JIPB Journal of Integrative Plant Biology

Research Article
https://doi.org/10.1111/jipb.13773

# An integrative framework reveals widespread gene flow during the early radiation of oaks and relatives in Quercoideae (Fagaceae)∞

Shui-Yin Liu[1,2], Ying-Ying Yang[1], Qin Tian[1,2], Zhi-Yun Yang[1], Shu-Feng Li[3], Paul J. Valdes[4,5], Alex Farnsworth[4,5], Heather R. Kates[6], Carolina M. Siniscalchi[7], Robert P. Guralnick[6], Douglas E. Soltis[6,8], Pamela S. Soltis[6], Gregory W. Stull[1]*, Ryan A. Folk[9]* and Ting-Shuang Yi[1,2]*

1. Germplasm Bank of Wild Species, Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Chinese Academy of Sciences Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China
4. School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK
5. State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
6. Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA
7. Mississippi State University Libraries, Mississippi State University, Mississippi State, Mississippi 39762, USA
8. Department of Biology, University of Florida, Gainesville, Florida 32611, USA
9. Department of Biological Sciences, Mississippi State University, Mississippi State, Mississippi 39762, USA

*Correspondences: Gregory W. Stull (gwstull@gmail.com); Ryan A. Folk (rfolk@biology.msstate.edu); Ting-Shuang Yi (tingshuangyi@mail.kib.ac.cn; Dr. Yi is fully responsible for the distribution of all the materials associated with this article)

Shui-Yin Liu

Ting-Shuang Yi

## ABSTRACT

Although the frequency of ancient hybridization across the Tree of Life is greater than previously thought, little work has been devoted to uncovering the extent, timeline, and geographic and ecological context of ancient hybridization. Using an expansive new dataset of nuclear and chloroplast DNA sequences, we conducted a multifaceted phylogenomic investigation to identify ancient reticulation in the early evolution of oaks (*Quercus*). We document extensive nuclear gene tree and cytonuclear discordance among major lineages of *Quercus* and relatives in Quercoideae. Our analyses recovered clear signatures of gene flow against a backdrop of rampant incomplete lineage sorting, with gene flow most prevalent among major lineages of *Quercus* and relatives in Quercoideae during their initial radiation, dated to the Early-Middle Eocene. Ancestral reconstructions including fossils suggest ancestors of *Castanea* + *Castanopsis*, *Lithocarpus*, and the Old World oak clade probably co-occurred in North America and Eurasia, while the ancestors of *Chrysolepis, Notholithocarpus*, and the New World oak clade co-occurred in North America, offering ample opportunity for hybridization in each region. Our study shows that hybridization—perhaps in the form of ancient syngameons like those seen today—has been a common and important process throughout the evolutionary history of oaks and their relatives. Concomitantly, this study provides a methodological framework for detecting ancient hybridization in other groups.

Keywords: ancient reticulation, biogeography, cytonuclear discordance, Fagaceae, fossils, niche evolution, phylogenomics, *Quercus*

**Liu, S.Y., Yang, Y.Y., Tian, Q., Yang, Z.Y., Li, S.F., Valdes, P.J., Farnsworth, A., Kates, H.R., Siniscalchi, C.M., Guralnick, R.P., et al.** (2024). An integrative framework reveals widespread gene flow during the early radiation of oaks and relatives in Quercoideae (Fagaceae). J. Integr. Plant Biol. **00:** 1–23.

## INTRODUCTION

Hybridization, both recent and ancient, is an important evolutionary process (Mallet, 2007; Stull et al., 2023). Evidence of hybridization has been increasingly documented across the Tree of Life (TOL) (Mason et al., 2019; Stull et al., 2020; Suvorov et al., 2022). However, detecting hybridization and distinguishing it from other biological phenomena (e.g., incomplete lineage sorting, ILS) that yield similar patterns of phylogenomic conflict remains challenging. Detecting ancient hybridization is particularly difficult, given that anecdotal evidence for hybridization (traditionally from morphology, overlapping geographic ranges, and chromosome numbers) is usually lacking, and phylogenomic signal is eroded over time (Stull et al., 2023). The integration of phylogenomic analysis, divergence time estimation, and ancestral range and ecological niche estimations provides a promising and effective framework for discovering ancient hybridization and reconstructing scenarios of reticulate evolution (e.g., Folk et al., 2018; Pavón-Vázquez et al., 2021; Yu et al., 2022). However, the integration of paleoecological perspectives with modern molecular data has been underutilized in evaluating the plausibility of ancient hybridization predating the Miocene (Folk et al., 2023). This gap is particularly noticeable in rapidly radiating and species-rich groups, preventing a broader synthesis of the importance of gene flow across the TOL.

The oaks (*Quercus*), currently distributed from Eurasia, North Africa, North America, and Central America to Colombia, are a diverse clade (ca. 435 spp.) of woody plants with vast ecological and economic importance (Nixon, 2006; Denk et al., 2017). Oaks have long been known for their propensity to hybridize across various phylogenetic and spatial scales (Hardin, 1975; Whittemore and Schaal, 1991; Crowl et al., 2020), and gene flow has been highlighted as an evolutionary asset in this group (Petit et al., 2003; Dodd and Afzal-Rafii, 2007; Leroy et al., 2020a). Where sympatric, or partially sympatric, oak species form syngameons, but they are able to maintain relatively clear species boundaries in a morphological sense despite active gene flow (Burger, 1975; Cannon and Petit, 2020; Leroy et al., 2020b). Hybridization among closely related oak species is well documented, such as hybridization within sections *Cyclobalanopsis* (An et al., 2017), *Ilex* (Feng et al., 2016), *Lobatae* (Moran et al., 2012), *Quercus* (Curtu et al., 2007), and *Virentes* (Eaton et al., 2015). Hybridization between species from different sections has also been reported, such as between sections *Cyclobalanopsis* and *Ilex* (Simeone et al., 2016), and between sections *Ponticae* and *Quercus* (Crowl et al., 2020).

Less attention has been paid to assessing the extent and biological significance of ancient hybridization during the early evolution of *Quercus* and relatives within the broader context of Fagaceae phylogeny (e.g., Zhou et al., 2022). While the broad intrageneric relationships within *Quercus* appear well-resolved, over 40% of oak species have not been included in previous phylogenetic studies, with subg. *Cerris* being particularly underrepresented (see Hipp et al., 2020). Additionally, high levels of gene tree discordance have been documented along the phylogenetic backbone, possibly stemming from a combination of ancient gene flow (Crowl et al., 2020; Hipp et al., 2020) and ILS (Zhou et al., 2022). Potential for ancient gene flow is further supported by extensive levels of deep discordance between chloroplast and nuclear phylogenies (i.e., cytonuclear discordance). In contrast to nuclear phylogenies that show *Quercus* as monophyletic, chloroplast phylogenies typically show *Quercus* as non-monophyletic, with oak sections showing various relationships with other genera of Fagaceae subfamily Quercoideae, which includes *Castanea*, *Castanopsis*, *Chrysolepis*, *Lithocarpus*, *Notholithocarpus*, and *Trigonobalanus* (Simeone et al., 2016; Yang et al., 2021; Zhou et al., 2022). Beyond observations of gene tree conflict, the relative contributions of gene flow and ILS to gene tree discordance in quantitative terms remain undocumented. Likewise, no synthetic examinations of the rich fossil record for Fagaceae have been conducted in combination with phylogenetic studies to clarify the temporal, geographic, and ecological contexts of putative ancient reticulation events. Hypotheses have been proposed for gene flow scenarios (e.g., Zhou et al., 2022), but quantitative tests of these hypotheses are needed to elucidate the conditions that promote hybridization as an evolutionary mechanism.

Here we present an integrative phylogenomic and paleobotanical study for reconstructing detailed scenarios of ancient hybridization during the early radiation of oaks. We aim to establish a framework useful not only for understanding the timing and location of deep reticulation in oaks, but also for addressing questions about the conditions in Earth history that may have promoted reticulation across the TOL more generally. To accomplish this goal, we generated a newly constructed phylogenomic dataset using both transcriptome and hybrid enrichment sequencing (Hyb-Seq) and including extensive nuclear and chloroplast loci across over 400 species of Fagaceae. Specifically, we first infer phylogenetic relationships of oaks and relatives and dissect the phylogenetic conflict among nuclear genes and between the nuclear and chloroplast genomes. Second, we identify the phylogenetic locations and temporal windows of ancient hybridization events between the ancestors of major *Quercus* lineages and those of other genera of Quercoideae. Third, we expand upon previous studies to evaluate the biogeographic and ecological plausibility of ancient hybridization in Quercoideae, while also integrating paleogeographic and paleoclimatic information from the extensive fossil record of Quercoideae.

## RESULTS

### Assembly and alignment

We recovered a mean number of 1,733,167 Hyb-Seq reads from the 100-locus NitFix panel (Kates et al., 2024; *SD* = 1,964,004) per sample with an average of 68% (*SD* = 12%) of reads on-target. Sequencing coverage across all samples ranged from 12× to 5,143× (mean = 977×). Two Hyb-Seq datasets were generated: the HYB-89MO dataset included 431 taxa with 89 orthologous loci inferred using the monophyletic outgroup (MO) approach; and the HYB-98RT dataset included 431 taxa with 98 orthologous loci inferred using the rooted ingroup (RT) approach. A mean number of 75 (*SD* = 15) and 78 (*SD* = 16) nuclear genes per taxon was recovered in the HYB-89MO and HYB-98RT datasets, respectively. The concatenated matrices of the HYB-89MO and HYB-98RT datasets had an aligned length of 96,122 bp and 105,564 bp, respectively (Table S1). These results indicated a good performance of the bait set, which was designed as universal for the rosid clade.

To complement the Hyb-Seq datasets, two transcriptome datasets were also generated using the same approach of orthology inference: the RNA-977MO dataset included 89 taxa and 977 MO orthologous loci; and the RNA-2821RT dataset included 89 taxa and 2,821 RT orthologous loci. The concatenated matrices of the RNA-977MO and RNA-2821RT datasets had an aligned length of 1,229,405 bp and 6,050,182 bp, respectively.

A chloroplast dataset was generated from a combination of Hyb-Seq, transcriptome, and GenBank data, and included 223 taxa after data cleaning. The concatenated plastome matrix had an aligned length of 135,243 bp (with one of the inverted repeats removed).

### Phylogenetic analyses and topological conflicts

Phylogenetic analyses were conducted based on two Hyb-Seq and two transcriptome datasets using maximum likelihood (ML) and coalescent-based methods. The phylogenetic results recovered largely the same, well-resolved topology of eight monophyletic genera in Fagaceae (subfamily Fagoideae comprising *Fagus*, and subfamily Quercoideae comprising *Castanea*, *Castanopsis*, *Chrysolepis*, *Lithocarpus*, *Notholithocarpus*, *Quercus*, and *Trigonobalanus*), and eight monophyletic sections in oaks (subg. *Cerris* comprising sections *Cerris*, *Cyclobalanopsis*, and *Ilex*, and subg. *Quercus* comprising sections *Lobatae*, *Ponticae*, *Protobalanus*, *Quercus*, and *Virentes*; Figures 1, S1–S8). However, genus- and section-level conflicts were recovered with low bootstrap support. These include the placement of *Chrysolepis*, which was weakly supported as sister to *Lithocarpus* (bootstrap value (BS) = 30), and sect. *Cyclobalanopsis*, which was placed as sister to *Notholithocarpus* (BS = 25) in the concatenated ML tree of the HYB-98RT dataset (Figure S4). Some well-supported relationships were recovered among subclades within oak sections such as sect. *Ilex*, and especially sect. *Cyclobalanopsis*. We recovered *Q. gaharuensis* as sister to the rest of sect. *Cyclobalanopsis* with full or high support (BS = 100; local posterior probability (LPP) = 0.92; Figures S1–S4).

Analyses employing phyparts and quartet sampling (QS) were performed to explore the topological concordance among different nuclear genes in the phylogeny of Fagaceae. The results suggested strong support for a single topology in several areas of the phylogeny (i.e., stem groups of *Fagus* and *Trigonobalanus*; and crown groups of *Castanea* + *Castanopsis* and a clade of all genera of Quercoideae except *Trigonobalanus*; Figures 1, S9). However, a high level of gene tree discordance was observed within the clade of *Chrysolepis*, *Lithocarpus*, *Notholithocarpus*, and *Quercus* (Figures S9–S11).

In comparison with nuclear trees, the chloroplast ML tree revealed significantly different relationships among genera in Quercoideae and among sections in *Quercus* (Figures 2A, S12). The chloroplast tree did not support the monophyly of *Chrysolepis*, *Quercus*, sect. *Cyclobalanopsis*, sect. *Ilex*, sect. *Ponticae*, sect. *Protobalanus*, sect. *Quercus*, or sect. *Virentes*. The chloroplast relationships largely track geography rather than evolutionary history per se as revealed by nuclear genes. For instance, *Lithocarpus* and the clade *Castanea* + *Castanopsis* were recovered as sisters (BS = 66), with these together nesting in subg. *Cerris*. All of these taxa (with the exception of several *Castanea* species) are presently restricted to the Old World. The clade containing the small genera *Notholithocarpus* and *Chrysolepis* from western North America was sister (BS = 72) to the clade composed of North American species in sect. *Protobalanus* and sect. *Quercus* s.l.; sect. *Quercus* s.l. refers to a clade of (sect. *Ponticae*, (sect. *Virentes*, sect. *Quercus*)).

### Divergence time estimation

Dating analyses based on chloroplast and nuclear DNA sequences were performed using treePL and showed that the divergence times among *Quercus* and other genera in Quercoideae were systematically younger in the chloroplast tree than the nuclear tree, a result that was robust to two alternative calibration scenarios (with two and four calibrations; Figure 2B). A nuclear maximum clade credibility (MCC) tree from the BEAST analysis included a more elaborate set of nine calibrations, recovering older divergence times than both the nuclear and chloroplast treePL analyses. The inferred divergence times of four key nodes from the dated chloroplast tree (four calibrations in treePL), the dated nuclear tree (four calibrations in treePL), and the dated nuclear MCC tree (nine calibrations in BEAST), respectively, are as follows: the split of (*Castanopsis* + *Castanea*) and subg. *Cerris* (47.0573, 54.1271, 64.2473 Ma); the split of *Lithocarpus* and subg. *Cerris* (47.0573, 51.9321, 59.7094 Ma), the split of *Chrysolepis* and subg. *Quercus* (46.9273, 51.1016, 57.1719 Ma), and the split of *Notholithocarpus* and subg. *Quercus* (46.9273, 49.9004, 55.3428 Ma).

### Phylogenetic signal of alternative topologies for five uncertain nodes

We quantified the distribution of the phylogenetic signal to explore why different phylogenomic datasets or different analyses of the same dataset yielded conflicting topologies, focusing on five deep nodes of Fagaceae that were identified in our results and in previous studies (Manos et al., 1999, 2001, 2008; Hubert et al., 2014). These nodes concern the phylogenetic positions of (1) *Castanea* + *Castanopsis*, (2) *Lithocarpus*, (3) *Chrysolepis*, (4)

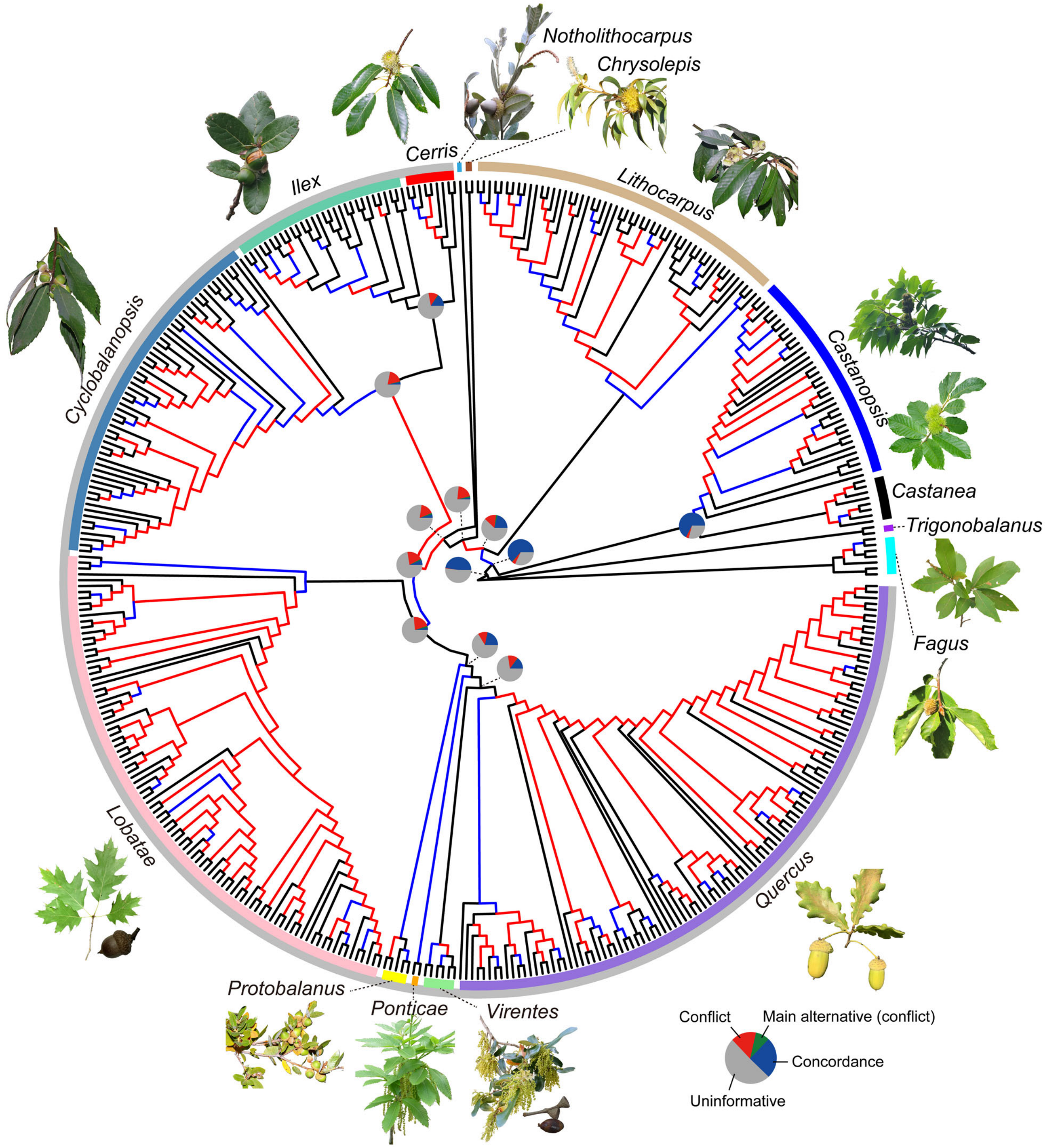


**Figure 1. Cladogram of the species tree of oaks and relatives inferred by ASTRAL-III based on the HYB-98RT dataset**

Tip labels are shown in Figure S2. Branches showing consistent relationships between ASTRAL-III and RAxML and for all four nuclear datasets (HYB-98RT, HYB-89MO, RNA-2821RT, and RNA-977MO) are colored black (local posterior probability (LPP) ≥ 0.95 or bootstrap value (BS) ≥ 95% in all eight species trees) and blue (LPP < 0.95 or BS < 95% in any one of the eight trees). Branches showing conflicting relationships among any of the eight trees are colored red. Pie charts for major clades show the phyparts results based on 2,821 nuclear gene trees (RNA-2821RT); see Figure S10–S11 for phyparts results of full taxa sets. Fagaceae genera (outer circle) and oak sections (inner circle) are indicated by colored bars, and their pictured representatives are: *Fagus grandifolia* by Bruce Kirchoff, *Chrysolepis chrysophylla* by J. Maughn, *Notholithocarpus densiflorus* by theforestprimeval, *Quercus chrysolepis* by copepodo, and *Q. pontica* by peganum from https://search.creativecommons.org/; *Trigonobalanus doichangensis* by Li Chen (with permission); *Q. rubra* and *Q. robur* by Yingying Yang; others by Shuiyin Liu.

*Notholithocarpus*, and (5) *Quercus* sect. *Cyclobalanopsis* (see Figures S1–S8, S13 and Table S2 for discordant topologies). Polytomy tests were first conducted across datasets and suggested rejecting the null hypothesis that each of these five uncertain nodes should be replaced by a polytomy ($P < 0.05$) except for *Chrysolepis* and sect. *Cyclobalanopsis* in the HYB-98RT dataset (Table S2). Approximately unbiased (AU) tests were used to examine

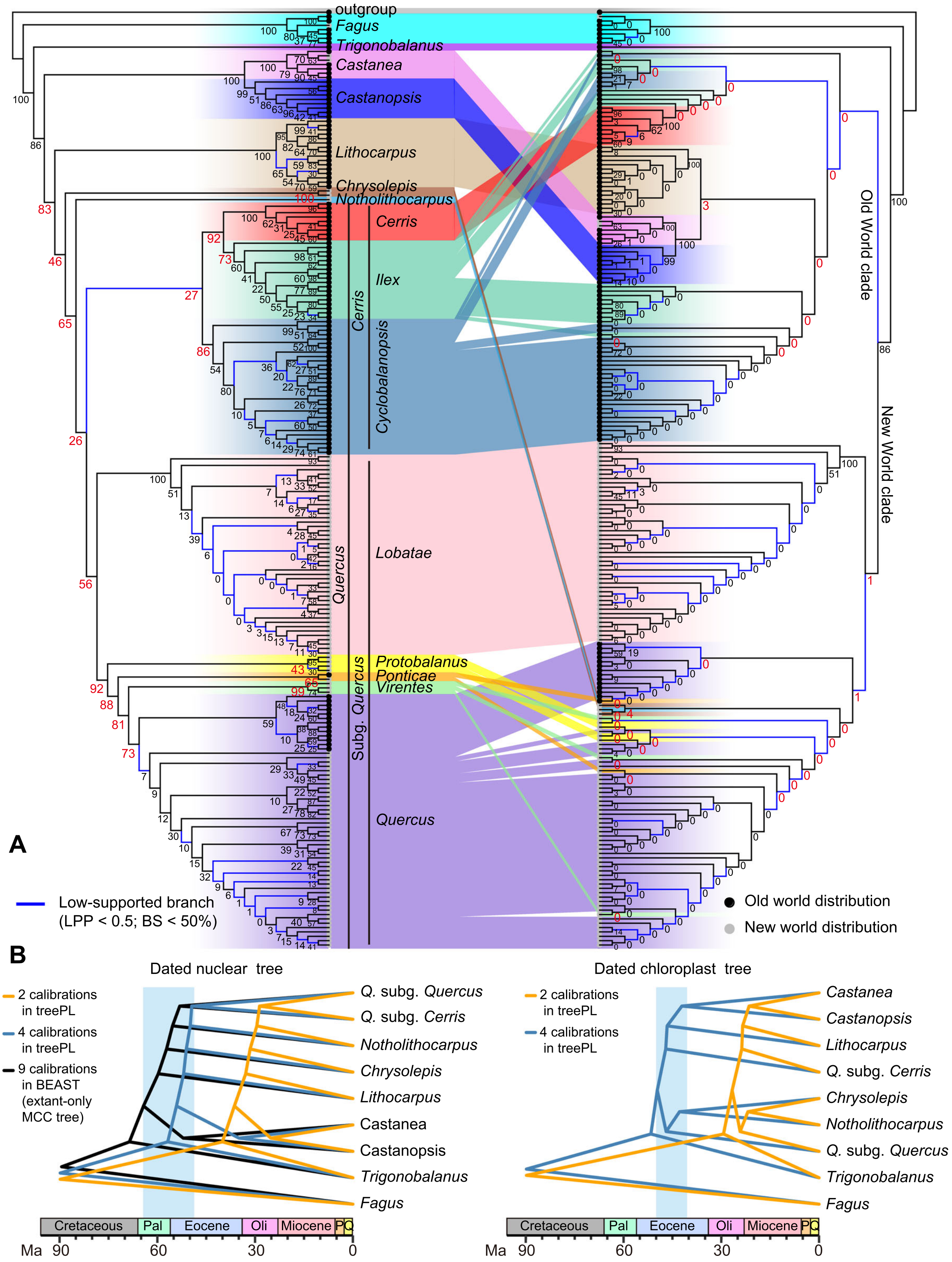


**Figure 2. Discordance between nuclear and chloroplast phylogenies for oaks and relatives**

**(A)** Cophylogeny showing incongruence between the nuclear ASTRAL (left; HYB-98RT) and unpartitioned chloroplast maximum likelihood (ML) (right) trees. Tip labels are shown in Figure S15. Clade frequencies of gene trees from the coalescent simulations are shown near the nodes; clade frequencies associated with deep cytonuclear discordances are red and enlarged. **(B)** Comparison of divergence times of *Quercus* and relatives in Quercoideae between the dated nuclear and chloroplast trees. The full dated trees are provided in Supplementary Materials. The light blue bar (ranging from the Early Paleocene to Middle Eocene) represents the timeframe of divergences of the main lineages in Quercoideae, based on analyses with four and nine calibrations. Oli, Oligocene; P, Pliocene; Pal, Paleocene; Q, Quaternary.

whether particular tree topologies could be significantly rejected by nuclear data. The chloroplast topology, the ML topology of the HYB-98RT dataset, and topologies from previous studies for five uncertain nodes were significantly worse ($P < 0.05$) than the nuclear topologyof all ASTRAL and most concatenated ML trees in this study (Table S2). Examination of the distribution of phylogenetic signal in alternative tree topologies of the five uncertain nodes (Figures S13, S14) showed that the proportions of genes and sites favoring a particular topology (85%–39% of genes and 91%–39% of sites) were usually greater than those favoring another topology (49%–15% of genes and 39%–9% of sites), but a large amount of conflicting signal was observed across all uncertain nodes and datasets. The proportions of supporting genes and sites of alternative topologies for five uncertain nodes are detailed in Supplementary Results.

### Simulating chloroplast trees under the coalescent

We applied coalescent simulations to investigate whether the instances of deep cytonuclear discordance that we observed were due to ancient gene flow or ILS. The organellar gene tree distribution simulated from the scaled ASTRAL tree under the coalescent (i.e., 1,000 simulated trees) showed that much of the nuclear backbone topology was within ILS predictions and had moderate to high clade probabilities (with some notable exceptions for sect. *Lobatae* and sect. *Quercus*); little of the chloroplast topology, except for some shallow regions, was within ILS predictions (clade probabilities mostly ~ 0%; Figures 2A, S15–S16). Focusing on the deep conflicting nodes between the nuclear and chloroplast trees (i.e., involving clades at the generic and sectional levels), clade probabilities of simulated trees evaluated on the chloroplast tree were less than 4% and 3% in the HYB-98RT and RNA-2821RT datasets, respectively (see red node labels in Figures 2A, S15, and S16), suggesting that the deep nodes of the chloroplast tree (that conflict with the nuclear species tree) are not within ILS expectations. Thus, gene flow is a better explanation for deep cytonuclear discordance in Quercoideae. A similar pattern was found in an organellar gene tree distribution simulated under the ASTRAL tree scaled by two (Figures S17, S18) and four (Figures S15, S16).

### Co-quantification of ILS, gene tree estimation error, and gene flow

We further estimated the relative contributions of ILS, gene flow, and gene tree estimation error to the observed gene tree discordance. Relative importance decomposition analysis demonstrated that linear models (on the basis of ILS, gene tree estimation error, and gene flow) explained 10.45% and 28.36% of the total gene tree variation in the HYB-98RT and RNA-2821RT datasets, respectively, across all internodes of Fagaceae, while the models explained 10.89% and 52.28% of gene tree variation in the HYB-98RT and RNA-2821RT datasets, respectively, across all internodes of *Quercus* (see $R^2$ in Figure 3). ILS and gene flow were the two main factors explaining gene tree conflict across the two datasets and taxonomic levels. Specifically, the relative contributions of ILS, gene flow, and estimation error across Fagaceae were 3.25%, 6.58%, and 0.62% in the HYB-98RT dataset (Figure 3A) and 19.64%, 5.15%, and 3.57% in the RNA-2821RT dataset (Figure 3C). Across *Quercus*, the relative contributions of ILS, gene flow, and estimation error were 7.79%, 2.19%, and 0.91% in the HYB-98RT dataset (Figure 3B) and 36.31%, 13.02%, and 2.95% in the RNA-2821RT dataset (Figure 3D).

### Introgression tests

To further investigate lineages potentially involved in deeper instances of introgression as well as the directionality of such introgression, we applied the five-taxon D-statistic ($D_{\mathrm{FOIL}}$) to oaks and other genera in Quercoideae. We found significant signals of introgression: 2.49% (132,435/5,316,699) in the $D_{\mathrm{FOIL}}$ tests for the HYB-98RT dataset and 3.58% (8,058/224,998) in the $D_{\mathrm{FOIL}}$ tests for the RNA-2821RT dataset (Table 1). Signatures of ancient introgression were inferred between *Quercus* and other genera in Quercoideae across the two datasets; however, signatures within *Quercus* were the most frequent overall (Figures S19, S20; Table S3). Furthermore, a summary of all positive $D_{\mathrm{FOIL}}$ tests (categorized by the two subgenera of *Quercus* and other genera in Quercoideae) showed a geographic pattern of introgression signatures, in which the number of significant $D_{\mathrm{FOIL}}$ tests between non-*Quercus* genera and subgenera of *Quercus* from the same continent was greater than that between non-*Quercus* genera and subgenera of oaks distributed on different continents (Table 1), consistent with the notion that species with overlapping distributions are more likely to hybridize.

### Phylogenetic network inference

We also conducted phylogenetic network analyses to explore whether reticulate evolution has occurred at deep locations in the Quercoideae phylogeny. Model selection suggested that any network was a better model than the strictly bifurcating species tree across the two datasets (Hyb-Seq and transcriptome; Tables S4, S5). The optimal networks of three representation strategies of two dataset types often presented incongruent scenarios, but all supported complex reticulations in the early evolutionary history of subfamily Quercoideae (Figure S21). Regarding the reticulation between oaks and their relatives, optimal networks recovered gene flow from oaks to the *Trigonobalanus excelsa* lineage, with an inheritance probability (IP) of 0.205 (Figure S21A), and from an ancestral lineage of *Quercus* + *Notholithocarpus* to the clade of sect. *Cerris* + sect. *Ilex* (IP = 0.353; Figure S21B) and to *Quercus* with a low IP of 0.001 (Figure S21D). Within oaks, prevalent gene flow was inferred among major lineages. For example, gene flow was detected from the ancestral lineage of the clade sect. *Virentes* + sect. *Quercus* to the *Q. palmeri* lineage of sect. *Protobalanus* (IP = 0.214; Figure S21C), and from the ancestral lineage of subg. *Quercus* to the clade sect. *Cerris* + sect. *Ilex* (IP = 0.319; Figure S21F).

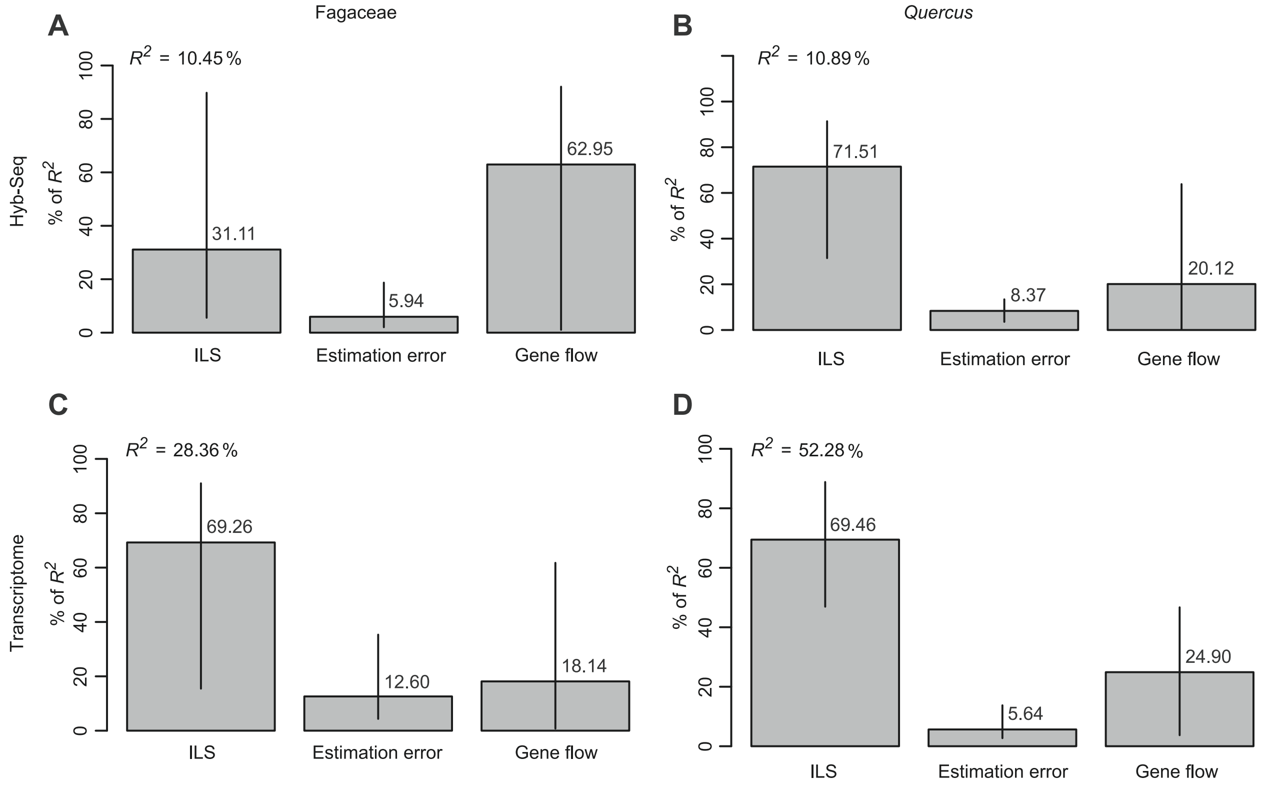


**Figure 3. Plots showing the relative contributions of incomplete lineage sorting (ILS), gene tree estimation error, and gene flow to the observed gene tree discordance, across two datasets and two taxonomic levels**

The Hyb-Seq **(A–B)** and transcriptome **(C–D)** datasets are represented here by the HYB-98RT and RNA-2821RT datasets, respectively. $R^2$ denotes the total proportion of gene tree variation explained by the model. Relative importance with 95% bootstrap confidence intervals is decomposed with the "lmg" method and regressors of log transformations and summed to 100%.

## Ancestral range estimations

Ancestral range and niche estimations were conducted to enable reconstruction of geographic and ecological scenarios of ancient hybridization during the early evolution of oaks. We also included paleogeographic information from the fossil record of oaks and their relatives, as a complement to the geographic and niche reconstructions using extant species and assessed the impact of their inclusion on the result. Biogeographic inferences under the DEC (dispersal-extinction-cladogenesis) model suggested that reconstructed ranges along the backbone of Fagaceae (particularly concerning the origin of Fagaceae) were not sensitive to the phylogenetic position of fossil taxa (i.e., changing positions within a particular lineage; Figure S22 vs. Figure S23; Figure S24 vs. Figure S25), but were sensitive to the inclusion of fossil taxa (Figure 4A vs. Figure 4B, C) and the range scoring scheme for fossil taxa (Figure 4B vs. Figure 4C). The extant-only and extant–extinct biogeographic results both reconstructed an ancestral distribution in the North American regions (Figure 4D, S26) for the stem and crown groups of three New World lineages of Quercoideae (*Chrysolepis*, *Notholithocarpus*, and subg. *Quercus*). The extant-only biogeographic result showed the estimated ancestral distributions in the East Asian region for stem and crown groups of four Old World lineages of Quercoideae (*Castanea, Castanopsis*, *Lithocarpus*, and subg. *Cerris*); in contrast, the extant–extinct biogeographic analysis recovered North America as the best-supported ancestral area for this group, with East Asia as a possible region of co-occurrence (Figure 4D).

## Ancestral niche estimations

Analysis of ancestral niche reconstruction based on extant species showed that the environmental spaces of the crown group of *Castanea* + *Castanopsis*, the crown group of *Lithocarpus*, and the crown group of subg. *Cerris* overlapped for all representative variables, which included aspects of climate, topography, soil, and landcover (Figures 5A, S27). Overlapping niche spaces were also inferred between the crown group of *Chrysolepis* and the crown group of subg. *Quercus*, as well as between *Notholithocarpus* and the crown group of subg. *Quercus* (Figure 5B). These findings were robust to the inclusion of fossil data (Figures 5C–H, S28–S31). The reconstructed temperature and precipitation values of these crown groups indicated the ancestors of these lineages lived in a warm and semi-moist to moist environment (Figures S27–S31).

## Paleoecological niche modeling

The paleoecological niche modeling (PaleoENM) analysis suggested that suitable habitats for the major lineages of Quercoideae were quite widespread across the Northern

**Table 1. Numbers of positive $D_{\text{FOIL}}$ introgression tests for *Quercus* and relatives in Quercoideae.**

| | HYB-98RT | | RNA-2821RT | |
|---|---|---|---|---|
| Introgression signatures | *Quercus* subg. Cerris | *Quercus* subg. *Quercus*# | *Quercus* subg. *Cerris* | *Quercus* subg. *Quercus*# |
| *Castanea* –> X | 92 | 6 | 5 | 19 |
| *Castanea/Castanea* <–> X | 2,931 | 1,064 | 170 | 487 |
| *Castanopsis* –> X | 145 | 0 | 3 | 0 |
| *Castanopsis/Castanea* <–> X | 7,253 | 1,000 | 5 | 36 |
| *Castanopsis/Castanopsis* <–> X | 19,355 | 534 | 12 | 33 |
| *Lithocarpus* –> X | 154 | 15 | 22 | 17 |
| *Lithocarpus/Lithocarpus* <–> X | 7,147 | 3,851 | 649 | 514 |
| *Chrysolepis*$^{\#}$/*Chrysolepis*$^{\#}$ <–> X | 201 | 730 | 0 | 0 |
| X –> *Castanea* | 65 | 1 | 1 | 12 |
| X –> *Castanopsis* | 95 | 0 | 2 | 13 |
| X –> *Lithocarpus* | 476 | 101 | 15 | 41 |
| X –> *Quercus* subg. *Cerris* | 39 | 175 | 4 | 125 |
| X –> *Quercus* subg. *Quercus*$^{\#}$ | 994 | 12 | 89 | 233 |
| X/X <–> *Castanea* | 483 | 409 | 15 | 330 |
| X/X <–> *Castanopsis* | 700 | 90 | 13 | 107 |
| X/X <–> *Lithocarpus* | 1,258 | 1,112 | 14 | 339 |
| X/X <–> *Quercus* subg. *Cerris* | 1,677 | 16,905 | 18 | 1,440 |
| X/X <–> *Quercus* subg. *Quercus*$^{\#}$ | 12,421 | 825 | 1,117 | 1,765 |
| Other introgression signal | 50,119 | | 393 | |
| No introgression signal | 5,184,264 | | 216,940 | |
| Total number of $D_{\text{FOIL}}$ tests | 5,316,699 | | 224,998 | |

*Note*: "X" means *Quercus* subg. *Cerris* (second and fourth columns) or *Quercus* subg. *Quercus* (third and fifth columns). The directionality of an introgression signature is indicated by the arrow. Taxa with "#" are only (or nearly only) distributed in the New World, while other taxa are only (or nearly only) distributed in the Old World.

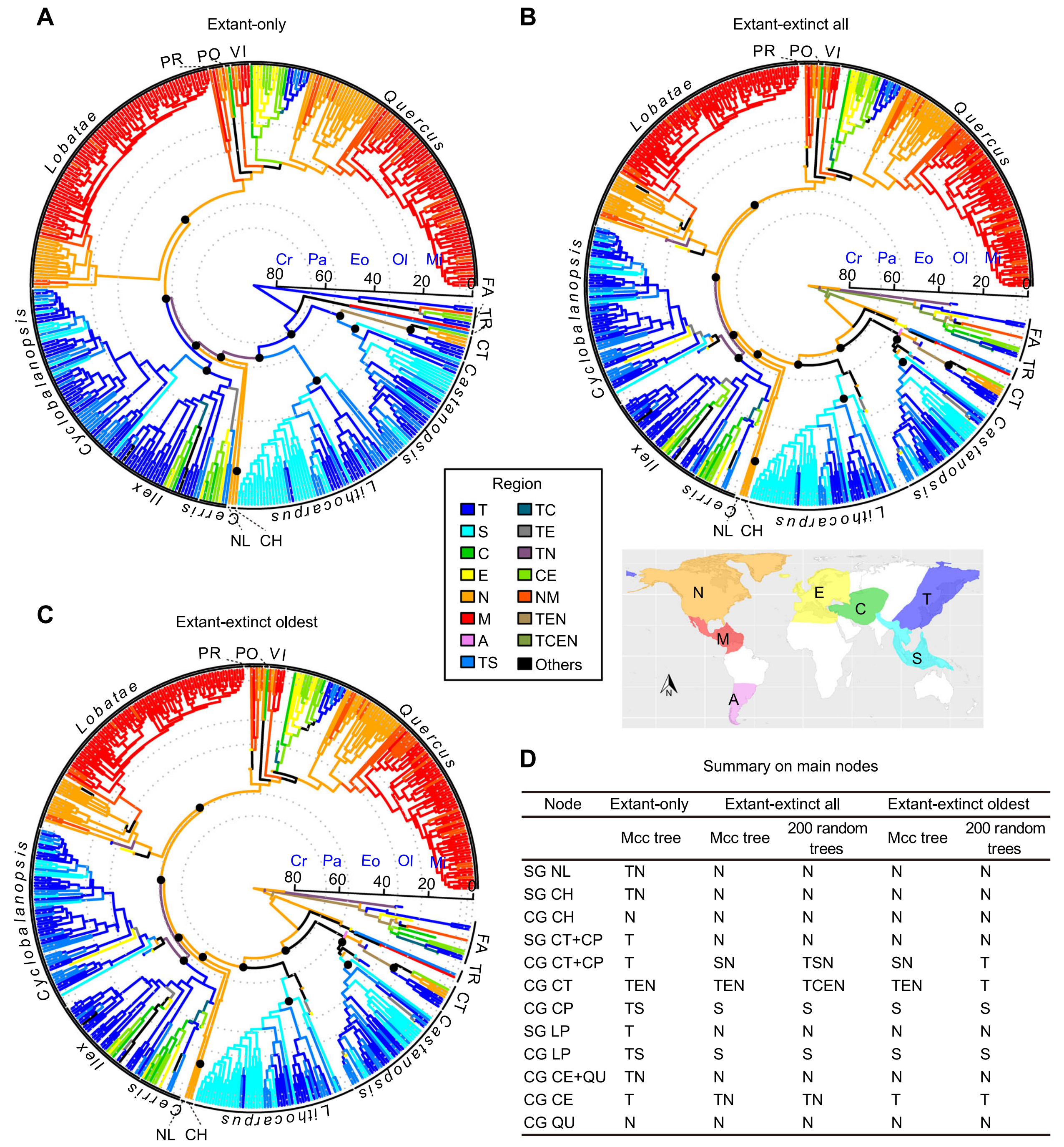
Hemisphere during the Early Paleogene. Four Old World lineages (*Castanea, Castanopsis*, *Lithocarpus*, and subg. *Cerris*) potentially co-occurred in North America and Eurasia during the Late Paleocene to the Late Eocene, and three New World lineages (*Chrysolepis*, *Notholithocarpus*, and subg. *Quercus*) potentially co-occurred in North America (Figures 6, S32). Toward the Oligocene, reconstructed potential distributions of these lineages were found to have expanded and the number of their fossil record have increased, particularly for temperate lineages like subg. *Quercus* (Figure S32), suggesting the rapid diversification of major crown lineages in Quercoideae after the Eocene.

# DISCUSSION

## New insights into the phylogeny of oaks and relatives

Our nuclear phylogenetic analyses resolved relationships among eight monophyletic genera in Fagaceae (i.e., *Castanea, Castanopsis, Chrysolepis*, *Fagus, Lithocarpus*, *Notholithocarpus*, *Quercus*, and *Trigonobalanus*) (Figures 1, S1–S8), with results generally consistent with previous studies (e.g., Manos et al., 2001; Hai et al., 2022; Zhou et al., 2022; Yang et al., 2023b). The exception is the placement of *Chrysolepis*, which is a key clade for the investigation of the evolution of reproductive traits, such as cupules, in Fagaceae (Manos et al., 2001, 2008). Three alternative nuclear topological placements were recovered for *Chrysolepis*: (1) sister to the clade of *Notholithocarpus* + *Quercus* (in our study, Zhou et al., 2022 using 91 Fagaceae species and 2,124 nuclear loci, and Yang et al., 2023b using 75 Fagaceae species and 643 nuclear loci); (2) sister to *Lithocarpus* (in our study, and Zhou et al., 2022); and (3) sister to the clade of (*Lithocarpus*, (*Notholithocapus*, *Quercus*)) (Zhou et al., 2022). The results of phylogenetic signal investigation (Figure S14), coalescent simulations (Figures 2A, S15–S18, S33, and S34), and $D_{\text{FOIL}}$ tests (Figures S19, S20; Tables 1, S3) indicate that the combination of both ILS and ancient hybridization is making it difficult, if not impossible, to confidently reconstruct a bifurcating species tree for *Chrysolepis* and its immediate relatives (*Lithocarpus*, and *Notholithocarpus* + *Quercus*). Future studies employing multifaceted analyses of complete nuclear genome sequences of *Chrysolepis* and close relatives might provide clearer resolution of the placement of this genus as well as further insight on the nature of the observed phylogenomic conflict.

The nuclear trees support the monophyly of *Quercus* and phylogenetic structure consistent with the taxonomic division of *Quercus* into two subgenera (*Cerris*, and *Quercus*) and eight sections (*Cerris*, *Cyclobalanopsis*, *Ilex*, *Lobatae*, *Ponticae*, *Protobalanus*, *Quercus*, and *Virentes*) (Figures 1, S1–S8). These findings are consistent with other recent studies (e.g., Denk et al., 2017; Hipp et al., 2020; Zhou et al., 2022; Yang et al., 2023b). The placement of sect. *Cyclobalanopsis*, which is characterized



D

Summary on main nodes

| Node | Extant-only | Extant-extinct all | | Extant-extinct oldest | |
|---|---|---|---|---|---|
| | Mcc tree | Mcc tree | 200 random trees | Mcc tree | 200 random trees |
| SG NL | TN | N | N | N | N |
| SG CH | TN | N | N | N | N |
| CG CH | N | N | N | N | N |
| SG CT+CP | T | N | N | N | N |
| CG CT+CP | T | SN | TSN | SN | T |
| CG CT | TEN | TEN | TCEN | TEN | T |
| CG CP | TS | S | S | S | S |
| SG LP | T | N | N | N | N |
| CG LP | TS | S | S | S | S |
| CG CE+QU | TN | N | N | N | N |
| CG CE | T | TN | TN | T | T |
| CG QU | N | N | N | N | N |

**Figure 4. Ancestral geographic regions that overlapped between ancestors of *Quercus* lineages and relatives in Quercoideae putatively involved in deep reticulation**

Tip labels of Figures 4A, 4B, and 4C are shown in Figures S26, S22, and S24, respectively. Biogeographic reconstructions on 200 randomly sampled extant–extinct trees are shown in Figures S23 and S25. The dashed gray circles indicate different geological periods. The black dots indicate the crown and stem nodes of major lineages in Quercoideae. Fagaceae genera (outer circle) and oak sections (inner circle) are indicated by black bars. Cr, Cretaceous; Eo, Eocene; Mi, Miocene; Ol, Oligocene; Pa, Paleocene. **(A)** Biogeographical reconstruction based on the extant-only maximum clade credibility (MCC) tree. **(B)** Biogeographic reconstruction based on the extant–extinct MCC tree and with the geographic region of each fossil taxon being scored by its entire fossil record (i.e., "all" records). **(C)** Biogeographic reconstruction based on the extant–extinct MCC tree and with the geographic region of each fossil taxon being scored by its oldest fossil record. **(D)** Reconstructed distributions for the crown (CG) and stem (SG) groups of major lineages in Quercoideae. CE, subg. *Cerris*; CH, *Chrysolepis*; CP, *Castanopsis*; CT, *Castanea*; FA, *Fagus*; LP, *Lithocarpus*; NL, *Notholithocapus*; PO, sect. *Ponticae*; PR, sect. *Protobalanus*; QU, subg. *Quercus*; TR, *Trigonobalanus*; and VI, sect. *Virentes*.

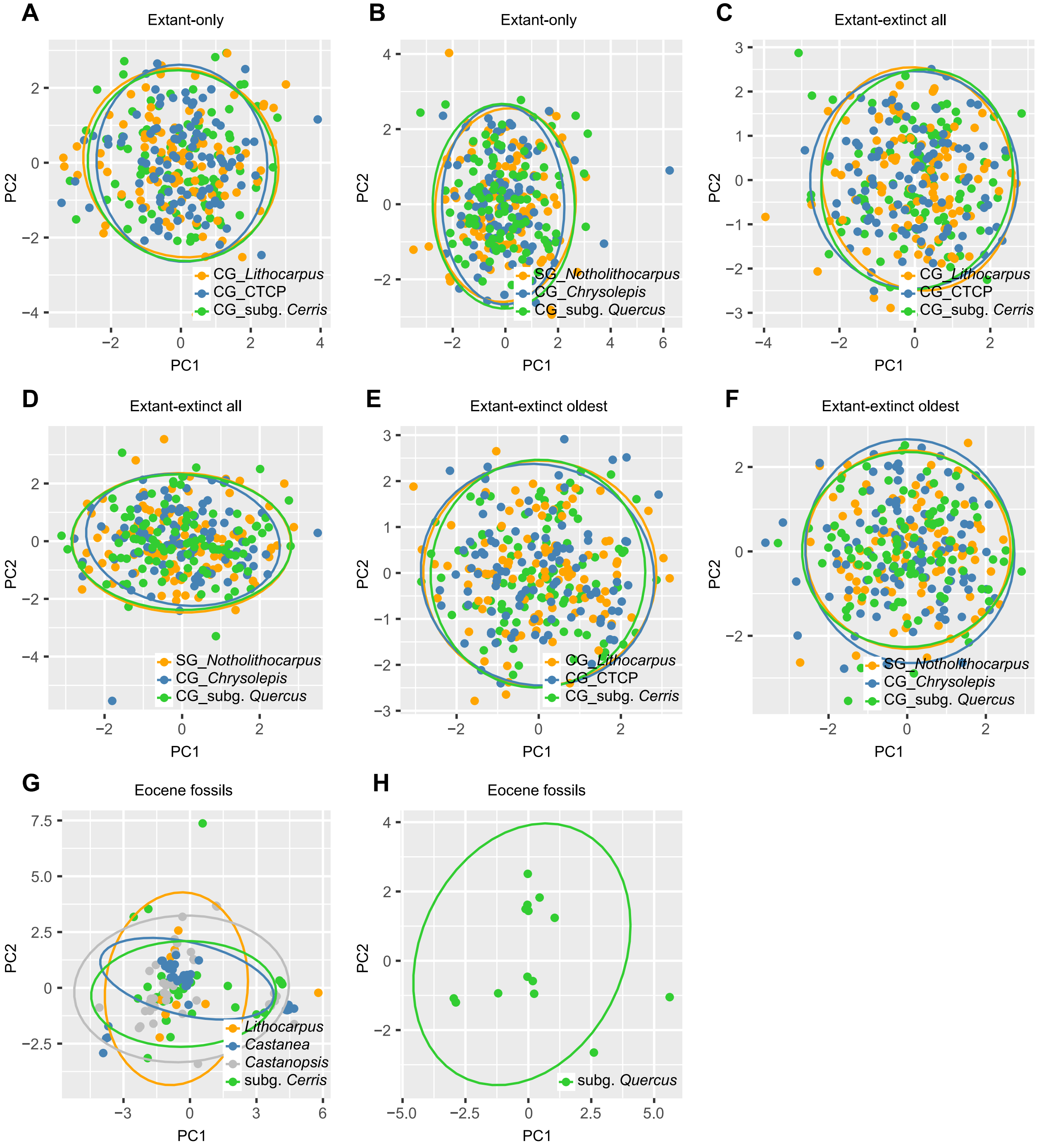


**Figure 5. Ancestral ecological niche space overlapping between ancestors of *Quercus* lineages and relatives in Quercoideae putatively involved in deep reticulation**

The first two components are shown through a principal component analysis on the estimated ancestral niche of focused lineages. **(A–B)** Ancestral niche reconstruction based on the extant-only maximum clade credibility (MCC) tree. **(C–D)** Ancestral niche reconstruction based on the extant–extinct MCC tree and with the niche of each fossil taxon being estimated by its entire fossil record ("all"). **(E–F)** Ancestral niche reconstruction based on the extant–extinct MCC tree and with the niche of each fossil taxon being estimated by its oldest fossil record. **(G–H)** Climatic niche space of Eocene fossil taxa of *Quercus* lineages and relatives. The Eocene niches of *Chrysolepis* and *Notholithocarpus* are not shown in Figure 5H due to the limited Eocene fossil record of these two genera. CG, crown group; CTCP, *Castanea* + *Castanopsis*; SG, stem group.

by concentric lamellae on the cupule, has long been controversial (Manos et al., 1999; Hubert et al., 2014; Deng et al., 2018). Once again, three alternative topological placements were recovered for sect. *Cyclobalanopsis*: (1) sister to the clade of sect. *Ilex* + sect. *Cerris* (in our study, Hubert et al., 2014; Deng et al., 2018; Hipp et al., 2020); (2) sister to the remaining oaks (Manos et al., 1999; Hubert et al., 2014); and (3) sister to *Notholithocarpus* (in our study). However, our analysis clearly

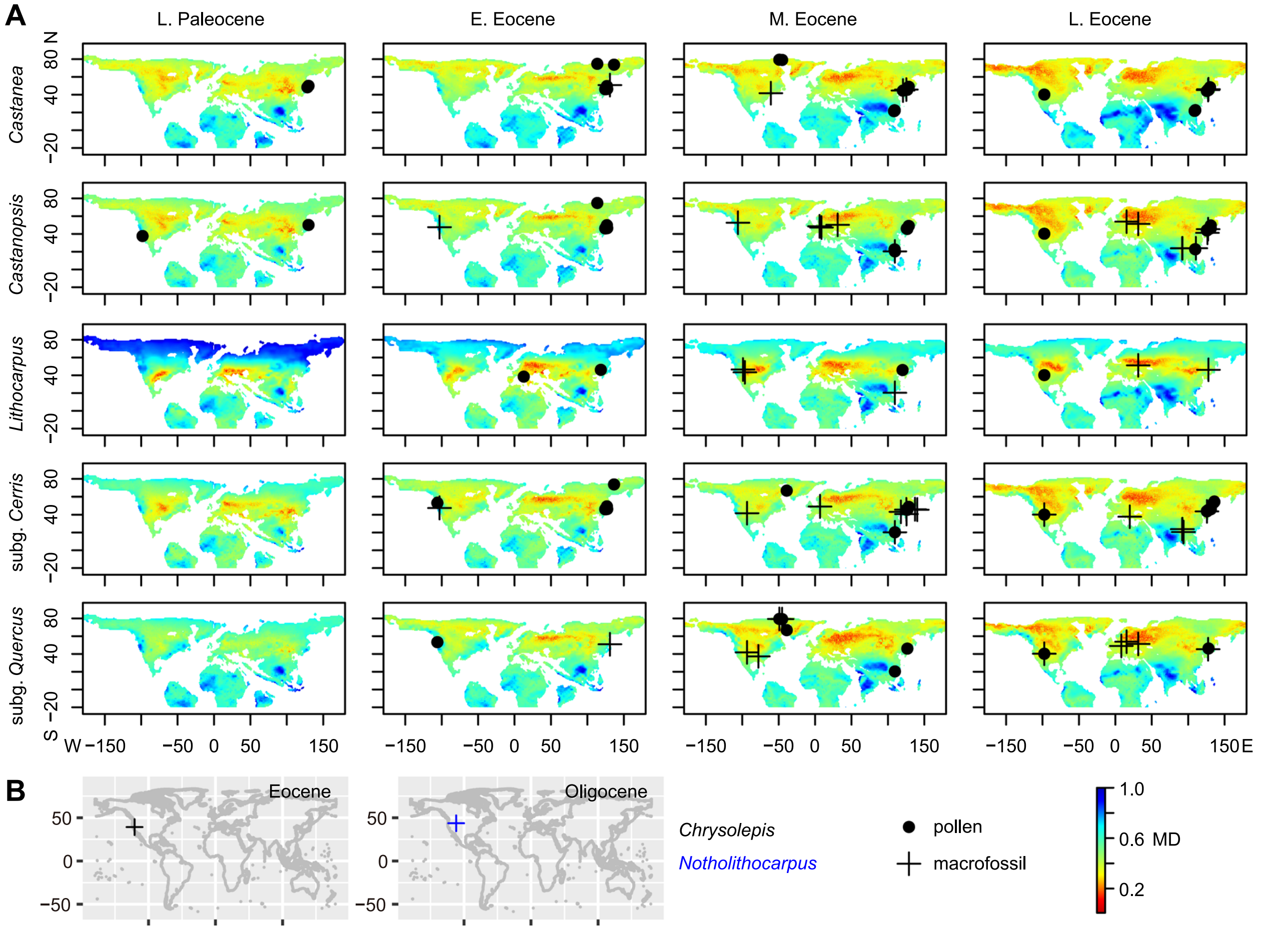


**Figure 6. Habitat suitability and fossil distribution of the Paleogene overlapping between ancestors of *Quercus* lineages and relatives in Quercoideae putatively involved in deep reticulation**

**(A)** Potential distributions of two subgenera of *Quercus* and other genera of Quercoideae from the Late Paleocene to the Late Eocene as inferred by paleoecological niche modeling. Maps were generated by projecting the climatic tolerances of Paleocene and Eocene fossil taxa onto four paleoclimate scenarios. Similar patterns of past distribution were obtained by projecting the climatic tolerances of Paleocene, Eocene, and Oligocene fossils onto six paleoclimate scenarios (Figure S32). A Mahalanobis distance (MD) score of <0.3 (red) corresponds with a highly suitable region for each Quercoideae lineage; climatically unsuitable regions are colored green to blue. E., Early; L., Late; M., Middle. **(B)** Fossil distributions of *Chrysolepis* and *Notholithocarpus* during the Paleogene.

supports the first topology as the most likely resolution of sect. *Cyclobalanopsis* with the highest gene-wise and site-wise phylogenetic signal (Figures S13, S14). We argue conflict involving sect. *Cyclobalanopsis* may represent not only ILS but also past introgression in the early history of oak diversification, based on the combined results of our coalescent simulations and species networks (Figure S21E, F) as well as previous analyses (Schnitzler et al., 2004; Mir et al., 2006; Zhou et al., 2022).

Regarding relationships among subclades within oak sections, we recovered some well-supported relationships that differed from previous studies (e.g., in sect. *Ilex*, and in sect. *Cyclobalanopsis*; Deng et al., 2018; Jiang et al., 2019; Hipp et al., 2020). Notably, we recovered *Q. gaharuensis*, a species inhabiting Borneo, Sumatra, and Malaysia that had not been previously included in any phylogenetic studies, as sister to the rest of sect. *Cyclobalanopsis* (Figures S1–S4). The compound trichome bases group defined by Deng et al., (2014, 2018) was previously believed to be sister to the rest of the section (Deng et al., 2018; Hipp et al., 2020). This section was previously inferred to have originated in the Paleotropics (Deng et al., 2018), and the placement of *Q. gaharuensis* sister to the rest of the section, as found here, corroborates this earlier finding. In other sections (e.g., sect. *Lobatae* and sect. *Quercus*), there is considerable conflict concerning phylogenetic relationships and support values estimated from different datasets and methods (Figures 1, S1–S8). Our work suggests that dataset size (or gene number) and analytical method may significantly affect the topologies and support values on these nodes with strong gene tree conflicts arising from ILS during rapid radiations as well as active and/or ancient gene flow among the constituent species, as previously reported for oaks (Hipp, 2015; Cannon and Petit, 2020; Kremer and Hipp, 2020).

## Dissecting causes of gene tree heterogeneity and deep cytonuclear discordance

We observed high levels of gene tree discordance across the Fagaceae phylogeny, particularly within the clade of *Chrysolepis*, *Lithocarpus*, *Notholithocarpus*, and *Quercus* (Figures 1, S9–S11), consistent with recent studies (Zhou et al., 2022; Yang et al., 2023b). ILS and gene flow explain most of the observed gene tree discordance in our analysis with gene tree estimation error accounting for a lower proportion (Figure 3). It is noteworthy that gene flow estimates differed between datasets, suggesting that taxon and locus sampling might partly determine how much gene flow is observable in any particular dataset. Similarly, we found that the gene tree variation explained by the model of these three factors decreased as taxon sampling increased and locus sampling decreased, similarly suggesting that observed gene tree discordance is influenced by taxon and locus sampling.

High levels of cytonuclear discordance were observed in both deep and shallow divergences within *Quercus* and across Quercoideae (Figures 2A, S15–S18), as has been shown by previous studies (e.g., Manos et al., 1999; Simeone et al., 2016; Pham et al., 2017; Yang et al., 2021; Zhou et al., 2022). However, our increased sampling compared to earlier efforts, with close taxonomic complementarity between the nuclear and chloroplast datasets, provided greater resolution of cytonuclear incongruence patterns across Quercoideae than in previous studies. In our chloroplast phylogeny, *Quercus* was recovered as non-monophyletic, with members of the genus forming two geographically distinct clades with the other genera of Quercoideae: a clade of (*Lithocarpus*, (*Castanea*, *Castanopsis*)) nested in the Old World oak clade (corresponding to the traditionally recognized subg. *Cerris*); and a clade with *Chrysolepis* + *Notholithocarpus* nested in the New World oak clade (corresponding to the traditionally recognized subg. *Quercus*). Given that both gene flow and ILS can result in discordance between nuclear and chloroplast phylogenies (Rieseberg and Soltis, 1991; Folk et al., 2017; Rose et al., 2021), we conducted analyses of coalescent simulations and $D_{\text{FOIL}}$ tests to identify the causes of the observed instances of cytonuclear discordance. Our coalescent simulations demonstrated that observed cytonuclear discordance was partially within ILS predictions at shallow levels, but at deeper levels the chloroplast backbone topology is not within ILS expectations and supports gene flow as the source of observed deep conflicts (Figure 2A). We found that chloroplast topologies of *Castanea* + *Castanopsis*, *Lithocarpus*, *Chrysolepis*, and *Notholithocarpus* were supported by appreciable proportions of nuclear genes and sites (Figures S13, S14), suggesting that both chloroplast haplotypes and nuclear alleles were introgressed during historic instances of gene flow between *Quercus* and relatives. This scenario is further supported by the $D_{\text{FOIL}}$ tests, which detected evidence of ancient introgression between some Eurasian genera—*Lithocarpus*, *Castanea,* and *Castanopsis*—and the Old World oak clade as well as between North American *Chrysolepis* and the New World oak clade (Table 1). While introgression among these genera was not detected in a previous effort using ABBA–BABA tests (Zhou et al., 2022), this discrepancy is most likely due to study-specific characteristics: a five-taxon test as implemented in $D_{\text{FOIL}}$ provides richer information for distinguishing hypotheses than a four-taxon test, and the relatively lower number of four-taxon tests conducted in Zhou et al. (2022) is a reflection of limited taxon sampling in the previous study. Overall, our analyses consistently show that ancient gene flow, rather than ILS, is the primary source of deep cytonuclear discordance in Quercoideae. Our findings support the previous viewpoint that cytonuclear or mito-nuclear discordance caused by gene flow is often closely associated with geographical adjacency (Acosta and Premoli, 2010; Toews and Brelsford, 2012).

Similar geographic patterns of ancient gene flow (including cytonuclear discordance) were also found between species of oaks from clades recognized as sections—for example, between *Q. pontica* (sect. *Ponticae*) and the Roburoid lineage (sect. *Quercus*), between species of sect. *Cyclobalanopsis* and some species of sect. *Ilex*, between sect. *Cerris* and some species of sect. *Ilex*, and between sect. *Protobalanus* and the Dumosae lineage (sect. *Quercus*) (all lineage names following Hipp et al., 2020). In these cases, the chloroplast topologies were not within ILS predictions (Figures 2A, S15–S18), suggesting instead the occurrence of historic asymmetrical gene flow between these lineages, leading to a local geographic structure in the chloroplast phylogenies at odds with the nuclear species tree reconstruction. Furthermore, an abundance of nodes with moderate or weak BS support in chloroplast trees, particularly toward the tips (Figure S12), suggests that plastomes have generally tracked geographic structure and have limited information on recent oak divergences, in agreement with previous studies (Yang et al., 2016; Pham et al., 2017). Collectively, these findings indicate that the plastome is not a reliable source of data for resolving species tree relationships at any evolutionary scale in Quercoideae, although when paired with the nuclear genome, the plastome is useful for shedding light on historic gene flow in the evolutionary history of oaks.

## Genomic, geographic, and ecological evidence for widespread gene flow during the Eocene

Given the complex modern and historical distribution of oaks and relatives (Zhou 1999; Barrón et al., 2017; Cannon et al., 2018), the question arises as to whether the lineages suggested by molecular data to have undergone hybridization had ancestral distributions and ecologies consistent with that scenario; that is, were the lineages that putatively hybridized in the distant past co-distributed in time, space, and environmental niche? The comparison of divergence times from chloroplast and nuclear estimates showed that the divergences of *Quercus* and the other main lineages of Quercoideae (*Notholithocarpus, Chrysolepis*, *Lithocarpus*, and *Castanopsis* + *Castanea*) in the chloroplast tree occurred in the Middle Eocene, following the inferred divergence times from the nuclear tree (Paleocene to Early Eocene; Figure 2B). Our inferred chloroplast ages are younger than those from a recent study (Early to Middle Eocene; Zhou et al., 2022) likely due to differences in calibration schemes.

Ancestral range estimations using only extant species (Figures 4A, S26), and both extant and extinct species (Figures 4B, C, S22–S25), generally supported the geographic possibility of historical overlap facilitating gene flow. The biogeographic analyses reconstructed: (i) North America as an ancestral region of co-occurrence for three New World lineages of Quercoideae (*Chrysolepis*, *Notholithocarpus*, and subg. *Quercus*) during the Eocene; and (ii) East Asian and/or North America as an ancestral region of co-occurrence for four currently Old World lineages of Quercoideae (*Castanea, Castanopsis*, *Lithocarpus*, and subg. *Cerris*) (Figure 4D; see also Siniscalchi et al., 2023). The PaleoENM analysis (which reconstructs suitable habitats rather than dispersal processes) suggested that three New World lineages potentially co-occurred in North America during the Paleocene-Eocene, and four currently Old World lineages potentially co-occurred in North America and Eurasia (Figures 6, S32). The distribution of known fossils is similarly broad (Figure 5; Kvaček and Walther, 1989; Zhou 1999; Grímsson et al., 2015, 2016; Barrón et al., 2017). The disagreement between the relatively narrow range reconstruction under biogeographic models and the broader range implied by the fossil record might be explained by the limitations of relying only on data for extant species and/or including only a few representative fossil taxa. High extinction rates in the Northern Hemisphere, particularly North America, Europe, and Central Asia since the Oligocene reflect cooling and drying trends that would have removed many lineages in these regions, obscuring historical distributions (Siniscalchi et al., 2023).

Our ancestral reconstructions suggest that the lineages involved in ancient gene flow not only showed broad geographic overlap but also had similar ecological preferences during the Eocene; this finding was robust to the inclusion of fossil data (Figures 5, S27–S31). In particular, the crown groups of four Old World lineages of Quercoideae were estimated to have co-occurred in warm and moist habitats characteristic of evergreen broadleaf forests or warm yet semi-moist to moist habitats of mixed forests. The crown groups of three New World lineages of Quercoideae were inferred to have co-occurred in warm and semi-moist to moist habitats of mixed forests. However, we note that these reconstructed ancestral overlaps could be an artifact of each lineage encompassing considerable niche breadth today, with intermediate conditions generally inferred for the ancestors, although the use of fossil data directly in the analysis partly allays this concern. Another challenge is that gene flow could homogenize ecological niches among reticulating lineages, obscuring any differences present prior to hybridization.

Collectively, these multifaceted analyses demonstrate that inferred geographic and ecological overlap during the Eocene facilitated widespread gene flow in two possible geographic centers—North America and Eurasia—during the initial radiation of oaks (Figures 4–6). *Quercus* and relatives may have been engaging in syngameon-type systems throughout their evolutionary history. One possible outcome of long-term and ongoing gene flow is that an Early Paleogene syngameon may have facilitated the transfer of adaptive alleles for key functional traits or resulted in enhanced genetic variation. This may have significantly contributed to the evolutionary success of oaks and their relatives (Hipp et al., 2020; Kremer and Hipp, 2020), allowing them to adapt to the myriad ecological challenges posed since the Oligocene, including widespread cooling and aridification and major topographical changes (Zachos et al., 2001; Antonelli et al., 2018).

## CONCLUSION

This work provides new insights into the evolutionary history of *Quercus*, documenting widespread ancient reticulations between major lineages of *Quercus* and relatives. Ancestral reconstructions based on data from extant species and the fossil record consistently support the plausibility of ancient reticulation events during the Early to Middle Eocene in two possible geographic centers, North America and Eurasia, among co-occurring lineages occupying overlapping ecological niche space. Given the inherent challenges of detecting ancient gene flow, we emphasize the importance of bringing together multiple lines of evidence and analytical approaches for confident reconstruction of ancient reticulation. This includes the dissection of conflicting signals in nuclear and chloroplast datasets using summary statistics, simulations, and network methods; molecular dating of separate genomic compartments; and ancestral range and niche reconstructions leveraging signals from both extant and fossil species. These types of syntheses clarify not only the players but also the time, place, and ecological context of ancient reticulation. This work reveals that hybridization is not only an important recent and ongoing evolutionary force in *Quercus* and relatives, but also an important process throughout the long history of this group. Our work also provides a methodological framework and guide for conducting similar studies in other groups.

## MATERIALS AND METHODS

### Taxon sampling

We analyzed 472 samples representing 420 species of Fagaceae, including 312 species of *Quercus* (Quercoideae) that represent ~72% of extant oak diversity and eight currently recognized sections, and 108 species of the six other genera of Quercoideae (beyond *Quercus*) and Fagoideae. The number and percentage of species sampled for each section and genus are summarized in Table S6. Of the Fagaceae samples, 391 were newly sequenced using Hyb-Seq; data for the remaining 81 samples were obtained from transcriptomes (Yang et al., 2021; GenBank, accessed 1 May 2021; Table S7). Most species were represented by one sample, with the exception of 52 species of Fagaceae that had two samples each; of these, three species have two samples with Hyb-Seq data, and 49 species have one from Hyb-Seq data and another from transcriptome data. For outgroups, we included 15 species (seven from the new Hyb-Seq data and eight from available transcriptomes) representing all other six families of Fagales (i.e., Betulaceae, Casuarinaceae, Juglandaceae, Myricaceae, Nothofagaceae, and Ticodendraceae). Our Hyb-Seq samples were collected from the field and the following 10 herbaria: A, BRIT, CAS, F, KUN, MO, NY, OS, TEX, and US (acronyms following *Index Herbariorum*; Thiers, 2016).

### Library preparation and sequencing

For hybrid enrichment, we used a set of exonic baits for 100 housekeeping genes designed for use across the rosid clade of angiosperms, but with a focus on the nitrogen-fixing clade, which includes Fagaceae. This locus panel, the "NitFix loci," has been successfully applied to multiple families/genera within the rosids (e.g., Fu et al., 2023; Yang et al., 2023a; Kates et al., 2024; Tian et al., 2024) and is described in detail by Folk et al. (2021). Genomic DNAs were extracted from herbarium or silica-dried materials using a cetyltrimethylammonium bromide protocol (Doyle and Doyle, 1987) modified to maximize the yield of total DNA. Isolated DNAs were submitted to Rapid Genomics (Gainesville, FL, USA) for quantification, library preparation, hybrid enrichment with the NitFix loci, and multiplex Illumina sequencing with 150-bp, paired-end reads. The sampling-to-sequencing workflow we followed, briefly described above, is described in detail by Folk et al. (2021).

### Read processing, assembly, and orthology inference

For the Hyb-Seq data, raw reads were cleaned using Trimmomatic v.0.36 (Bolger et al., 2014) and assembled using HybPiper v.1.3.1 (Johnson et al., 2016). Orthology inference was carried out using several phylogenetic methods (i.e., MO and RT approaches) from the pipelines of Yang and Smith (2014) and Morales-Briones et al. (2021). For the transcriptome data, raw reads were corrected, trimmed, and filtered using Rcorrector v.1.0.4 (Song and Florea, 2015), Trimmomatic v.0.36, and Bowtie v.2.4.2 (Langmead and Salzberg, 2012). We conducted *de novo* assembly of cleaned reads using Trinity v.2.10.0 (Grabherr et al., 2011), removed low-quality and chimeric transcripts using Transrate v.1.0.3 (Smith-Unna et al., 2016), clustered cleaned transcripts as putative genes using Corset v.1.07 (Davidson and Oshlack, 2014), translated transcripts using TransDecoder v.5.3.0 (Haas et al., 2013), removed identical coding sequences using CD-HIT v.4.7 (Fu et al., 2012), and finally performed a BLASTN search to generate the initial homologs using MCL v.14-137 (van Dongen, 2000). The orthology inference was conducted using the same approach in the Hyb-Seq dataset. This workflow resulted in three Hyb-Seq datasets ("HYB-89MO," "HYB-114RT," and "HYB-98RT") and four transcriptome datasets ("RNA-2150MO," "RNA-4853RT," "RNA-977MO," and "RNA-2821RT") from different strategies of ortholog filtering. Further details on assembly, gene cluster processing, homolog tree inference, tip trimming, orthology inference, and ortholog cleaning are provided in Supplementary Methods; an overview of how these Hyb-Seq and transcriptome datasets were used in downstream analyses is presented in Figures S35 and S36. As described further in Supplementary Methods, a single ortholog-filtered dataset was chosen for the Hyb-Seq data ("HYB-98RT") and one dataset for the transcriptomic data ("RNA-2821RT") based on relatively more loci and higher taxon coverage in each for examination of gene tree discordance and gene flow unless otherwise indicated.

### Plastome assembly

We assembled plastomes from 391 Hyb-Seq and 60 transcriptome samples using a modified bash script wrapper (https://github.com/ryanafolk/Assembly-tools/) with read mapping using BWA v.0.7.12 (Li, 2013) and consensus sequence calling using SAMtools v.0.1.18 and BCFtools v.0.1.17 (Danecek et al., 2021). Forty-seven complete plastomes of Fagaceae and one of Betulaceae (used as an outgroup) were also downloaded from GenBank (accessed 5 April 2021). After filtering samples with over 80% plastome missing data and showing abnormal long branches in a preliminary tree, our final dataset included 222 plastomes of Fagaceae, consisting of 174 species of *Quercus* (one sample per species except for *Q. acrodonta* having two samples) and 47 species from other genera of Fagaceae, as well as one plastome from *Carpinus monbeigiana* of Betulaceae used to root the tree (Table S8). All eight currently recognized genera in Fagaceae and eight sections in *Quercus* were included in the final plastome dataset. The number and percentage of species sampled for each genus and section are summarized in Table S6. More details on plastome assembly and filtering are provided in Supplementary Methods. An overview of the plastome dataset used in our downstream analyses is provided in Figure S37.

### Phylogenetic analyses

We generated phylogenetic trees using ML and coalescent-based methods. Preliminary ML analyses of the concatenated nuclear matrices were conducted in RAxML v.8.2.11 (Stamatakis, 2014) with bootstrap support estimated by 100 fast bootstrap replicates and showed similar topologies under an unpartitioned GTR-GAMMA model and a partitioned GTR-GAMMA model with optimal partition scheme estimated in PartitionFinder v.2.1.1 (Lanfear et al., 2016) (results of which are provided in Supplementary Materials). The final ML analyses of the concatenated nuclear matrices were conducted using an unpartitioned GTR-GAMMA model and 1,000 fast bootstrap replicates (hereafter, "concatenated ML nuclear tree"); individual nuclear gene trees were also generated using this same approach. Coalescent-based species tree inference was conducted using the nuclear gene trees in ASTRAL-III v.5.6.3 (Zhang et al., 2018), and support values were estimated with local posterior probabilities (Sayyari and Mirarab, 2016) (hereafter, “ASTRAL nuclear tree”). The entire plastome has long been considered to comprise a single evolutionary unit from a coalescent point of view (Birky, 1995; Doyle, 2022); plastome sequences are therefore usually concatenated to amplify the phylogenetic signal. Thus, the chloroplast phylogeny was reconstructed through ML analyses of the concatenated plastome matrix with 1,000 bootstrap replicates under an unpartitioned GTR-GAMMA model.

### Divergence time estimation

To generate a dated nuclear tree of extant species, we first identified a set of 20 clock-like genes using SortaDate v.1.0 (Smith et al., 2018). A Bayesian dating analysis was then performed in BEAST v.2.6 (Bouckaert et al., 2014) under an uncorrelated lognormal relaxed clock, using a concatenated matrix of the 20 nuclear genes and all 423 Fagaceae taxa, with species outside Fagaceae excluded. We employed a Yule speciation process and GTR-GAMMA substitution model and fixed the topology according to the concatenated ML nuclear tree. Ten well-vetted fossils were used to calibrate the ages of nine nodes: crown group (CG) of Fagaceae, CG of *Fagus*, stem group (SG) of *Castanopsis*, SG of *Quercus*, and five major lineages of *Quercus* (see Supplementary Methods and Table S9 for detailed justification and prior setting). Monte Carlo Markov chains (MCMC) were run for 600 million generations, sampling every 10,000 generations. Convergence was assessed in Tracer v.1.7 (Rambaut et al., 2018) with effective sample size (ESS) of each parameter higher than 200. The MCC tree was summarized in TreeAnnotator v.2.6 (Bouckaert et al., 2014) with the first 20% of the trees discarded as burn-in (hereafter, “extant-only MCC tree”).

To assess the impact of fossil taxon inclusion on the geographic and niche reconstructions, we estimated a dated nuclear tree including extinct species as tips using a tip-dating analysis in BEAST under the fossilized birth-death process (Heath et al., 2014). The dataset contained 423 extant taxa of Fagaceae with 20 nuclear genes and 66 extinct macrofossil taxa of Fagaceae selected to represent the taxonomic and geographic breadth of the Fagaceae fossil record. We fixed the topology of extant species according to the concatenated ML nuclear tree, while the positions of extinct taxa were constrained to particular lineages based on the literature (e.g., Larson-Johnson, 2016; Siniscalchi et al., 2023; Yang et al., 2023b; see detailed validation information for each fossil taxon in Supplementary Materials). The age of origin of each fossil taxon was sampled from its stratigraphical age range under a uniform distribution. We used the same substitution model and nine node calibrations as in the extant-Fagaceae-only BEAST analysis. We ran six independent MCMC analyses each for 600 million generations and one independent MCMC analysis for one billion generations, sampling every 10,000 generations. After removing the first 20% of the trees from each independent run as burn-in, we combined the results of the seven MCMC analyses using LogCombiner v.2.6 (Bouckaert et al., 2014). The ESS values exceeded 200 for all parameters. We summarized this “extant–extinct MCC tree” using TreeAnnotator.

Given that the size of the plastome dataset precluded analysis in BEAST, dating of the chloroplast tree was performed using penalized likelihood in treePL v.1.0 (Smith and O'Meara, 2012). Comparable treePL analyses were also performed on the extant-only nuclear ML tree to facilitate a comparison of the ages of deep divergences in the nuclear and chloroplast trees, which can indicate the possible temporal window for ancient hybridization (see Supplementary Methods for more details on the rationale). We used two calibration strategies: (i) two fossil calibrations at concordant nodes: CG of Fagaceae, and CG of *Fagus*; and (ii) four fossil calibrations at concordant nodes: CG of Fagaceae, CG of *Fagus*, SG of the East Asian clade in sect. *Cerris* (monophyletic in chloroplast tree), and SG of sect. *Lobatae* (monophyletic in chloroplast tree) for these analyses, given that alternative relationships in the chloroplast tree precluded the use of several calibration points used in the nuclear dating analysis.

### Dissecting discordance and detecting gene flow

Concordance analyses

We performed phyparts analysis (Smith et al., 2015) using gene trees with low-supported branches (i.e., BS < 70%) collapsed; these gene trees were mapped against the ASTRAL species tree. Visualization of results was carried out using the script “phypartspiecharts.py” (https://github.com/mossmatters/phyloscripts/). Moreover, to distinguish weakly supported branches from those with strong conflict, we conducted QS analysis (Pease et al., 2018) with 1,000 replicates and using the ASTRAL species tree and concatenated alignment as the inputs.

Tree topology tests

For the conflicting deep nodes, we first tested whether insufficient resolution was responsible for the conflicting topologies by performing a polytomy test (Sayyari and Mirarab, 2018) in ASTRAL-III, asking whether hard polytomies could be rejected at these uncertain nodes. Second, we used the AU test (Shimodaira, 2002) to examine whether particular tree topologies could be significantly rejected by nuclear data using IQ-TREE v.2.1.2 (Minh et al., 2020) with 10,000 RELL replicates. Third, we quantified the phylogenetic signal of each topology using the pipeline of Shen et al. (2017). Phylogenetic signal here is defined as the difference in log-likelihood scores between two (or three) alternative resolutions (T1 and T2/T3) of a given node in a tree. We calculated the difference in the site-wise log-likelihood scores (ΔSLS) and in the gene-wise log-likelihood scores (ΔGLS). Specifically, we estimated the site-wise log-likelihood scores for T1 and T2 (and T3) using the concatenated nuclear matrix and RAxML v.8.2.11 with –f G. ΔSLS was then calculated as the difference in the site-wise log-likelihood scores of T1 versus T2 (vs. T3) for every site in the data matrix. ΔGLS was calculated as the sum of ΔSLS of all sites in a given gene.

### Coalescent simulations to examine cytonuclear discordance

Following previous studies (e.g., Folk et al., 2017; Morales-Briones et al., 2018), 1,000 organellar gene trees were simulated under the coalescent model in the Python package "DendroPy" (Sukumaran and Holder, 2010) with the ASTRAL species tree as the guide tree. The guide tree was pruned to match the taxon set of the chloroplast tree and scaled by a factor of two or four to approximate the branch lengths expected under organellar inheritance (for species with hermaphroditic and dioecious flowers, respectively). We then summarized the frequency at which each clade was observed in the simulated trees by mapping the simulated trees against the chloroplast and nuclear species trees using RAxML v.8.2.11 with -f b. Nuclear tree branches with high frequencies of alternative topologies (including the empirical chloroplast topology) are consistent with expectations under ILS. However, if the empirical chloroplast topology is observed rarely (or never) in the simulated organellar trees (vs. the nuclear species tree), this suggests that the empirical chloroplast topology is likely a result of gene flow (García et al., 2017; Morales-Briones et al., 2018; Stull et al., 2020).

### Quantifying gene tree discordance due to ILS, estimation error, and gene flow

We used the pipeline of Cai et al. (2021) to estimate the relative contributions of ILS, gene flow, and gene tree estimation error to observed gene tree discordance. Briefly, we first estimated values of these four variables (see details in Supplementary Methods) and formatted them into a matrix of a dependent (i.e., gene tree discordance) and three independent (i.e., ILS, gene tree estimation error, and gene flow) variables across all internodes with each row representing these values for one node. The relative importance of three regressors was assessed using a linear regression method implemented by the functions "boot.relimp" and "booteval.relimp" in the R package "relaimpo" (Grömping, 2006).

### $D_{\mathrm{FOIL}}$ analyses

We applied the $D_{\mathrm{FOIL}}$ tests to oaks and their relatives using Ex-$D_{\mathrm{FOIL}}$ (Lambert et al., 2019). $D_{\mathrm{FOIL}}$ tests are a system of applying *D*-statistics (Durand et al., 2011) to a symmetric five-taxon phylogeny (((P1,P2),(P3,P4)),O), with extremely low false-positive rates (Pease and Hahn, 2015). To achieve computational efficiency, we first reduced the 423 sampled Fagaceae taxa of HYB-98RT dataset to 150 taxa (representing *Castanea*, *Castanopsis*, *Lithocarpus*, *Chrysolepis*, *Notholithocarpus*, and *Quercus*) by selecting 20 representatives of each non-*Quercus* genus and each section of *Quercus* (for genera/sections with more than 20 sampled species); we selected species that maximized data quality (i.e., the most loci recovered from assembly) as well as phylogenetic breadth for each genus/section. We only tested the four-taxon combinations of the above six genera due to the strong cytonuclear discordance observed particularly in Quercoideae. *Trigonobalanus doichangensis*, a member of *Trigonobalanus* that is sister to the rest of Quercoideae, was selected as the outgroup, and the concatenated nuclear alignment was used for site counts. Second, we used a script "DFOIL_Picker.R" of Ex$D_{\mathrm{FOIL}}$ to yield all unique combinations of four taxa that are arranged in a symmetric tree based on the reduced 150-taxa list and the dated nuclear MCC tree. We removed those combinations with sampling of taxa from the same genus (or, for *Quercus,* sampling of the same section) because our focus was on deep introgression. We also conducted analyses using the full RNA-2821RT dataset given its more limited taxon sampling. In total, 5,316,699 and 224,998 $D_{\mathrm{FOIL}}$ tests were performed for the Hyb-Seq and transcriptome datasets, respectively.

### Phylogenetic network analyses

We conducted phylogenetic network analyses in PhyloNet v.3.8.2 (Than et al., 2008; Wen et al., 2018). Given the intensive computational requirements of these analyses, three strategies were used to generate three separate reduced-representation datasets. One strategy involved selecting two species of each genus of Fagaceae and one outgroup (hereafter "planA"). The second strategy involved reducing our taxon sampling to one species of each non-*Quercus* genus of Fagaceae, one species of each section in *Quercus,* and one outgroup (hereafter "planB"). The third strategy involved selecting two species of each section in *Quercus* and one outgroup (hereafter "planC"). The first two strategies were aimed at identifying possible reticulation between oaks and other genera, while the third one was aimed at identifying possible deep reticulation within oaks. Rooted gene trees were pruned from the HYB-114RT and RNA-4853RT datasets (as these two datasets included the most loci of the Hyb-Seq and transcriptome datasets, respectively) to generate each reduced-representation dataset (i.e., planA, planB, and planC) using a custom R script. Only pruned gene trees with >33% of the species for each reduced taxon set were used for PhyloNet analyses. For each dataset, we ran five network searches allowing one to six reticulation events and 10 runs using a maximum pseudo-likelihood approach (command "InferNetwork_MPL"). We inferred the optimal number of reticulations (0 to 6, with the strictly bifurcating species tree used to represent 0 reticulation; the likelihood score of this tree was calculated using the command "CalGTProb" in PhyloNet) based on the lowest Akaike information criterion (AIC) correction (Sugiura, 1978) and Bayesian information criterion (Schwarz, 1978). For these model fit statistics, the number of parameters was set to the sum of the number of branch lengths and inheritance probabilities being estimated, and the sample size was set to the number of gene trees for finite sample correction.

## Ancestral range estimations

We delimited seven biogeographic regions based on extant and extinct distributions of Fagaceae species: East Asia, South and Southeast Asia, Central and Western Asia, Europe (including a portion of North Africa), North America (north of Mexico), tropical America (Mexico to northern South America), and southern South America (Figure 4). The estimates of ancestral biogeographic range were conducted in the R package

"BioGeoBEARS" (Matzke, 2013) under the DEC, DIVALIKE, and BAYAREALIKE models, using the extant-only MCC tree. The three biogeographic models were compared using AIC to select the optimal one (i.e., the DEC model; Table S10). To assess the impact of fossil inclusion and the phylogenetic uncertainty of fossil placements on ancestral range estimation, we also performed the ancestral range estimation under the DEC model with the extant–extinct MCC tree and 200 extant–extinct trees randomly sampled from the BEAST output. The biogeographic region of each extinct taxon was scored using two different schemes, based on the oldest record and all records of that fossil taxon (Supplementary Materials). The maximum ancestral range size was constrained to three regions, except for the analysis with extinct taxa range being scored with all fossil records that was set as four, reflecting the maximum number of observed regions in extant and extinct species.

## Ancestral niche estimations

We first collected occurrence records for each sampled extant species from GBIF (https://www.gbif.org/), iDigBio (https://www.idigbio.org/), NSII (http://nsii.org.cn/2017/home-en.php), and the literature (last accessed 16 November 2021). For each species, we removed records that were cultivated and outside the native distribution and only retained one record per 1-km grid cell (i.e., 30-s resolution). This resulted in 187,117 occurrences for subsequent analyses. Fourteen Fagaceae species with fewer than three records after cleaning were excluded from the ancestral niche reconstructions. We assembled 35 environmental layers at 30-s resolution from WorldClim (https://worldclim.org/data/worldclim21.html), SoilGrids1km (https://soilgrids.org/), GTOPO30 (https://lta.cr.usgs.gov/GTOPO30), and EarthEnv (https://www.earthenv.org/landcover). We extracted environmental data from these 35 layers and cleaned occurrences for 409 extant Fagaceae taxa using the R package "Raster" (http://CRAN.R-project.org/package=raster). After the removal of highly correlated variables (Pearson's $r \geq 0.7$) using the R package "PKUss" (https://github.com/filBe87/PKUss/), we were left with 12 representative variables covering aspects of climate, topography, soil, and land cover: mean annual temperature, temperature annual range, mean annual precipitation, precipitation seasonality, elevation, slope, bulk density, mean pH, mean organic carbon content, evergreen broadleaf landcover percentage, deciduous broadleaf landcover percentage, and mixed trees landcover percentage.

To integrate the statistical distribution of niche space occupancy of all species, predicted niche occupancy profiles (PNOs, Evans et al., 2009) were applied to ancestral niche estimation using the "contMap" function from the R package "phytools" (Revell, 2012), implemented using a custom R script. A PNO profile is a density histogram that integrates cumulative probabilities of suitability for each value on a single environmental layer. Here, for each species and layer, we generated a PNO profile based on observed environmental values—that were extracted directly from environmental layers underlying the occurrence records—under a uniform distribution (i.e., the probability of each environmental value is $1/n$ where $n$ is the number of occurrences). Briefly, we first proportionally sampled environmental values in 100 replicates from each PNO profile for each present-day species and variable and subsequently used each sample as the observed value of each species in a single ancestral niche estimation analysis, leading to 1,200 ancestral reconstruction analyses. Each analysis was run on the extant-only MCC tree. To reduce the dimension in environmental niche space for visualization, we extracted the estimated ancestral niche for focal lineages from reconstruction results (i.e., 1,200 estimated values for each lineage; 12 variables × 100 replicates). The first two components were then shown through a principal component analysis on the estimated ancestral niche of each lineage.

To assess the impact of fossil inclusion, we applied the same approach to the ancestral niche reconstructions on the extant–extinct MCC tree. We did not conduct analyses on the 200 randomly sampled extant–extinct trees as was done in ancestral range reconstructions, given the heavy computation burden that would be required and similar results from a preliminary analysis. We here used eight paleoclimatic variables from paleoclimatic models. These variables are mean annual temperature, warmest month mean surface air temperature, coldest month mean surface air temperature, warmest month–coldest month temperature difference, mean annual precipitation, wettest month precipitation, driest month precipitation, and wettest month–driest month precipitation difference. For each extant species, the PNO profile was generated from the extant-only occurrences dataset and eight modern climate layers (at 30-s resolution) that were downloaded or calculated from WorldClim. For each extinct taxon, we first converted its modern coordinates to paleo-coordinates using the "reconstruct" function from the R package "chronosphere" (Kocsis and Raja, 2020). The eight paleoclimate layers for 26 time points (from 127.2 to 0 Ma; at 5-min resolution) were derived from Valdes et al. (2021). We extracted paleoclimate values for each paleo-coordinate from the corresponding paleoclimate layers (i.e., closest in geological time). Two strategies were then used to generate a PNO profile for each extinct taxon, based on the paleoclimate values of the oldest record and all records of that fossil taxon.

## Paleoecological niche modeling

We also used fossil-based PaleoENM analysis (Meseguer et al., 2015; Myers et al., 2015; Li et al., 2022) to project the potential distribution of oaks and relatives into the past. We first compiled comprehensive Paleogene fossil distribution data for oaks and relatives from the literature and online resources (CAD, Xing et al., 2016; PBDB, https://paleobiodb.org) (last accessed 21 November 2022). Fossil data were binned by major lineages (i.e., *Castanea*, *Castanopsis*, *Chrysolepis*, *Lithocarpus*, *Notholithocarpus*, subg. *Cerris*, and subg. *Quercus*), based on determinations from the literature involving comparisons with nearest-living relatives (obtained from the literature or the Paleoflora database, http://www.palaeoflora.de/). Fossil data with ambiguous affinities were excluded from further analysis (e.g., the oldest oak pollen fossil from western Europe). When not provided in the literature, modern coordinates of fossils were obtained from Google Earth based on the described fossil localities. The final fossil dataset included 467 occurrences, representing 70 records of *Castanea*, 75 records of *Castanopsis*, one record of *Chrysolepis*, 17 records of *Lithocarpus*, one record of *Notholithocarpus*, 187 records of subg. *Cerris*, and 116 records of subg. *Quercus* (Supplementary Materials).

The PaleoENM analysis was performed following the workflow of Meseguer et al. (2015). We converted modern coordinates of all cleaned fossil occurrences to paleocoordinates and accordingly extracted paleoclimate values from the same eight paleoclimate variables using the same methods as in the extant–extinct niche reconstructions. These paleoclimate values were used for multiple comparisons of niches among oak lineages and closely related genera during the Paleogene using the R package "agricolae" (de Mendiburu and Yaseen, 2020). In addition, given the requirement of adequate fossil occurrences to model past distribution and the global temperature change in the Paleogene, we used two strategies to estimate past climate tolerance for each lineage: (i) by grouping fossil records of the Paleocene and Eocene into one time slice (hereafter "PE"); and (ii) by grouping fossil records of the Paleocene, Eocene, and Oligocene into one time slice (hereafter "PEO") to estimate a single past climate tolerance. A past climate tolerance is the climate range in shared grid cells that satisfies the requirement of each climatic variable—within the range of maximum and minimum values of each variable. Assuming niche conservatism through the Paleogene, we subsequently projected the past climate tolerance of each lineage into a suite of paleoclimate models (i.e., layers), covering seven stages in the Paleogene: Early Paleocene (66.0 Ma), Late Paleocene (60.6 Ma), Early Eocene (55.8 Ma), Middle Eocene (44.5 Ma), Late Eocene (35.9 Ma), Early Oligocene (31.0 Ma), and Late Oligocene (25.6 Ma). The Mahalanobis distance (MD) score was calculated to represent the environmental suitability in each grid cell. The MD score was scaled from 0 to 1; a smaller MD score indicates an area with higher suitability and vice versa. The PaleoENM analysis was not performed for *Chrysolepis* and *Notholithocarpus* due to the limited occurrences; the past distributions of these two genera were represented directly by fossil distribution maps.

### Data availability statement

The resulting DNA alignments, trees, and custom Python and R scripts are available at GitHub (https://github.com/ShuiyinLIU/oaks_ancient_hybri). Raw sequence data are available at the National Center for Biotechnology Information Sequence Read Archive (transcriptomes in PRJNA910851; Hyb-Seq data in PRJNA913201) and the National Genomics Data Center Genome Sequence Archive (PRJCA029210). Paleoclimate models can be accessed at the University of Bristol, BRIDGE repository (https://www.paleo.bristol.ac.uk/ummodel/scripts/papers/Valdes_et_al_2021.html).

## ACKNOWLEDGEMENTS

This work was funded by the National Natural Science Foundation of China with a key international (regional) cooperative research project (No. 31720103903 to T.-S.Y. and D.E.S), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (grant No. XDB31000000 to D.-Z.L.), the Science and Technology Basic Resources Investigation Program of China (2019FY100900 to H.-L.L.), the National Natural Science Foundation of China (No. 32270247 to R.Z.), the Yunling International High-end Experts Program of Yunnan Province, China (grant No. YNQR-GDWG-2017-002 to P.S.S. and T.-S. Y. and YNQR-GDWG-2018-012 to D.E.S. and T.-S.Y.), the USA Department of Energy (grant DE-SC0018247 to P.S.S., D.E.S., R.P.G., and R.A.F.), the USA National Science Foundation (grant DEB-1916632 to R.A.F., R.P.G., D.E.S., and P.S.S.), the Natural Environment Research Council of the UK (grant No. NE/X015505/1 to A.F. and P.J.V.), and Leverhulme Research Project (grant RPG-2019-365 to A.F. and P.J.V.). We thank the following herbaria and their curators and staff for kindly providing herbarium or silica-dried materials: Botanical Research Institute of Texas; California Academy of Sciences; Field Museum of Natural History; Harvard University; Kunming Institute of Botany, CAS; Missouri Botanical Garden; The Ohio State University; Smithsonian Institution; New York Botanical Garden; and University of Texas at Austin. We thank Dr. Lin Bai and Jiajin Wu for help with sampling. We thank the Germplasm Bank of Wild Species at the Kunming Institute of Botany, including the iFlora high performance computing center, for facilitating this study. We thank anonymous reviewers for their helpful suggestions for improving this manuscript.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## AUTHOR CONTRIBUTIONS

S.L., T.Y., R.A.F., G.W.S., P.S.S., D.E.S., and R.P.G. designed the research. All authors contributed to the taxon sampling. S.L., T.Q., and H.R.K. conducted the DNA extraction. Y.Y. performed the assembly of transcriptomes, and S.L. conducted the remaining data analysis. S.L. wrote the first draft. All authors read, revised, and approved the manuscript.

**Edited by:** Jianquan Liu, Lanzhou University, China

**Received** Dec. 20, 2023; **Accepted** Aug. 16, 2024; **Published** Sep. 19, 2024

**OO:** OnlineOpen

## REFERENCES

**Acosta, M.C., and Premoli, A.C.** (2010). Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). Mol. Phylogenet. Evol. **54:** 235–242.

**An, M., Deng, M., Zheng, S.S., Jiang, X.L., and Song, Y.G.** (2017). Introgression threatens the genetic diversity of *Quercus austrocochinchinensis* (Fagaceae), an endangered oak: A case inferred by molecular markers. Front. Plant Sci. **8:** 229.

**Antonelli, A., Kissling, W.D., Flantua, S.G.A., Bermúdez, M.A., Mulch, A., Muellner-Riehl, A.N., Kreft, H., Linder, H.P., Badgley, C., Fjeldså, J., et al.** (2018). Geological and climatic influences on mountain biodiversity. Nat. Geosci. **11:** 718–725.

**Barrón, E., Averyanova, A., Kvaček, Z., Momohara, A., Pigg, K.B., Popova, S., Postigo-Mijarra, J.M., Tiffney, B.H., Utescher, T., and Zhou, Z.K.** (2017). The fossil history of *Quercus*. In *Tree physiology. Oaks physiological ecology. Exploring the functional diversity of genus Quercus L*. Gil-Pelegrín, E., Peguero-Pina, J.J., Sancho-Knapik, D., eds, (Cham, Switzerland: Springer), pp. 39–105.

**Birky, Jr., C.W.** (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc. Natl. Acad. Sci. U. S. A. **92:** 11331–11338.

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

**Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J.** (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. **10:** e1003537.

**Burger, W.C.** (1975). The species concept in *Quercus*. Taxon **24:** 45–50.

**Cai, L., Xi, Z., Lemmon, E.M., Lemmon, A.R., Mast, A., Buddenhagen, C.E., Liu, L., and Davis, C.C.** (2021). The perfect storm: Gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. Syst. Biol. **70:** 491–507.

**Cannon, C.H., Brendel, O., Deng, M., Hipp, A.L., Kremer, A., Kua, C.S., Plomion, C., Romero-Severson, J., and Sork, V.L.** (2018). Gaining a global perspective on Fagaceae genomic diversification and adaptation. New Phytol. **218:** 894–897.

**Cannon, C.H., and Petit, R.J.** (2020). The oak syngameon: More than the sum of its parts. New Phytol. **226:** 978–983.

**Crowl, A.A., Manos, P.S., McVay, J.D., Lemmon, A.R., Lemmon, E.M., and Hipp, A.L.** (2020). Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). New Phytol. **226:** 1158–1170.

**Curtu, A.L., Gailing, O., and Finkeldey, R.** (2007). Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. BMC Evol. Biol. **7:** 218.

**Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al.** (2021). Twelve years of SAMtools and BCFtools. Gigascience **10:** giab008.

**Davidson, N.M., and Oshlack, A.** (2014). Corset: Enabling differential gene expression analysis for *de novo* assembled transcriptomes. Genome Biol. **15:** 410.

**Deng, M., Hipp, A.L., Song, Y.G., Li, Q.S., Coombes, A., and Cotton, A.** (2014). Leaf epidermal features of *Quercus* subgenus *Cyclobalanopsis* (Fagaceae) and their systematic significance. Bot. J. Linn. Soc. **176:** 224–259.

**Deng, M., Jiang, X.L., Hipp, A.L., Manos, P.S., and Hahn, M.** (2018). Phylogeny and biogeography of East Asian evergreen oaks (*Quercus* section *Cyclobalanopsis*; Fagaceae): Insights into the Cenozoic history of evergreen broad-leaved forests in subtropical Asia. Mol. Phylogenet. Evol. **119:** 170–181.

**Denk, T., Grimm, G.W., Manos, P.S., Deng, M., and Hipp, A.L.** (2017). An updated infrageneric classification of the oaks: Review of previous taxonomic schemes and synthesis of evolutionary patterns. In *Tree physiology. Oaks physiological ecology. Exploring the functional diversity of genus Quercus L*. Gil-Pelegrín, E., Peguero-Pina, J.J., Sancho-Knapik, D., eds, (Cham, Switzerland: Springer), pp. 13–38.

**Dodd, R.S., and Afzal-Rafii, Z.** (2007). Selection and dispersal in a multispecies oak hybrid zone. Evolution **58:** 261–269.

**Doyle, J.J.** (2022). Defining coalescent genes: Theory meets practice in organelle phylogenomics. Syst. Biol. **71:** 476–489.

**Doyle, J.J., and Doyle, J.L.** (1987). Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem. Bull. **19:** 11–15.

**Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M.** (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. **28:** 2239–2252.

**Eaton, D.A., Hipp, A.L., Gonzalez-Rodriguez, A., and Cavender-Bares, J.** (2015). Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution **69:** 2587–2601.

**Evans, M.E.K., Smith, S.A., Flynn, R.S., and Donoghue, M.J.** (2009). Climate, niche evolution, and diversification of the “bird-cage” evening primroses (*Oenothera*, sections *Anogra* and *Kleinia*). Am. Nat. **173:** 225–240.

**Feng, L., Zheng, Q.J., Qian, Z.Q., Yang, J., Zhang, Y.P., Li, Z.H., and Zhao, G.F.** (2016). Genetic structure and evolutionary history of three alpine sclerophyllous oaks in East Himalaya-Hengduan Mountains and adjacent regions. Front. Plant Sci. **7:** 1688.

**Folk, R.A., Gaynor, M.L., Engle-Wrye, N.J., O'Meara, B.C., Soltis, P.S., Soltis, D.E., Guralnick, R.P., Smith, S.A., Grady, C.J., Okuyama, Y., et al.** (2023). Identifying climatic drivers of hybridization with a new ancestral niche reconstruction method. Syst. Biol. **72:** 856–873.

**Folk, R.A., Kates, H.R., LaFrance, R., Soltis, D.E., Soltis, P.S., and Guralnick, R.P.** (2021). High-throughput methods for efficiently building massive phylogenies from natural history collections. Appl. Plant Sci. **9:** e11410.

**Folk, R.A., Mandel, J.R., and Freudenstein, J.V.** (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. Syst. Biol. **66:** 320–337.

**Folk, R.A., Visger, C.J., Soltis, P.S., Soltis, D.E., and Guralnick, R.P.** (2018). Geographic range dynamics drove ancient hybridization in a lineage of angiosperms. Am. Nat. **192:** 171–187.

**Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W.** (2012). CD-HIT: Accelerated for clustering the nextgeneration sequencing data. Bioinformatics **28:** 3150–3152.

**Fu, X.G., Liu, S.Y., Velzen, R.V., Stull, G.W., Tian, Q., Li, Y.X., Folk, R.A., Guralnick, R.P., Kates, H.R., Jin, J.J., et al.** (2023). Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-nuclear discordance and provides new insights into generic relationships. J. Syst. Evol. **61:** 806–826.

**García, N., Folk, R.A., Meerow, A.W., Chamala, S., Gitzendanner, M.A., Oliveira, R.S., Soltis, D.E., and Soltis, P.S.** (2017). Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). Mol. Phylogenet. Evol. **111:** 231–247.

**Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.** (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29:** 644–652.

**Grímsson, F., Grimm, G.W., Zetter, R., and Denk, T.** (2016). Cretaceous and Paleogene Fagaceae from North America and Greenland: Evidence for a late Cretaceous split between *Fagus* and the remaining Fagaceae. Acta Palaeobot. **56:** 247–305.

**Grímsson, F., Zetter, R., Grimm, G.W., Pedersen, G.K., Pedersen, A.K., and Denk, T.** (2015). Fagaceae pollen from the early Cenozoic of West

Greenland: Revisiting Engler's and Chaney's arcto-tertiary hypotheses. Plant Syst. Evol. **301:** 809–832.

**Grömping, U.** (2006). Relative importance for linear regression in R: The package relaimpo. J. Stat. Softw. **17:** 43893.

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

**Hai, L., Li, X.Q., Zhang, J.B., Xiang, X.G., Li, R.Q., Jabbour, F., Ortiz, R.D.C., Lu, A.M., Chen, Z.D., and Wang, W.** (2022). Assembly dynamics of East Asian subtropical evergreen broadleaved forests: New insights from the dominant Fagaceae trees. J. Integr. Plant Biol. **64:** 2126–2134.

**Hardin, J.W.** (1975). Hybridization and introgression in *Quercus alba*. J. Arnold Arbor. **56:** 336–363.

**Heath, T.A., Huelsenbeck, J.P., and Stadler, T.** (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. U.S.A. **111:** E2957–E2966.

**Hipp, A.L.** (2015). Should hybridization make us skeptical of the oak phylogeny. Int. Oaks **26:** 9–18.

**Hipp, A.L., Manos, P.S., Hahn, M., Avishai, M., Bodenes, C., Cavender-Bares, J., Crowl, A.A., Deng, M., Denk, T., Fitz-Gibbon, S., et al.** (2020). Genomic landscape of the global oak phylogeny. New Phytol. **226:** 1198–1212.

**Hubert, F., Grimm, G.W., Jousselin, E., Berry, V., Franc, A., and Kremer, A.** (2014). Multiple nuclear genes stabilize the phylogenetic backbone of the genus *Quercus*. Syst. Biodivers. **12:** 405–423.

**Jiang, X.L., Hipp, A.L., Deng, M., Su, T., Zhou, Z.K., and Yan, M.X.** (2019). East Asian origins of European holly oaks (*Quercus* section *Ilex* Loudon) via the Tibet-Himalaya. J. Biogeogr. **46:** 2203–2214.

**Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C., and Wickett, N.J.** (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. **4:** 1600016.

**Kates, H.R., O'Meara, B.C., LaFrance, R., Stull, G.W., James, E.K., Liu, S.Y., Tian, Q., Yi, T.S., Conde, D., Kirst, M., et al.** (2024). Shifts in evolutionary lability underlie independent gains and losses of root-nodule symbiosis in a single clade of plants. Nat. Commun. **15:** 4262.

**Kocsis, A.T., and Raja, N.B.** (2020). chronosphere: Earth system history variables. Available from: https://cran.r-project.org/web/packages/chronosphere/index.html

**Kremer, A., and Hipp, A.L.** (2020). Oaks: An evolutionary success story. New Phytol. **226:** 987–1011.

**Kvaček, Z., and Walther, H.** (1989). Paleobotanical studies in Fagaceae of the European Tertiary. Plant Syst. Evol. **162:** 213–229.

**Lambert, S.M., Streicher, J.W., Fisher-Reid, M.C., Mendez de la Cruz, F.R., Martinez-Mendez, N., Garcia-Vazquez, U.O., Nieto Montes de Oca, A., and Wiens, J.J.** (2019). Inferring introgression using RADseq and $D_{\mathrm{FOIL}}$: Power and pitfalls revealed in a case study of spiny lizards (*Sceloporus*). Mol. Ecol. Resour. **19:** 818–837.

**Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., and Calcott, B.** (2016). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. **34:** 772–773.

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

**Larson-Johnson, K.** (2016). Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. New Phytol. **209:** 418–435.

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

**Leroy, T., Rougemont, Q., Dupouey, J.L., Bodenes, C., Lalanne, C., Belser, C., Labadie, K., Le Provost, G., Aury, J.M., Kremer, A., et al.** (2020b). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. New Phytol. **226:** 1183–1197.

**Li, H.** (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997

**Li, W.C., Huang, J., Chen, L.L., Spicer, R.A., Li, S.F., Liu, J., Gao, Y., Wu, F.X., Farnsworth, A., Valdes, P.J., et al.** (2022). *Podocarpium* (Fabaceae) from the late Eocene of central Tibetan Plateau and its biogeographic implication. Rev. Palaeobot. Palynol. **305:** 104745.

**Mallet, J.** (2007). Hybrid speciation. Nature **446:** 279–283.

**Manos, P.S., Cannon, C.H., and Oh, S.H.** (2008). Phylogenetic relationships and taxonomic status of the paleoendemic Fagaceae of western North America: Recognition of a new genus, *Notholithocarpus*. Madroño **55:** 181–190.

**Manos, P.S., Doyle, J.J., and Nixon, K.C.** (1999). Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* Subgenus *Quercus* (Fagaceae). Mol. Phylogenet. Evol. **12:** 333–349.

**Manos, P.S., Zhou, Z.K., and Cannon, C.H.** (2001). Systematics of Fagaceae: Phylogenetic tests of reproductive trait evolution. Int. J. Plant Sci. **162:** 1361–1379.

**Mason, A.J., Grazziotin, F.G., Zaher, H., Lemmon, A.R., Lemmon, E.M., and Parkinson, C.L.** (2019). Reticulate evolution in nuclear Middle America causes discordance in the phylogeny of palm-pitvipers (Viperidae: *Bothriechis*). J. Biogeogr. **46:** 833–844.

**Matzke, N.J.** (2013). BioGeoBEARS: BioGeography with Bayesian and likelihood evolutionary analysis in R scripts [online]. Available from: http://cran.r-project.org/web/packages/BioGeoBEARS/

**de Mendiburu, F., and Yaseen, M.** (2020). agricolae: Statistical procedures for agricultural research. Available from: https://cran.r-project.org/package=agricolae

**Meseguer, A.S., Lobo, J.M., Ree, R., Beerling, D.J., and Sanmartín, I.** (2015). Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of *Hypericum* (Hypericaceae). Syst. Biol. **64:** 215–232.

**Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R.** (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37:** 1530–1534.

**Mir, C., Toumi, L., Jarne, P., Sarda, V., Di Giusto, F., and Lumaret, R.** (2006). Endemic North African *Quercus afares* Pomel originates from hybridisation between two genetically very distant oak species (*Q. suber* L. and *Q. canariensis* Willd.): Evidence from nuclear and cytoplasmic markers. Heredity **96:** 175–184.

**Morales-Briones, D.F., Gehrke, B., Huang, C.H., Liston, A., Ma, H., Marx, H.E., Tank, D.C., and Yang, Y.** (2021). Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in *Alchemilla* s.l. (Rosaceae). Syst. Biol. **71:** 190–207.

**Morales-Briones, D.F., Liston, A., and Tank, D.C.** (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytol. **218:** 1668–1684.

**Moran, E.V., Willis, J., and Clark, J.S.** (2012). Genetic evidence for hybridization in red oaks (*Quercus* sect. *Lobatae*, Fagaceae). Am. J. Bot. **99:** 92–100.

**Myers, C., Stigall, A., and Lieberman, B.** (2015). PaleoENM: Applying ecological niche modeling to the fossil record. Paleobiology **41:** 226–244.

**Nixon, K.C.** (2006). Global and neotropical distribution and diversity of oak (Genus *Quercus*) and oak forests. In *Ecology and conservation of neotropical montane oak forests. Ecological studies*. Kappelle, M., ed., (Berlin, Heidelberg: Springer), pp. 3–12.

**Pavón-Vázquez, C.J., Brennan, I.G., Keogh, J.S., and Solis-Lemus, C.** (2021). A comprehensive approach to detect hybridization sheds light on the evolution of earth's largest lizards. Syst. Biol. **70:** 877–890.

**Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E., and Smith, S.A.** (2018). Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. **105:** 385–403.

**Pease, J.B., and Hahn, M.W.** (2015). Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. **64:** 651–662.

**Petit, R., Bodénès, C., Ducousso, A., Roussel, G., and Kremer, A.** (2003). Hybridization as a mechanism of invasion in oaks. New Phytol. **161:** 151–164.

**Pham, K.K., Hipp, A.L., Manos, P.S., and Cronn, R.C.** (2017). A time and a place for everything: Phylogenetic history and geography as joint predictors of oak plastome phylogeny. Genome **60:** 720–732.

**Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A.** (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. **67:** 901–904.

**Revell, L.J.** (2012). Phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. **3:** 217–223.

**Rieseberg, L.H., and Soltis, D.E.** (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants **5:** 65–84.

**Rose, J.P., Toledo, C.A.P., Lemmon, E.M., Lemmon, A.R., and Sytsma, K.J.** (2021). Out of sight, out of mind: Widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. Syst. Biol. **70:** 162–180.

**Sayyari, E., and Mirarab, S.** (2016). Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. **33:** 1654–1668.

**Sayyari, E., and Mirarab, S.** (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes (Basel) **9:** 132.

**Schnitzler, J.P., Steinbrecher, R., Zimmer, I., Steigner, D., and Fladung, M.** (2004). Hybridization of European oaks (*Quercus ilex* × *Q. robur*) results in a mixed isoprenoid emitter type. Plant Cell Environ. **27:** 585–593.

**Schwarz, G.** (1978). Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

**Shen, X.X., Hittinger, C.T., and Rokas, A.** (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. **1:** 126.

**Shimodaira, H.** (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51:** 492–508.

**Simeone, M.C., Grimm, G.W., Papini, A., Vessella, F., Cardoni, S., Tordoni, E., Piredda, R., Franc, A., and Denk, T.** (2016). Plastome data reveal multiple geographic origins of *Quercus* Group *Ilex*. PeerJ **4:** e1897.

**Siniscalchi, C.M., Correa-Narvaez, J., Kates, H.R., Soltis, D.E., Soltis, P.S., Guralnick, R.P., Manchester, S.R., and Folk, R.A.** (2023). Fagalean phylogeny in a nutshell: Chronicling the diversification history of Fagales. bioRxiv. https://doi.org/10.1101/2023.03.06.531381

**Smith, S.A., Brown, J.W., and Walker, J.F.** (2018). So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. PLoS ONE **13:** e0197433.

**Smith, S.A., Moore, M.J., Brown, J.W., and Yang, Y.** (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. **15:** 150.

**Smith, S.A., and O'Meara, B.C.** (2012). treePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics **28:** 2689–2690.

**Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M., and Kelly, S.** (2016). TransRate: Reference-free quality assessment of *de-novo* transcriptome assemblies. Genome Res. **26:** 1134–1144.

**Song, L., and Florea, L.** (2015). Rcorrector: Efficient and accurate error correction for Illumina RNAseq reads. Gigascience **4:** 48.

**Stamatakis, A.** (2014). RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:** 1312–1313.

**Stull, G.W., Pham, K.K., Soltis, P.S., and Soltis, D.E.** (2023). Deep reticulation: The long legacy of hybridization in vascular plant evolution. Plant J. **114:** 743–766.

**Stull, G.W., Soltis, P.S., Soltis, D.E., Gitzendanner, M.A., and Smith, S.A.** (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. Am. J. Bot. **107:** 790–805.

**Sugiura, N.** (1978). Further analysts of the data by akaike's information criterion and the finite corrections. Commun. Stat. Theor. Methods **7:** 13–26.

**Sukumaran, J., and Holder, M.T.** (2010). DendroPy: A Python library for phylogenetic computing. Bioinformatics **26:** 1569–1571.

**Suvorov, A., Scornavacca, C., Fujimoto, M.S., Bodily, P., Clement, M., Crandall, K.A., Whiting, M.F., Schrider, D.R., and Bybee, S.M.** (2022). Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. Syst. Biol. **71:** 526–546.

**Than, C., Ruths, D., and Nakhleh, L.** (2008). PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary histories. BMC Bioinform. **9:** 322.

**Thiers, B.** (2016). Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/

**Tian, Q., Stull, G.W., Kellermann, J., Medan, D., Nge, F.J., Liu, S.Y., Kates, H.R., Soltis, D.E., Soltis, P.S., Guralnick, R.P., et al.** (2024). Rapid in-situ diversification rates in Rhamnaceae explain the parallel evolution of high diversity in temperate biomes from global to local scales. New Phytol. **241:** 1851–1865.

**Toews, D.P.L., and Brelsford, A.** (2012). The biogeography of mitochondrial and nuclear discordance in animals. Mol. Ecol. **21:** 3907–3930.

**Valdes, P.J., Scotese, C.R., and Lunt, D.J.** (2021). Deep ocean temperatures through time. Clim. Past **17:** 1483–1506.

**van Dongen, S.M.** (2000). *Graph Clustering by Flow Simulation*. PhD Dissertation. Ultrecht University.

**Wen, D.Q., Yu, Y., Zhu, J.F., and Nakhleh, L.** (2018). Inferring phylogenetic networks using PhyloNet. Syst. Biol. **67:** 735–740.

**Whittemore, A.T., and Schaal, B.A.** (1991). Interspecific gene flow in sympatric oaks. Proc. Natl. Acad. Sci. U.S.A. **88:** 2540–2544.

**Xing, Y.W., Gandolfo, M.A., Onstein, R.E., Cantrill, D.J., Jacobs, B.F., Jordan, G.J., Lee, D.E., Popova, S., Srivastava, R., Su, T., et al.** (2016). Testing the biases in the rich Cenozoic Angiosperm macrofossil record. Int. J. Plant Sci. **177:** 371–388.

**Yang, C.X., Liu, S.Y., Zerega, N.J.C., Stull, G.W., Gardner, E.M., Tian, Q., Gu, W., Lu, Q., Folk, R.A., Kates, H.R., et al.** (2023a). Phylogeny and biogeography of *Morus* (Moraceae). Agronomy **13:** 2021.

**Yang, Y., and Smith, S.A.** (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. **31:** 3081–3092.

**Yang, Y.C., Zhou, T., Duan, D., Yang, J., Feng, L., and Zhao, G.F.** (2016). Comparative analysis of the complete chloroplast genomes of five *Quercus* species. Front. Plant Sci. **7:** 959.

**Yang, Y.Y., Qu, X.J., Zhang, R., Stull, G.W., and Yi, T.S.** (2021). Plastid phylogenomic analyses of Fagales reveal signatures of conflict and ancient chloroplast capture. Mol. Phylogenet. Evol. **163:** 107232.

**Yang, Y.Y., Stull, G.W., Qu, X.J., Zhao, L., Hu, Y., Wang, Z.H., Ma, H., Li, D.Z., Smith, S.A., and Yi, T.S.** (2023b). Genome duplications, genomic conflict, and rapid phenotypic evolution characterize the Cretaceous radiation of Fagales. bioRxiv. https://doi.org/10.1101/2023.06.11.544004

**Yu, J.R., Niu, Y.T., You, Y.C., Cox, C.J., Barrett, R.L., Trias-Blasi, A., Guo, J., Wen, J., Lu, L.M., and Chen, Z.D.** (2022). Integrated phylogenomic analyses unveil reticulate evolution in *Parthenocissus* (Vitaceae), highlighting speciation dynamics in the Himalayan–Hengduan Mountains. New Phytol. **238:** 888–903.

**Zachos, J., Pagani, M., Sloan, L., Thomas, E., and Billups, K.** (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. Science **292:** 686–693.

**Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S.** (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. **19:** 153.

**Zhou, B.F., Yuan, S., Crowl, A.A., Liang, Y.Y., Shi, Y., Chen, X.Y., An, Q.Q., Kang, M., Manos, P.S., and Wang, B.S.** (2022). Phylogenomic analyses highlight innovation and introgression in the continental radiations of Fagaceae across the Northern Hemisphere. Nat. Commun. **13:** 1320.

**Zhou, Z.K.** (1999). Fossils of the Fagaceae and their implications in systematics and biogeography. Acta Phytotax. Sin. **37:** 369–385.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.13773/suppinfo

**Figure S1.** Species tree of oaks and relatives inferred by ASTRAL-III based on the HYB-89MO dataset including 89 loci from 431 taxa
**Figure S2.** Species tree of oaks and relatives inferred by ASTRAL-III based on the HYB-98RT dataset including 98 loci from 431 taxa
**Figure S3.** Maximum likelihood (ML) tree of oaks and relatives inferred by RAxML based on the concatenated supermatrix (the HYB-89MO dataset) including 89 loci from 431 taxa under an unpartitioned GTR-GAMMA model
**Figure S4.** ML tree of oaks and relatives inferred by RAxML based on the concatenated supermatrix (the HYB-98RT dataset) including 98 loci from 431 taxa under an unpartitioned GTR-GAMMA model
**Figure S5.** Species tree of oaks and relatives inferred by ASTRAL-III based on the RNA-977MO dataset including 977 loci from 89 taxa
**Figure S6.** Species tree of oaks and relatives inferred by ASTRAL-III based on the RNA-2821RT dataset including 2821 loci from 89 taxa
**Figure S7.** ML tree of oaks and relatives inferred by RAxML based on the concatenated supermatrix (the RNA-977MO dataset) including 977 loci from 89 taxa under an unpartitioned GTR-GAMMA model
**Figure S8.** ML tree of oaks and relatives inferred by RAxML based the concatenated supermatrix (the RNA-2821RT dataset) including 2821 loci from 89 taxa under an unpartitioned GTR-GAMMA model
**Figure S9.** Levels of intergenic discordance and uninformativeness with respect to major lineages of oaks and other genera in Quercoideae
**Figure S10.** Phyparts results based on the 98 gene trees from the HYB-98RT dataset, mapped against the ASTRAL species tree
**Figure S11.** Phyparts result based on the 2821 gene trees of the RNA-2821RT dataset, mapped against the ASTRAL species tree
**Figure S12.** Cladogram of the chloroplast ML tree of oaks and relatives inferred by RAxML based on the concatenated plastome supermatrix including 223 taxa
**Figure S13.** The distribution of gene-wise and site-wise phylogenetic signal for alternative topologies of five uncertain deep branches based on the HYB-98RT dataset
**Figure S14.** The distribution of gene-wise and site-wise phylogenetic signal for alternative topologies of five uncertain deep branches based on the RNA-2821RT dataset
**Figure S15.** Cophylogeny showing incongruence between the nuclear ASTRAL (left; the HYB-98RT dataset) and chloroplast ML (right) trees with coalescent simulation results from the nuclear guide tree scaled by a factor of four
**Figure S16.** Tanglegram comparing the nuclear ASTRAL (left; the RNA-2821RT dataset) and chloroplast (right) phylogenies optimized in Dendroscope, with coalescent simulation results from the nuclear guide tree scaled by a factor of four
**Figure S17.** Cophylogeny showing incongruence between the nuclear ASTRAL (left; the HYB-98RT dataset) and chloroplast (right) ML trees with coalescent simulation results from the nuclear guide tree scaled by a factor of two
**Figure S18.** Cophylogeny showing incongruence between the nuclear ASTRAL (left; the RNA-2821RT dataset) and chloroplast (right) ML trees with coalescent simulation results from the nuclear guide tree scaled by a factor of two
**Figure S19.** Proportions of positive tests for various introgression signatures visualized on the time-calibrated tree based on the HYB-98RT dataset
**Figure S20.** Proportions of positive tests for various introgression signatures visualized on the time-calibrated tree based on the RNA-2821RT dataset
**Figure S21.** Species networks from reduced-representation data sets of oaks and Fagaceae
**Figure S22.** Ancestral range estimation of Fagaceae under the DEC model, based on the extant-extinct MCC tree, and with the geographic region of fossil taxon being scored by its all fossil record
**Figure S23.** Ancestral range estimation of Fagaceae under the DEC model, based on the 200 randomly sampled extant-extinct trees, and with the geographic region of fossil taxon being scored by its all fossil record
**Figure S24.** Ancestral range estimation of Fagaceae under the DEC model, based on the extant-extinct MCC tree, and with the geographic region of fossil taxon being scored by its oldest fossil record
**Figure S25.** Ancestral range estimation of Fagaceae under the DEC model, based on the 200 randomly sampled extant-extinct trees, and with the geographic region of fossil taxon being scored by its oldest fossil record
**Figure S26.** Ancestral range estimation of Fagaceae under the DEC model, based on the extant-only MCC tree of the Hyb-Seq dataset
**Figure S27.** Ancestral niche estimation of Fagaceae based on the extant-only MCC tree of the Hyb-Seq dataset
**Figure S28.** Ancestral niche estimation of Fagaceae based on the extant-extinct MCC tree, and with the niche of fossil taxon being estimated by its all fossil record
**Figure S29.** Ancestral niche estimation of Fagaceae based on the extant-extinct MCC tree, and with the niche of fossil taxon being estimated by its oldest fossil record
**Figure S30.** Boxplots showing climatic niche space of fossil taxa of *Quercus* lineages and relatives in Quercoideae during the Eocene
**Figure S31.** Boxplots showing climatic niche space from Early Eocene to Late Oligocene of fossil taxa of *Quercus* lineages and relatives in Quercoideae
**Figure S32.** Potential distribution of *Quercus* lineages and relatives in Quercoideae from Late Paleocene to Late Oligocene as inferred by paleoecological niche modeling under PEO strategy
**Figure S33.** Clade frequencies of the chloroplast trees simulated under the coalescent (with the guide tree scaled by a factor of four), mapped on the cladogram of the concatenated ML tree (the HYB-98RT dataset)
**Figure S34.** Clade frequencies of the chloroplast trees simulated under the coalescent (with the guide tree scaled by a factor of two), mapped on the cladogram of the concatenated ML tree (the HYB-98RT dataset)
**Figure S35.** Overview of the Hyb-Seq datasets (red boxes) and analyses conducted in this study
**Figure S36.** Overview of the transcriptome datasets (red boxes) and analyses conducted in this study
**Figure S37.** Overview of plastome datasets (red box) and analyses conducted in this study
**Table S1.** A summary of all datasets used in this study
**Table S2.** The five uncertain deep nodes and their alternative topologies as well as AU tests and polytomy tests
**Table S3.** Number of positive tests from the $D_{\text{FOIL}}$ tests categorized by genera in Quercoideae
**Table S4.** Model selection among species networks recovered for oaks and their relatives based on the Hyb-Seq dataset
**Table S5.** Model selection among species networks recovered for oaks and their relatives based on the transcriptome dataset

**Table S6.** Taxon sampling percentage for each genus of Fagaceae and each section of oaks in the Hyb-Seq, transcriptome, and plastome datasets

**Table S7.** Information for taxon sampling in Hyb-Seq and transcriptome datasets

**Table S8.** Information for taxon sampling in plastome dataset

**Table S9.** Prior settings and estimated ages of nine calibrated nodes in BEAST analysis

**Table S10.** Comparison of three biogeographical models based on the Hyb-Seq dataset