang YY, Wang B. Linked selection, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence in *Quercus dentata*. *J Syst Evol*. 2022a:60(6):1344–1357. doi:10.1111/jse.12817.

Zhou BF, Yuan S, Crowl AA, Liang YY, Shi Y, Chen XY, An QQ, Kang M, Manos PS, Wang B. Phylogenomic analyses highlight innovation and introgression in the continental radiations of Fagaceae across the Northern Hemisphere. *Nat Commun*. 2022b:13(1):1320. doi:10.1038/s41467-022-28917-1.

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*. 2022c:606(7914):527–534. doi:10.1038/s41586-022-04808-9.