Evolutionary dynamics driving continental radiations of Fagaceae forests across the Northern Hemisphere

Biao-Feng Zhou
South China Botanical Garden, Chinese Academy of Sciences
Shuai Yuan
South China Botanical Garden, Chinese Academy of Sciences
Andrew Crowl
Department of Biology, Duke University
Yi-Ye Liang
South China Botanical Garden, Chinese Academy of Sciences
Yong Shi
South China Botanical Garden, Chinese Academy of Sciences
Xue-Yan Chen
South China Botanical Garden, Chinese Academy of Sciences
Qing-Qing An
South China Botanical Garden, Chinese Academy of Sciences
Ming Kang
South China Botanical Garden, Chinese Academy of Sciences https://orcid.org/0000-0002-4326-7210
Paul Manos
Department of Biology, Duke University
Baosheng Wang (baosheng.wang@scbg.ac.cn)
South China Botanical Garden, Key Laboratory of Plant Resources Conservation and Sustainable Utilization, Chinese Academy of Sciences

Article

Keywords: macroevolutionary events, genome-wide signatures, hybridization

Posted Date: October 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-968321/v1

Version of Record: A version of this preprint was published at Nature Communications on March 14th, 2022. See the published version at https://doi.org/10.1038/s41467-022-28917-1.

Evolutionary dynamics driving continental radiations of Fagaceae forests across the Northern Hemisphere

Biao-Feng Zhou[1,2,5]; ORCID: 0000-0002-2782-4160
Shuai Yuan[1,5]; ORCID: 0000-0002-9851-6384
Andrew A. Crowl[3,5]; ORCID: 0000-0002-1745-0687
Yi-Ye Liang[1]; ORCID: 0000-0002-4992-3784
Yong Shi[1]
Xue-Yan Chen[1]
Qing-Qing An[1]
Ming Kang[1,4]; ORCID: 0000-0002-4326-7210
Paul S. Manos[3*]; ORCID: 0000-0003-3122-1538
Baosheng Wang[1,4*]; ORCID: 0000-0002-0934-1659

Affiliations:

[1]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

[2]University of the Chinese Academy of Sciences, Beijing 100049, China

[3]Department of Biology, Duke University, Durham, NC 27708, USA

[4]Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Guangzhou 510650, China

[5]These authors contributed equally to this work

[*]Corresponding authors: pmanos@duke.edu, baosheng.wang@scbg.ac.cn

Introductory paragraph

Northern Hemisphere forests changed drastically in the early Eocene with the diversification of the oak family (Fagaceae). Cooling climates over the next 20 million years fostered the spread of temperate biomes that became increasingly dominated by oaks and their chestnut relatives. Here we investigate the timing and pattern of major macroevolutionary events and ancient genome-wide signatures of hybridization across Fagaceae. An unparalleled transformation of forest dynamics began with the rapid diversification of major lineages within 15 million years following the K-Pg extinction. Innovations related to seed and pollen dispersal are implicated in triggering waves of continental radiations, while fungal symbioses fortified a competitive edge underground. We detected introgression at multiple time scales, including ancient events predating the origination of genus-level diversity. As oak lineages moved into newly available temperate habitats in the early Miocene, secondary contact between previously isolated species occurred. This resulted in adaptive introgression, further amplifying global proliferation.

Main text

Northern Hemisphere forests and shrublands are now dominated by species comprising temperate and subtropical lineages, marking one of the greatest floristic transitions in the vegetation history of the Cenozoic[1-3]. Paleobotanical reconstructions suggest that a cooling global climate afforded ecological opportunities to plant groups that were physiologically predisposed to disperse into and radiate within broadening and often repeated seasonal biomes across what would become the Americas and Eurasia[4-7]. Central to this pattern of floristic replacement with significant ecological consequence are the roughly 900 species currently recognized within Fagaceae (oak, beech, chestnut, stone oak). Important components of the timing and pattern of macroevolutionary events and the role of ancient hybridization, however, have yet to be sufficiently described across Fagaceae.

The oak family plays a major ecological role in terms of sheer abundance of standing biomass[6,8-16] and a variety of mutualistic associations involving ectomycorrhizal fungi[17-19], gall-forming insects[20-22], and seed-dispersing vertebrates[23-28]. Interactions between Fagaceae and their co-distributed biota suggests degrees of host specificity and the potential for co-evolution, reciprocal diversification, and expansion of range size.

Fossil analogs of modern Fagaceae are well represented in the Northern Hemisphere, indicating long-term presence and differential patterns of diversification[29-37]. Recent studies integrating these fossils within phylogenies of modern taxa have provided essential context to estimate divergence times[38-40]. With a minimum age of ca. 80 million years ago (Ma) and precise aging of new fossilized pollen and macrofossils assigned to some modern groups by 50 Ma, the evolution of major lineages appears to be unusually rapid for forest tree species[41,42]. However, the diversification history of Fagaceae remains incompletely understood, with the exception of modern lineages of *Quercus*[39,43-45]. Therefore, a complete historical account of this continental radiation is needed to bring to light the dynamics of speciation through the genomes of these ecologically important tree species.

Oaks have a long history of divergence in spite of gene flow. Recent estimates of phylogeny using next generation sequencing of nuclear DNA resolve the main oak groups while demonstrating that oak species are generally not of hybrid origin[39,46]. However, more targeted phylogenomic studies have shown that hybridization has left its signature: one of unstable lineages and taxa, the likely result of ancient hybridization, another of intermediate position between parental lineages as expected by recent-generation hybrids[47-50]. New insights into nuclear genomic architecture of hybridization complement various datasets derived from the maternally-inherited plastome and suspected cases of plastome capture and resulting cytoplasmic-nuclear discordance have been shown at various phylogenetic depths in *Quercus*[51-53]. Now the timing and impact of these events within *Quercus*, as well as within and between other lineages, is within reach: chronograms for both genomes along with thorough interrogation of the nuclear genome using modern analytical approaches provides the framework needed to estimate the timing of hybridization events, identify the signatures of gene flow, and detect evidence for adaptive evolution.

Phylogenomic analyses of nuclear and plastid genomes reveal a complex history of divergence and gene flow in deep time across Fagaceae. To test specific hypotheses of ancient hybridization, we constructed time-calibrated phylogenies to pinpoint major divergent and reticulate events across a broad sample of 122 individual plants representing 91 species from all recognized genera, using 2,124 nuclear loci and full plastomes (Supplementary Tables 1 and 2). With these data, we characterize the diversification of Fagaceae and identify admixed genomes due to ancient gene flow within the first complete family-wide phylogenetic context.

Results and Discussion

Time-calibrated phylogeny based on nuclear data

Maximum likelihood (ML) and Bayesian analyses of the concatenated dataset and coalescent analyses using ASTRAL-III and SVDquartets produced similar trees with strong support (BS > 90 and BI > 0.95) for all nodes except a few branches (Fig. 1; Supplementary Fig. 1). All genera of Fagaceae were inferred to be monophyletic with fully resolved interrelationships. Our phylogenetic estimate unambiguously supports three early-diverging lineages of Fagaceae – *Fagus*, *Trigonobalanus*, and two castaneoid lineages, *Castanea* + *Castanopsis* – along with a novel resolution for a crown clade comprising the three remaining castaneoid genera, *Chrysolepis*, *Lithocarpus*, and *Notholithocarpus*, which in turn is sister to *Quercus* (Fig. 1; Supplementary Fig. 1). Resolution of castanoid taxa (*Chrysolepis*, *Lithocarpus*, and *Notholithocarpus*) as sister to *Quercus* settles long-standing questions on the origin of the wind-pollinated oaks: they are derived from insect-pollinated ancestors that already possessed single rounded fruit seated within a valveless cupule[54]. Within *Quercus*, our analyses confirmed the phylogenetic structure resolved by previous studies based on sequences derived from RAD-seq datasets[39,50] and nuclear loci[47]. Despite phylogenetic congruence across methods, high levels of gene-tree conflict within the nuclear genome were observed, likely due to incomplete lineage sorting (ILS; Supplementary Figs. 2, 3 and 4). This would be expected given the rapid evolution of crown clade genera as inferred here (see below).

We constrained nodes with eight fossil calibrations (Fig. 1; Supplementary Table 3) to estimate divergence times and diversification dynamics within Fagaceae. Two early-diverging lineages of Fagaceae originated by the late Cretaceous, with *Fagus* and *Trigonobalanus* diverging at 82.6 Ma (95% CI = 84.2 – 80.1 Ma) and 70.7 Ma (95% CI = 69.1 – 62.4 Ma), respectively (Fig. 1). Subsequent branching events in the early Cenozoic suggest that the six genera (*Castanea*, *Castanopsis*, *Chrysolepis*, *Lithocarpus*, *Notholithocarpus* and *Quercus*; the hypogeous seed or "HS" clade hereafter) that comprise 98.8% (N = 893) of the modern species originated during the Paleocene. The ancestor of the HS clade split at 64.5 Ma (95% CI = 69.1 – 62.4 Ma) followed by the rapid origination of extant genera within a 15 Ma window (Fig. 1). These events closely follow the Cretaceous-Paleogene (K-Pg) boundary dated at 66 Ma[55].

Accelerated diversification following the K-Pg mass extinction event has been documented in plants[56,57], birds[58], frogs[59], fish[60], and mammals[61], most likely a generalized consequence of ecological opportunities following the mass extinction. An increase in speciation rate just after the K-Pg boundary was confirmed for Fagaceae by diversification rate analyses, with a net speciation rate shift detected along the branch leading to the ancestor of the HS clade (Fig. 1). This result is robust to different calibration sets and reference trees for molecular dating (Supplementary Fig. 5).

Ecological correlates of diversification

The HS clade shares the derived feature of hypogeous germination, as defined by the first leaves of the embryo remaining in the seed as storage organs that contribute to enhancing seedling survivorship[26]. This condition is often correlated with larger seeds that are biotically dispersed by various specialized animal groups whereas the two early-branching lineages (*Fagus* and *Trigonobalanus*) share the plesiomorphic condition of smaller seeds and the generalized state of epigeal germination[62]. Previous phylogenetic studies including fossils have revealed several transitions to biotic dispersal across fagalean lineages during its ca. 95 million-year history[63,64]. Biotically dispersed lineages have larger range sizes and higher diversification rates than abiotically dispersed lineages. Innovations associated with seed morphology coincide with an increase in diversification rate of the HS clade after the K-Pg boundary (Fig. 1). Time-calibrated phylogenies of the main groups of modern HS seed dispersers, specifically scatter-hoarding Sciuridae (squirrels), Covidae (jays), and Picidae (woodpeckers) contrast sharply. Evolution of rodent-mediated dispersal closely follows the origin of the HS clade and other large-seeded biotically dispersed fagalean lineages supporting a generalized co-evolution with early-diverging Sciuridae[61,65-67]. In contrast, the relative timing for the diversification of bird lineages associated with the dispersal of HS seed is at least 20 million years later[68,69]. This suggests a second phase of mutualistic response driven by HS seed production generated patterns of co-distribution between granivorous birds, best exemplified in *Quercus*, that likely began during the Miocene[24,70].

Ecological success of Fagaceae is often attributed to symbiosis with at least three main ectomycorrhizal (ECM) lineages of basidiomycetes: Russulales, Boletales, and Agaricales[71]. This mutualism represents an ancient resource-sharing mechanism

that contributes heavily to ecosystem processes and dominance of Fagaceae[17,72,73]. The estimated number of global ECM fungal species is c. 6,000 with Fagaceae accounting for 45% of the associated 2,000 species of host seed plant diversity[71]. While the stem lineages of the main ECM clades date back to the Jurassic, crown clade diversification and inferred shifts in speciation rate occur contemporaneously in many of the lineages associated with Fagaceae[74]. Multiple increases in speciation rate postdate the K-Pg boundary by at least 20 million years, suggesting that transition to Fagaceae forests in the Oligocene may have contributed to species radiations linked to symbiosis. Indeed, secondary increases in speciation rate spanning the Oligocene and Miocene were detected in three clades, *Lithocarpus* from southeast Asia, the Eurasian subclade of section *Quercus*, and section *Lobatae* which is endemic to the Americas (Fig. 1; Supplementary Fig. 5). Previous studies based on global sampling of oak species reported four shifts of diversification during the Miocene[39], including the two events we observed within *Quercus*.

Rapid radiation of the genus *Quercus* is consistent with global temperature cooling associated with the onset of temperate habitats during the Oligocene (Fig. 1A). Our phylogenetic analyses confirm that *Quercus* evolved from within a clade formed by all five insect-pollinated castaneoid genera, and diverged from them approximately 56 Ma (Fig 1; Supplementary Fig. 5). Fossilized pollen assignable to modern oak sections is found at high latitudes well before *Quercus* migrated to middle latitudes[32]. Thus, the origin of wind-pollination in *Quercus* preceded the explosive radiations of oaks in the Oligocene to early Miocene. Shift to wind-pollination alone did not increase the diversification rate of oak species immediately, but instead served as a predisposed neutral change that later facilitated rapid radiation of this genus during the expansion of seasonal climates (Fig. 1). Consistent with this expectation, oaks have their highest species richness in cool-temperate areas in middle latitudes and montane areas at lower latitudes of the Americas, where they form ecologically dominant forests[6,46].

Ancient hybridization explains cytoplasmic-nuclear gene tree conflict

Plastome-based analyses using various phylogenetic methods (ML or BI analyses), data partitioning schemes (un-partitioned or partitioned by gene and codon position), and alignments (nucleotide or amino acid sequences) yielded largely congruent topologies (Supplementary Fig. 6). Major nodes along the backbone of the plastid tree

were highly supported (BS > 80% and BI > 95%) and consistent with the nuclear trees in the resolution of *Fagus* and *Trigonobalanus* as early-branching lineages (Fig. 2). The plastome topology, however, differs markedly from the trees obtained with nuclear loci in regards to the composition and placement of major lineages within the HS clade (Fig. 2). While most plastome subclades comprise related species, several combine disparate taxonomic groups. We failed to recover monophyly of two genera, *Quercus* and *Notholithocarpus*, and five sections of *Quercus* (*Quercus*, *Virentes*, *Ponticae*, *Protobalanus*, *Ilex* and *Cerris*). The structure of the plastome reconstruction within the HS clade is largely geographic, consistent with previous studies[52,53,75,76], with the taxonomic diversity divided into two major clades we treat here as New World (NW) and Old World (OW) (Fig. 2).

The NW-OW pattern recovered in our plastome analyses suggests an early geographic homogenization of cytoplasm across lineages generating the observed cytoplasmic-nuclear discordance at the deepest level of the HS clade (Fig. 2). While the most likely source of cytoplasmic-nuclear discordance is hybridization, incomplete lineage sorting (ILS) could produce a similar pattern. To discriminate between these two hypotheses, we performed coalescent-based simulations. We found the plastid tree discordance to be significantly higher than the expected distribution under a strict coalescent process (Supplementary Fig. 7) and conflicting plastid bipartition frequencies at or near zero in the 10,000 simulated organellar gene trees (Supplementary Fig. 8). ILS alone is therefore insufficient to explain the observed cytoplasmic-nuclear incongruence recovered in these datasets and a scenario of historical gene flow must be invoked.

Hybridization is a widespread phenomenon within modern lineages of Fagaceae, especially between species within sections of *Quercus*[77], and plastome capture events are well documented between sympatric species[52,78-80]. Hybridization is also prevalent between closely related species across many genera within other fagalean families[81-83]. However, gene flow between modern genera is without precedent. When we applied molecular dating methods to the full plastome data, we found estimated divergence times for the deepest splits to generally fall within the rapid diversification phase for the HS clade based on nuclear data (Fig. 1; Supplementary Fig. 9). Without invoking non-sexual processes such as transmission between incompatible species through intimate physical contact, e.g., plant-plant parasitism and natural root grafts[84], ancient hybridization is the most likely source of

deep cytoplasmic-nuclear conflict in Fagaceae. Taken together, our results indicate this pattern of geographic division of reciprocally monophyletic plastome types is best explained as a vestige of widespread ancient hybridization among ancestral populations of the HS clade that became spatially isolated by paleogeographic barriers to gene flow dated minimally to the early Paleocene (Fig. 2; Supplementary Fig. 9).

We additionally found evidence for more recent plastome capture events resulting from hybridization within *Quercus* between the late Miocene to Pliocene (Fig. 2; Supplementary Fig. 9). As expected, the pattern of discordance between plastome and nuclear genomes uncovers multiple instances where species from phylogenetically distinct clades, but overlapping geographic ranges, share plastome types, for example between species of sections: *Ilex* and *Cerris*, *Virentes* and *Quercus*, *Protobalanus* and *Quercus*, and *Ponticae* and *Quercus* (Fig. 2). While inferring ancient hybridization events using cytoplasmic-nuclear gene tree conflict provides some evidence of reticulate evolutionary history, satisfactorily confirming and characterizing ancient gene flow requires a detailed investigation into the nuclear genome.

Ancient gene flow and adaptive introgression detected in the nuclear genome

Extensive investigation using a *D*-statistic (ABBA-BABA) test detected significant gene flow on 236 (0.911%) of 25882 trios extracted from the species tree ($P < 0.01$ after Bonferroni correction) (Supplementary Table 4). Not surprisingly, most cases of gene flow appeared to be recent in origin and between closely related species from within genera or sections of *Quercus* (Fig. 3A). Ancient gene flow, however, was detected between Eurasian white oaks (section *Quercus*) and *Q. pontica* (section *Ponticae*) and between North American white oaks (section *Quercus*) and the ancestor of section *Virentes* (Fig. 3), consistent with the results of gene tree analyses from the two genomes (Fig. 2). Network analyses using SNaQ confirmed historical gene flow between *Q. pontica* and Eurasian white oaks inferred in the current study (Supplementary Fig. 10) and previous studies[47].

We also assessed the distribution of alternative topologies within our 2124 nuclear gene dataset and found introgressed signals to be widely scattered across the genome (Supplementary Fig. 11). This is expected given that long-term recombination tends to fragment introgressed stretches of DNA following initial hybridization events [85]. However, positive selection has been shown to maintain long

introgressed haplotypes in populations of humans and maize[86,87], with the length of introgressed fragments increasing with stronger selection[88]. Our investigation of putatively ancient hybridization events between sections of *Quercus* yielded haplotypes that were significantly longer than expected under neutrality. Identity-by-descent (IBD) analyses based on whole genome SNP data clearly detected a large number of shared haplotype blocks (see Methods) for three lineage-pairs, i.e., *Q. pontica* vs. European and Asian white oaks, North American white oaks vs. section *Virentes*, and North American white oaks vs. *Q. sadleriana* (Fig. 4B-D). However, we did not find IBD blocks of similarly long lengths between other *Quercus* sections in which we documented plastome capture events (Supplementary Table 5).

Within the long sets of shared IBD regions, the *D*-statistic test revealed gene flow between oak sections. In addition, the recombination rate in the same IBD regions was not different from genomic background (*W* ranges from 82978 to 321198, $P = 0.66 – 0.73$, Mann-Whitney *U* test; Supplementary Table 6), and the length of the IBDs was not associated with recombination rate (Spearman's $\rho = -0.19 – 0.14$, $P = 0.10 – 0.79$; Supplementary Fig. 12). Therefore, these haplotypes shared between *Quercus* sections are most likely due to historical inter-sectional hybridization instead of the maintenance of ancestral polymorphisms in regions with reduced recombination rates.

To test this prediction, we calculated the probability of maintaining selectively neutral haplotypes of a given length in both oak sections after introgression using methods developed to study introgression in humans[86] and using generation times and mutation and recombination rates derived for oak species[89-91]. We determined that 166 IBD blocks (11724 -113757 bp) were significantly longer than expected if the introgressed fragments were selectively neutral ($P < 0.05$; Fig. 4d; see details in Methods), suggesting that the IBDs identified here provide convincing evidence of adaptive introgression. Multiple GO categories with important metabolic processes and molecular functions (e.g., terpene metabolic processes, sesquiterpenoid metabolic processes) were overrepresented for genes located in these IBD regions (Supplementary Table 7), further suggesting a diverse set of genes and functional categories may have contributed to adaptive introgression of oak species. Adaptive introgression between closely related species has recently been documented in *Quercus*[92]. Our study posits that introgressed elements between divergent oak sections could be preserved for millions of years by natural selection.

306 With the exception of the few cases involving sections *Quercus*, *Ponticae* and
307 *Virentes*, we found no corroborating evidence of hybridization within the nuclear
308 genome of the remaining lineages exhibiting cytoplasmic-nuclear gene tree conflict.
309 The occurrence of plastome capture events in the absence of detectable nuclear
310 introgression is not unexpected, and could be due to the early phases of hybrid zone
311 dynamics[93,94]. For example, extensive backcrossing with one parental species after
312 initial hybridization could sweep out signals of reticulation events in the nuclear
313 genome and recombination over long evolutionary time could have degraded signals
314 of ancient hybridization[95,96]. In oaks, backcrossing is preferentially unidirectional[97,98]
315 and linkage disequilibrium typically declines to background quickly (within 1kb)[99,100],
316 blurring the signals of past introgression in the nuclear genome. Alternatively, as
317 mentioned above, plastomes can be captured through non-hybridizing means such as
318 intimate physical contact, e.g. plant-plant parasitism and natural grafts[84], which would
319 leave no signal in the nuclear genome.
320
321 ***Genomic footprints of a changing temperate forest***
322 We show that the story of the evolutionary diversification of modern Fagaceae can be
323 told through the lens of two unlinked genomes, each contributing unique inferences to
324 disentangle a complex combination of divergent and reticulate historical events that
325 unfolded through the Cenozoic. Further, historical migration events in temperate
326 lineages were revealed by discovery of three exceptions to the NW-OW plastome
327 pattern (Fig. 2). Chestnuts (*Castanea*) currently distributed across the Holarctic arose
328 in the OW and moved to the NW, while the Eurasian oak sublineages of sections
329 *Quercus* and *Ponticae* are NW in origin, consistent with RAD-seq analysis[50]. These
330 bidirectional land bridge crossings with unambiguous origins document the timing of
331 limited, but key dispersal events leading to the spread of modern Fagaceae forests
332 across the northern hemisphere[101]. While an untold number of extinctions will escape
333 this level of scrutiny, the reciprocal migrations of oak and chestnut species during the
334 Oligocene provide evidence for the origins of ecologically significant components of
335 northern hemisphere forests. The cascading ecological implications of biotic
336 exchanges of this magnitude at the organismal level await future study.
337 Hybridization, common throughout Fagales, may be adaptive at various stages
338 of diversification depending on patterns of persistent interfertility and range overlap
339 among lineages[102]. For Fagaceae, an early stage of widespread hybridization among

ancestral elements of the HS clade is suggested by an enduring paleophylogeographic signal in the plastomes of modern lineages. Soon after, a rapid burst of cladogenesis at the base of the HS clade, tracked by the nuclear genome, generated the extant lineages as resolved here. As reproductive isolation evolved across most of these lineages, divergent evolution generated sets of exclusive plastome haplotypes within the broader phylogeographic pattern observed here, except for the instances where interfertile oak lineages experienced secondary contact during the Miocene. Within several clades of *Quercus*, cytoplasmic-nuclear gene tree incongruities support previous studies indicating an expanded role of hybridization in flowering plant evolution[103-106].

We suggest that oaks and their chestnut relatives have been hybridizing for millions of years. In Fagaceae, this is facilitated by small and evolutionarily stable genomes, high levels of synteny, and a consistent chromosome number across taxa [107-112]. In addition to conserved genomes and maintenance of some level of interfertility, these lineages share other life-history traits with diverse and often tropical tree genera that suggest the syngameon is functionally adaptive. Fagaceae species in particular share evolutionary and ecological characteristics that may promote adaptive introgression including generalized pollination systems, high levels of fecundity, and widespread sympatry[41,108,110].

Consequently, we document three main geographic areas of historical introgression between oak sections as evidenced by plastome capture: western North America, southeastern North America, and Eurasia. These areas are known to be centers of phylogenetic diversity for the genus[39] with extensive zones of sympatry and evidence for convergent evolution of form in response to climate[10,46]. In two of these areas, specifically Eurasia where the ranges of sections *Ponticae* and *Quercus* once overlapped and the American southeast where sections *Virentes* and *Quercus* are still known to hybridize, we present evidence from the nuclear genome that ancient hybridization has left a signature of adaptive evolution. While more detailed study is necessary to fully appreciate the impact these introgressed alleles may have had on the modern oak landscape in these regions, ancient hybridization between the relictual *Q. pontica* sublineage of sect. *Ponticae* and the widespread Eurasian sublineage of sect. *Quercus* appears to have contributed to an increased diversification rate in section *Quercus* during the Miocene (Fig 1; see also Hipp *et al.*[39]). This uptick in speciation and the ecological opportunity available to the white oaks, a relative

newcomer to the Old World, marks the rise and spread of a dominant deciduous lineage bearing an introgressed nuclear genome into the forested ecosystems across Eurasia.

Methods

Taxon sampling, DNA extraction and whole genome sequencing

We constructed a comprehensive Fagaceae dataset consisting of 122 individuals from 91 species representing all eight currently recognized genera[54,113]. Complete taxon sampling was achieved for three small genera: *Chrysolepis* (2 species), *Notholithocarpus* (2 species) and *Trigonobalanus* (3 species). For the remaining genera, representative samples for all major lineages were included: *Fagus* (2), *Castanea* (5), *Castanopsis* (12) and *Lithocarpus* (10). For the well-studied genus *Quercus*, extensive sampling (54 species) was conducted to represent all eight recognized sections[39,44]. Several species were represented by multiple accessions collected from different natural populations or cultivated plants. *Betula pendula* was selected as an outgroup due to its close relationship to Fagaceae and the availability of an assembled genome [114]. Accession information is provided in Supplementary Table 1.

Total genomic DNA was extracted from silica-dried leaf tissue using BioTeke Genomic DNA Extraction Kit (Beijing, China). High quality DNA was used to constructed paired-end sequencing libraries with an insert size of 500-600 bp according to the Illumina library preparation protocol. Sequencing was carried out on the NovaSeq platform at Novogene (Beijing, China) to a coverage of 25-40× for all samples. Short reads (150 bp pair-end) have been deposited to Genbank under accession numbers XX-XX.

Orthologous gene identification and nuclear alignment matrix assembly

To obtain orthologous genes (OGs) for phylogenetic analysis, we performed a series of critical search and filtering processes. There are four high quality genome assemblies (chromosome-level) available in Fagaceae: *Fagus sylvatica*[115], *Castanea mollissima*[116], *Quercus robur*[91] and *Quercus lobata*[117]. These four assemblies together with the outgroup species *B. pendula*[114] were used to identify putative OGs in OrthoFinder v2.3.12[118] with an E-value of $1E^{-5}$. Orthologous groups containing only one sequence from each examined species were retained to minimize paralogs in

subsequent phylogenetic analyses. Single copy genes (SCGs) identified by OrthoFinder may still have duplicates, either as pseudogenes or un-annotated functional genes in the genome. To identify and remove additional multiple-copy genes, we blasted coding sequences (CDS) of SCGs against each of the five genomes using BLAST+[119]. We filtered alignments using the following thresholds: E-value < $1E^{-5}$, alignment length ≥ 80% of the query sequence, and identity ≥ 80%. We kept CDS with only one hit in each of the five species. The retained CDS regions identified as belonging to a single gene were concatenated for subsequent phylogenetic analyses.

To generate the nuclear DNA sequences, we sequenced whole genomes of 117 individuals (Supplementary Table 1) to a coverage of 25-40× using the Illumina NovaSeq platform and called genotypes in SCGs regions. We trimmed and filtered raw reads using trimmomatic v0.39[120], mapped high quality reads to a reference genome using BWA[121], and called genotypes via HaplotypeCaller in GATK v4.2[122]. A simulation study found that the inclusion of nonpolymorphic positions in the alignment and mapping short reads to multiple references could improve the accuracy of phylogenetic inference[123]. Thus, we called all available sites (both variants and invariants). To reduce the effects of reference bias, we used three reference genomes for mapping and SNP calling in related species. The genome of *F. sylvatica* was used as reference for genus *Fagus*, the *Q. robur* genome was used for genus *Quercus*, and the *Castanea mollissima* genome was used for the remaining six genera. We only considered sites with mapping quality ≥ 30 and base quality ≥ 30, and further filtered variants using the following criteria: (1) homozygous genotypes with depth < 4 or heterozygous genotypes with depth < 20 were assigned as missing; (2) sites with mean depth < 5 or > 100 across all individuals were discarded; (3) sites with proportion of heterozygous genotypes > 50% were excluded.

To obtain an aligned matrix of SCGs, we generated a 6-way whole genome alignment based on the four reference genomes and two additional assemblies (*Q. lobata* and *Q. suber*) following a lastZ/Multiz pipeline[124,125]. We used *Q. robur* as a reference genome for genome alignment, and merged genotypes from mapping to different references or extracted from different assemblies together according to their relative positions on the *Q. robur* genome. The data matrix was then filtered by excluding sites containing ≥ 10% missing data, and SCGs with length < 200 bp. Alignments with divergent paralogous genes usually show elevated levels of

polymorphism, thus we further excluded SCGs with polymorphism in the top 95th percentile (cutoff = 43.8%). Every OG was presented in all sampled individuals with no missing data. Our final dataset included 2124 SCGs with a total length of 1,689,974 bp for data analyses (Supplementary Table 2; Dryad Data Archive).

Evaluating the impacts of reference genomes on the accuracy of SNP calling and phylogenetic reconstruction

We applied both empirical and simulation analyses to assess the impacts of the reference genome on the accuracy of SNP calling and phylogenetic reconstruction. The assembly of *Castanea mollissima* was used as the reference genome for SNP calling in *Castanea* and the five genera (*Chrysolepis*, *Castanopsis, Lithocarpus, Notholithocarpus* and *Trigonobalanus*) without available genome assemblies. To test whether reference bias was introduced by using a divergent reference genome, we re-called SNPs for these five genera by using *Q. robur* as reference, and compared genotypes called from *Q. robur* with those from *C. mollissima*. Despite the slightly higher rate of missing data (9.29 – 9.72%) observed with using *Q. robur* as reference genome compared to *C. mollissima* (3.96 – 4.27%), 95.62 - 95.84% genotypes were identical between these two datasets (Supplementary Table 8). Identical tree topologies also were generated based on the two datasets when using the same phylogenetic method (data not shown), suggesting weak reference bias in our data.

To further monitor the accuracy of genotyping in the query dataset with different divergence levels from the reference genome, we generated mutated sequences (henceforth referred to as "mutated-sequence") by randomly adding 0.25% − 20% mutations to the longest chromosome of *Q. robur* (chromosome 2, henceforth referred to as "reference-sequence"). Next, we simulated 150bp pair-end reads from each mutated-sequence with 30× coverage (close to our sequencing depth 25-40×). Simulated reads were mapped to the reference-sequence, and SNPs were called and filtered using the same protocol as described above. For each simulated dataset, we compared genotype calls to the mutated-sequence from which the datasets were generated. To mimic the real data, SNPs called from the repetitive regions was excluded from data analyses. The true positive (TP) rate was defined as TP/(TP+FP), where TP is position identical to mutated-sequence, and FP (false positives) are called genotypes different from mutated-sequence. The missing rate (MR) was defined as MISS/SIZE, where MISS is non-genotyped sites and SIZE is total sites (~51.2 Mb) in

476 the reference-sequence after excluding masked repetitive regions. High TP rate (>
477 97.7%) and low MR (< 1.5%) were found in datasets with divergence levels from
478 reference-sequence no more than 10% (Supplementary Fig. 12). By extracting
479 sequences of the 2124 SCGs from the 6-way whole genome alignment, nucleotide
480 divergence was estimated as 7.46% - 7.69% between most divergent genera (i.e.
481 *Fagus* vs. *Quercus* and *Castanea*) (Supplementary Table 9), genotypes called from
482 SCGs by using a divergent reference (e.g. using *C. mollissima* for other genera)
483 would not result in strong reference bias.
484
485 ***Plastome assembly and alignment***
486 We assembled 117 plastomes during the course of this study and obtained five
487 additional plastomes from Genbank (Supplementary Table 1). Raw reads from whole
488 genome sequencing were used for *de novo* assembly of plastomes in NOVOPlasty
489 v4.2[126]. A ribulose-bisphosphate carboxylase (*rbc*L) gene sequence from *Quercus*
490 *rubra* was used as the seed sequence for assembly. Assembled plastomes were
491 annotated using the program PGA[127]. The boundaries of inverted repeats and coding
492 regions of each annotated gene were determined in Geneious 7.1.4[128] by using the *Q.*
493 *rubra* plastome as a reference. Coding regions of 76 protein-coding genes present in
494 all species were extracted from the assemblies (Supplementary Table 10), aligned
495 using MAFFT 7.221[129], and manually adjusted using Bioedit v.7.2
496 (https://bioedit.software.informer.com). Based on plant plastid genetic code, the
497 codon alignment was translated into amino acid sequences. A preliminary
498 phylogenetic analysis found two *Q. ilex* samples were placed as a sister group to all
499 other Fagaceae species except the genera *Fagus* and *Trigonobalanus*. This is likely an
500 artificial of clustering, because previous analyses with extensive sampling (26
501 individuals) spanning the geographic distribution of *Q. ilex* placed this species within
502 a clade formed by Eurasian oaks and genera *Castanea* and *Castanopsis* based on
503 plastid data[53]. Therefore, we excluded these two *Q. ilex* samples from subsequent
504 plastome analyses. Removing these two samples did not change the topology among
505 other species (data not shown). The plastome alignment is 65,814 bp in length, of
506 which 11,058 characters were polymorphic. A list of the 76 genes is presented in
507 Supplementary Table 10, the alignment of these genes can be found in Dryad-
508 archived data, and the plastome assemblies are deposited in Genbank under accession
509 numbers XX-XX.

Phylogenetic analyses

Phylogenetic analyses were conducted using Maximum Likelihood (ML) and Bayesian approaches for concatenated nuclear and plastome data. Partitioned ML analysis was performed using RAxML v8.2.12[123]. The best-scoring ML tree was found from 1000 ML trees, and topological robustness was evaluated by using 1000 non-parametric bootstrap replicates. Bayesian analysis was conducted in MrBayes v3.2.6[130]. Markov chain Monte Carlo (MCMC) runs were performed for 10 million generations, and trees were sampled every 100 generations. The first 2,5000 (25%) trees were discarded as burn-in to ensure that the chains were stationary. The remaining trees were used to generate a strict consensus tree and to calculate posterior probabilities for each node.

PartitionFinder2[131] was used to determine the optimal partitioning strategy and evolutionary model of each partition under the Akaike Information Criterion (AIC)[132]. For nuclear DNA data, partitioning by gene yielded 35 partitions in the best scheme. For plastome DNA data, full partitioning scheme by both locus and codon position (each of the three codon positions in each gene as one partition) was examined, and the best scheme contained 24 partitions. For plastome amino acid data, each gene was considered as one partition, resulting in 12 partitions in the optimal scheme. In ML analyses, the GTRGAMMA model was used for all DNA sequence partitions, and the evolutionary models chosen by PartitonFinder2 were used for amino acid partitions. For Bayesian analyses, the evolutionary model identified by PartitionFinder2 was used for each DNA and amino acid partition. The models, partitions, and alignments used for phylogenetic analysis can be found in Dryad Data Archive.

Two coalescent-based approaches were used to infer a species tree for Fagaceae. First, we applied a summary statistic method using ASTRAL-III v5.7.3[133]. Gene trees were estimated from single-gene alignments using RAxML with GTRGAMMA model and 1000 fast bootstrap replicates. Individual gene trees (best trees) and bootstrap replicates were used to estimate a species tree in ASTRAL-III with 1000 coalescent bootstrap replicates. Following Zhang *et al.*[133], branches with low support were removed to improve the accuracy of tree inference. We tested different thresholds by collapsing branches with support less than 10%, 20%, 30%, 40% and 50%, and obtained near-identical tree topologies (data not shown). The tree generated by ASTRAL-III with 50% threshold is presented.

SVDquartets[134], a method based on site pattern frequencies and algebraic statistics implemented in PAUP v4.0a152[135] was additionally used to estimate a species tree. This method was originally designed for SNP data, but also performed well on large multiple-locus datasets[134]. The concatenated nuclear data matrix was used as input for SVDquartets. All possible quartets were evaluated, and clade support was assessed using 500 bootstrap replicates.

Divergence time and diversification rate estimation

Divergence time estimation was conducted for both plastome and nuclear datasets using MCMCTree v4.9j in the PAML package[136]. MCMCTree estimates divergence times using an approximate likelihood method, and is computationally efficient with large genomic data[137]. The MCMC chains were first run for 3 million generations as burn-in, and then were sampled every 400 generations until a total of 25,000 samples were collected (10 million generations). Tracer and LogCombiner were used to confirm the convergence across each run and ensure the ESS of all parameters were greater than 200. For each of the plastome and nuclear datasets, three independent runs with different seeds were compared for convergence, and similar results were generated.

For nuclear DNA data, we divided the 2124 nuclear genes into three partitions according to substitution rates estimated by Baseml (package in PAML) with a strict molecular clock and then applied an uncorrelated rate model (clock = 2 in MCMCTree) to infer divergence times. We used priors of G (1, 6.1677) for the overall substitution rates (rgene_gamma), G (2, 5, 1) for the rate-drift parameter (sigma2_gamma). Because concatenated and coalescent analyses revealed different relationships among genera *Quercus*+*Notholithocarpus*, *Lithocarpus* and *Chrysolepis*, we constrained each alternative topology and constructed the ML reference tree for dating. For plastome data, we treated all 76 cp genes as one partition, and estimated divergence times by using the ML tree as reference under an uncorrelated rate model. We set priors of rgene_gamma and sigma2_gamma parameters as G (1, 41.667) and G (2, 5, 1), respectively.

Based on results of Xiang *et al.*[64], the root age of Fagaceae was constrained to 95.5 - 101.2 MYA for both plastome and nuclear data. For nuclear data, we further added six additional widely accepted fossil calibrations (Supplementary Table 3). For the plastome analysis, only two calibrations could be used due to non-monophyletic

lineages in the plastome tree (Supplementary Table 3). For species with multiple samples, we chose one individual for dating the nuclear DNA tree, while retaining all individuals for dating the plastome tree because many species were not monomorphic for their plastome.

To estimate the diversification rate of Fagaceae, we applied Bayesian Analysis of Macroevolutionary Mixture (BAMM)[138]. The time tree estimated by MCMCtree was used as an input tree. To account for incomplete taxon sampling, we calculated sampling fraction of each genus and each section of genus *Quercus* based on the number of species recorded in previous reports[39,113], and then added un-sampled taxa to a random position in each corresponding lineage (Supplementary Table 11). The BAMM analyses were run for 10 million generations, saving every 1000 generations. The first 30% samples were discarded as burn-in, and the remaining samples were summarized and plotted using BAMMtools[138].

Topological concordance analyses

To evaluate the conflicts between nuclear gene trees and the species tree, we first calculated the internode certainty all (ICA) to quantify the degree of conflict on each node between a target tree and gene trees[139]. ICA values close to 1 indicate strong concordance for the bipartition defined by a given internode, while ICA values close to 0 indicate strong conflict. Negative ICA values indicate that the defined bipartition conflict with other high frequent bipartitions. The ICA values were estimated in RAxML and the species tree found by ASTRAL-III was used as the target tree. We further summarized the number of conflicting and concordant bipartitions with PHYPARTS[140], using the species tree estimated by ASTRAL-III and the individual gene trees.

Evaluation of substitutional saturation and codon-usage bias within the chloroplast dataset

To investigate whether base substitution saturation biased the accuracy of phylogenetic inference in plastome phylogenetic analyses, we estimated the amount of substitution saturation using methods detailed in Xia *et al.*[141]. This involved employing critical index of substitution saturation (ISSc) that defines a threshold for significant saturation in the data. From the data of 76 chloroplast genes, we assessed the level of substitution saturation for codon12 and codon3 using the program

DAMBE7[142], and found that there was sufficient phylogenetic information at all codon positions (Supplementary Table 12).

To investigate how synonymous codon usage varies among Fagaceae species, and whether synonymous codon biases have resulted in artificial and random phylogenetic inference, we measured Relative Synonymous Codon Usage (RSCU) values using GCUA[143]. RSCU is defined as the ratio of the observed codon appearance to the number expected given that all synonymous codons appear with uniform frequency. We found similar level of GC content and variation in codon bias across Fagaceae species (Supplementary Fig. 13). These results suggested that Fagaceae plastid genomes are highly conserved, and the plastid-based analyses would be not biased due to substitution saturation or compositional heterogeneity among species.

Coalescent simulation

To test whether incomplete lineage sorting (ILS) alone could explain the incongruence between plastome tree and nuclear species tree, we followed Folk *et al*.[144] to simulate 10,000 plastome trees under the coalescent model using DENDROPY v.4.1.0[145]. The ASTRAL-III tree was used as a guide tree for the simulation. To simulate plastome trees, branch lengths were scaled by a factor of four to account for the haploidy and maternal inheritance of the plastome. Clade frequencies of simulated trees were summarized using PHYPARTS[140]. In the scenario of ILS alone, the topology from our empirical plastome tree should be present in simulated trees with high frequency; if gene flow is present, the topology recovered in our empirical tree should be absent or at very low frequency in the simulated trees. Following previous studies[96,146], we also counted the number of extra lineages in observed and simulated trees using the function deep-coal_count in Phylonet v2.4[147]. In the case that gene flow is present, more extra lineages are expected in the observed trees relative to simulated trees.

Gene flow analyses

To detect potential gene flow between species, we performed ABBA-BABA statistic tests in Dsuite[148]. These analyses take advantage of a four taxon statement ((H1, H2)H3)H4). With H4 as the outgroup, H1 and H2 are treated as a pair of sister species and H3 is tested as the species with potential gene flow with H1 or H2. The number

of sites with allele patterns of ABBA and BABA are tallied. The *D* statistic is derived from calculating *D* = (nABBA – nBABA)/(nABBA + nBABA), where nABBA and nBABA are the total number of sites with patterns of ABBA and BABA, respectively[149,150]. A negative *D* value indicates gene flow between H1 and H3, a positive *D* value indicates gene flow between H2 and H3, and *D* = 0 indicates no gene flow[149,150]. Because ABBA-BABA test assumes a sister relationship between H1 and H2, we restricted our analyses by sampling H1 and H2 from same genera, or same sections within genus *Quercus*. In addition, because H1 and H2 are sister species, the sites with the pattern of BBAA are expected to be larger than ABBA and BABA patterns. We further filtered trios that violated this assumption, and applied ABBA-BABA test to 25882 trios extracted from the species tree. To account for multiple testing, we corrected *P*-values with Benjamini-Hochberg FDR[151]. For a pair of species involved in multiple tested trios (for example, while H2 and H3 are fixed, there may be different H1 taxa available, thus different *D* values for H2 and H3 may be generated), the estimated *D* value with lowest FDR was retained. An individual of *Trigonobalanus doichangensis* was used as an outgroup for all tests. To test how outgroup choice influenced the analysis, we also used an individual of *Notholithocapus densiflorus* in tests within *Quercus* and obtained results similar to those using *T. doichangensis* (data not shown).

To further explore the reticulate evolutionary histories within Fagaceae, we inferred species networks using SNaQ[152] implemented in the package PhyloNetworks[153]. SNaQ is a pseudolikelihood method, which estimates a phylogenetic network while accounting for both ILS and gene flow[152]. We reduced the dataset to a computationally tractable size[154], and generated four sub-datasets each with 15-17 taxa sampled. For each sub-dataset, we sampled species showing inconsistent placement between nuclear and plastome trees. The first one focused on relationships within subgenus *Quercus* and a sample of 16 species (Supplementary Fig. 10). The second one focused on the relationship within subgenus *Cerris* and a sample of 15 species (Supplementary Fig. 10). The third one focused on the relationships among genera *Castanea*, *Castanopsis*, *Lithocarpus*, and *Quercus*, and the forth one other focused on the relationship among genera *Chrysolepis*, *Notholithocarpus*, and *Quercus* (Supplementary Fig. 10). One individual gene trees generated by RAxML were used as input, and nested analyses were performed allowing for zero (h = 0) to four (h = 4) hybridization events. Each nested analysis

679 was optimized by 10 independent runs, and the best fitting model was selected based
680 on the log pseudolikelihood score.
681 To investigate the genomic pattern of introgressed loci, we quantified the
682 distribution of phylogenetic signal for conflicting topologies across nuclear gene
683 trees, and then mapped loci supporting alternative partitions to the *Q. robur* genome[91].
684 Following Shen *et al.*[155], we calculated site-wise log-likelihood scores for the primary
685 and alternative topologies in our concatenated matrix using the "-f G" command in
686 RAxML. After that, the difference in site-wise log-likelihood scores (ΔSLS) between
687 topologies were summed across sites in each gene, generating gene-wise log-
688 likelihood scores (ΔGLS). For each node of interest, the primary topology was
689 defined as the species tree recovered by ASTRAL-III, and the alternative topologies
690 were ML trees constrained to recover the most common conflicting bipartitions.
691
692 ***Identity-by-descent (IBD) analyses***
693 We performed IBD analyses based on genome-wide SNP data in the genus *Quercus*.
694 By using a same SNP calling and filtering procedure described above (see section
695 "*Orthologous gene identification and nuclear alignment matrix assembly*"). Raw
696 reads of *Quercus* species were trimmed using Trimmomatic v0.38[120], aligned to *Q.*
697 *robur* reference genome assembly[91] using BWA[121], and called genotypes using
698 GATK v4.1[156]. We applied a strict filtering process to remove low quality SNPs. We
699 removed all sites located in repetitive regions of the *Q. robur* reference genome[91], and
700 discarded all indels and multiallelic SNPs. We further set genotypes supported by less
701 than four reads as missing data, and deleted SNPs with mean depth <5 or >100, or
702 genotyped in less than half of individuals, or proportion of called heterozygous
703 genotypes >50%. Finally, we obtained 34,250,467 high quality SNPs for IBD
704 analyses.
705 We used Beagle v4.1[157] to phase and impute the SNP data, and uncover shared
706 IBD blocks between species. The following parameters were used for IBD analyses in
707 Beagle: window = 100,000; overlap = 10,000; ibdtrim = 100; ibdlod = 5. To compare
708 the recombination rate between IBD blocks and genomic background, we used a
709 genetic map of *Q. robur* developed by Plomion *et al.*[91]. We smoothed the
710 recombination rate across the genome to 200 kb, and then mapped IBD blocks to the
711 genetic map. For each IBD block, we obtained the recombination rate on middle

712 points of the block, and then used this value as the recombination rate for the whole
713 IBD block.
714 To test whether the IBDs shared between sections are under selection, we
715 calculated the probability of a selectively neutral haplotype with a given length shared
716 by two sections after introgression. If the IBD blocks were significant longer than the
717 neutral haplotype, they were most likely maintained by selection after introgression.
718 Following Huerta-Sanchez *et al.*[86], the probability for each shared IBD block was
719 estimated as: 1- GammaCDF (*L*, shape = 2, rate = lambda), where the GAmmaCDF is
720 the Gamma distribution function and arguments are given in parentheses. The rate
721 parameter lambda was estimated as: lambda = r * (T/G), where r is recombination
722 rate, T is the time of gene flow occurred, and G is the generation time. To calculate
723 the time of gene flow introduced shared IBDs between oak sections, we calculated the
724 genetic divergence (d_{XY}) between sections (i.e. *Q. pontica vs.* European and Asian
725 white oaks, North American white oaks *vs.* section *Virentes* and *Q. sadleriana*) on
726 shared IBD blocks. The estimated mean values of d_{XY} was 0.011 – 0.018, which was
727 transformed to 2.8 – 4.5 millions years based on a mutation rate of 2×10^{-9} per site per
728 year[10]. Thus, we roughly used 3 million years for the time of introgression. Using a
729 recombination rate of 1×10^{-8} estimated from *Q. robur* genetic map (total length of
730 genetic map is 740 cM, and the genome size is 804 Mb)[91], and assuming a generation
731 time of 50 years, we get lambda = $1 \times 10^{-8} \times (3 \times 10^{6}/50) = 6 \times 10^{-4}$. We calculated
732 the probability for each IBD blocks and corrected multiple testing using Benjamini–
733 Hochberg FDR[151].
734 To examine whether functional classes of genes were overrepresented in IBD
735 blocks under selection, we performed GO analyses using the R package topGO 2.43.0
736 (http://www.bioconductor.org/). We applied Fisher's exact test to estimate the
737 statistical significance of enrichment, and corrected multiple testing by Benjamini–
738 Hochberg FDR[151]. A cutoff of FDR < 0.01 was used to determine the significance of
739 GO enrichment.
740
741 **References**
742 1 Sun, J. *et al.* Synchronous turnover of flora, fauna, and climate at the Eocene-
743 Oligocene Boundary in Asia. *Scientific Reports* **4**, 7463 (2014).

2 Tiffney, B. H. Perspectives on the origin of the floristic similarity be tween eastern Asia and eastern North America. *Journal of the Arnold Arboretum* **66**, 73-94 (1985).

3 Tiffney, B. H. The Eocene North Atlantic land bridge: its importance in Tertiary and modern phytogeography of the Northern Hemisphere. *Journal of the Arnold Arboretum* **66**, 243–273 (1985).

4 Donoghue, M. J. A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 11549-11555 (2008).

5 Edwards, E. J. *et al.* Convergence, consilience, and the evolution of temperate deciduous forests. *American Naturalist* **190**, S87-S104 (2017).

6 Segovia, R. A. *et al.* Freezing and water availability structure the evolutionary diversity of trees across the Americas. *Science Advances* **6**, eaaz5373 (2020).

7 Tiffney, B. H. & Manchester, S. R. The use of geological and paleontological evidence in evaluating plant phylogeographichypotheses in the Northern Hemisphere Tertiary. *International Journal of Plant Sciences* **162**, S3-S17 (2001).

8 Axelrod, D. I. Biogeography of oaks in the Arcto-Tertiary province. *Annals of the Missouri Botanical Garden* **70**, 629-657 (1983).

9 Axelrod, D. I., Ai-Shehbaz, I. & Raven, P. H. *History of the modern flora of China*. (Springer, 1996).

10 Cavender-Bares, J. Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. *New Phytologist* **221**, 669-692 (2019).

11 Delcourt, H. R. & A., D. P. *North American terrestrial vegetation*. (Cambridge University Press, 2000).

12 Olson, J. S., Watts, J. A. & Allison, L. J. Carbon in live vegetation of major world ecosystems. (1983).

13 Soepadmo, E. *Flora Malesiana Series I.* Vol. 7 (Noordhoff International Publishing, 1972).

14 Vogt, K. A. *et al.* Review of root dynamics in forest ecosystems grouped by climate, climatic forest type and species. *Plant and Soil* **187**, 159-219 (1996).

15 Whitmore, T. C. *Tropical rain forest of the Far East*. (Oxford University Press, 1984).

16 Zhu, H. Ecological and biogeographical studies on the tropical rain forest of south Yunnan, SW China with a special reference to its relation with rain forests of tropical Asia. *Journal of Biogeography* **24**, 647-662 (1997).

17 Averill, C., Bhatnagar, J. M., Dietze, M. C., Pearse, W. D. & Kivlin, S. N. Global imprint of mycorrhizal fungi on whole-plant nutrient economics. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 23163-23168 (2019).

18 Martin, F., Kohler, A., Murat, C., Veneault-Fourrey, C. & Hibbett, D. S. Unearthing the roots of ectomycorrhizal symbioses. *Nature Reviews Microbiology* **14**, 760-773 (2016).

19 Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis, 2nd edn*. (Academic Press, 1997).

20 Abrahamson, W. G. & Melika, G. Gall-inducing insects (Cynipinae) provide insights into plant systematic relationships. *American Journal of Botany* **85**, 111-111 (1998).

21 Raman, A. Nutritional diversity in gall-inducing insects and their evolutionary relationships with flowering plants. *International Journal of Ecology and Environmental Sciences* **22**, 133-143 (1996).

22 Stone, G. N. *et al.* Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution* **63**, 854-869 (2009).

23 Johnson, W. C. & Webb, T. The role of bluejays (*Cyanocitta cristata* L.) in the postglacial dispersal of fagaceous trees in eastern North America. *Journal of Biogeography* **16**, 561-571 (1989).

24 Koenig, W. D. & Haydock, J. Oaks, acorns, and the geographical ecology of acorn woodpeckers. *Journal of Biogeography* **26**, 159-165 (1999).

25 Payne, J. & Francis, C. M. *A Field Guide to the Mammals of Borneo*. (Sabah Society with World Wildlife Fund Malaysia, 1985).

26 Steele, M. A. *Oak Seed Dispersal*. (The Johns Hopkins University Press, 2021).

27 Vander Wall, S. B. The evolutionary ecology of nut dispersal. *Botanical Review* **67**, 74-117 (2001).

28 Vander Wall, S. B. How plants manipulate the scatter-hoarding behaviour of seed-dispersing animals. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 989-997 (2010).

29 Barrón, E. *et al.* in *Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L.* (eds Eustaquio Gil-Pelegrín, José Javier Peguero-Pina, & Domingo Sancho-Knapik) 39-105 (Springer International Publishing, 2017).

30 Crepet, W. L. & Nixon, K. C. Earliest megafossil evidence of Fagaceae: phylogenetic and biogeographic implications. *American Journal of Botany* **76**, 842-855 (1989).

31 Denk, T. & Grimm, G. W. Significance of pollen characteristics for infrageneric classification and phylogeny in *Quercus* (Fagaceae). *International Journal of Plant Sciences* **170**, 926-940 (2009).

32 Denk, T., Grímsson, F. & Zetter, R. Fagaceae from the early Oligocene of Central Europe: Persisting new world and emerging old world biogeographic links. *Review of Palaeobotany and Palynology* **169**, 7-20 (2012).

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

34 Jones, J. H. Evolution of the Fagaceae: the implications of foliar features. *Annals of the Missouri Botanical Garden* **73**, 228-275 (1986).

35 Manchester, S. R. Biogeographical relationships of North American Tertiary floras. *Annals of the Missouri Botanical Garden* **86**, 472-522 (1999).

36 Sadowski, E. M., Schmidt, A. R. & Denk, T. Staminate inflorescences with in situ pollen from Eocene Baltic amber reveal high diversity in Fagaceae (oak family). *Willdenowia* **50**, 405-517 (2020).

37 Bouchal, J., Zetter, R., Grimsson, F. & Denk, T. Evolutionary trends and ecological differentiation in early Cenozoic Fagaceae of western North America. *American Journal of Botany* **101**, 1332-1349 (2014).

38 Gandolfo, M. A., Nixon, K. C., Crepet, W. L. & Grimaldi, D. A. A late Cretaceous fagalean inflorescence preserved in amber from New Jersey. *American Journal of Botany* **105**, 1424-1435 (2018).

39 Hipp, A. L. *et al.* Genomic landscape of the global oak phylogeny. *New Phytologist* **226**, 1198-1212 (2020).

40 Wilf, P., Nixon, K. C., Gandolfo, M. A. & Cuneo, N. R. Eocene Fagaceae from Patagonia and Gondwanan legacy in Asian rainforests. *Science* **364**, eaaw5139 (2019).

41 Petit, R. J. & Hampe, A. Some evolutionary consequences of being a tree. *Annual Review of Ecology Evolution and Systematics* **37**, 187-214 (2006).

42 Smith, S. A. & Donoghue, M. J. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322**, 86-89 (2008).

43 Denk, T. & Grimm, G. W. The biogeographic history of beech trees. *Review of Palaeobotany and Palynology* **158**, 83-100 (2009).

44 Denk, T., Grimm, G. W., Manos, P. S., Deng, M. & Hipp, A. L. in *Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L.* Vol. 7 *Tree Physiology* (eds E. GilPelegrin, J. J. PegueroPina, & D. SanchoKnapik) 13-38 (Springer International Publishing, 2017).

45 Deng, M., Jiang, X. L., Hipp, A. L., Manos, P. S. & Hahn, M. 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. *Molecular Phylogenetics and Evolution* **119**, 170-181 (2018).

46 Hipp, A. L. *et al.* Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. *New Phytologist* **217**, 439-452 (2018).

47 Crowl, A. A. *et al.* Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). *New Phytologist* **226**, 1158-1170 (2020).

48 Hauser, D. A., Keuter, A., McVay, J. D., Hipp, A. L. & Manos, P. S. The evolution and diversification of the red oaks of the California Floristic Province (*Quercus* section *Lobatae*, series *Agrifoliae*). *American Journal of Botany* **104**, 1581-1595 (2017).

49 McVay, J. D., Hauser, D., Hipp, A. L. & Manos, P. S. Phylogenomics reveals a complex evolutionary history of lobed-leaf white oaks in western North America. *Genome* **60**, 733-742 (2017).

50 McVay, J. D., Hipp, A. L. & Manos, P. S. A genetic legacy of introgression confounds phylogeny and biogeography in oaks. *Proceedings of the Royal Society B* **284**, 20170300 (2017).

878 51 Manos, P. S., Doyle, J. J. & Nixon, K. C. Phylogeny, biogeography, and
879 processes of molecular differentiation in *Quercus* subgenus *Quercus*
880 (Fagaceae). *Molecular Phylogenetics and Evolution* **12**, 333-349 (1999).
881 52 Pham, K. K., Hipp, A. L., Manos, P. S. & Cronn, R. C. A time and a place for
882 everything: Phylogenetic history and geography as joint predictors of oak
883 plastome phylogeny. *Genome* **60**, 720-732 (2017).
884 53 Simeone, M. C. *et al.* Plastome data reveal multiple geographic origins of
885 *Quercus* Group *Ilex*. *PeerJ* **4**, e1897 (2016).
886 54 Oh, S. H. & Manos, P. S. Molecular phylogenetics and cupule evolution in
887 Fagaceae as inferred from nuclear *CRABS CLAW* sequences. *Taxon* **57**, 434-
888 451 (2008).
889 55 Renne, P. R. *et al.* Time scales of critical events around the Cretaceous-
890 Paleogene boundary. *Science* **339**, 684-687 (2013).
891 56 Koenen, E. J. M. *et al.* The origin of the Legumes is a complex paleopolyploid
892 phylogenomic tangle closely associated with the Cretaceous–Paleogene (K–
893 Pg) mass extinction event. *Systematic Biology* **70**, 508-526 (2021).
894 57 Wang, W. *et al.* Menispermaceae and the diversification of tropical rainforests
895 near the Cretaceous-Paleogene boundary. *New Phytologist* **195**, 470-478
896 (2012).
897 58 Suh, A., Smeds, L. & Ellegren, H. The dynamics of incomplete lineage sorting
898 across the ancient adaptive radiation of Neoavian birds. *PLoS Biology* **13**,
899 e1002224 (2015).
900 59 Feng, Y. J. *et al.* Phylogenomics reveals rapid, simultaneous diversification of
901 three major clades of Gondwanan frogs at the Cretaceous-Paleogene
902 boundary. *Proceedings of the National Academy of Sciences of the United*
903 *States of America* **114**, E5864-E5870 (2017).
904 60 Alfaro, M. E. *et al.* Explosive diversification of marine fishes at the
905 Cretaceous-Palaeogene boundary. *Nature Ecology & Evolution* **2**, 688-696
906 (2018).
907 61 Meredith, R. W. *et al.* Impacts of the Cretaceous terrestrial revolution and
908 KPg extinction on mammal diversification. *Science* **334**, 521-524 (2011).
909 62 Martinez, I. & Gonzalez-Taboada, F. Seed dispersal patterns in a temperate
910 forest during a mast event: performance of alternative dispersal kernels.
911 *Oecologia* **159**, 389-400 (2009).

912 63 Larson-Johnson, K. Phylogenetic investigation of the complex evolutionary
913 history of dispersal mode and diversification rates across living and fossil
914 Fagales. *New Phytologist* **209**, 418-435 (2016).
915 64 Xiang, X. G. *et al.* Large-scale phylogenetic analyses reveal fagalean
916 diversification promoted by the interplay of diaspores and environments in the
917 Paleogene. *Perspectives in Plant Ecology, Evolution and Systematics* **16**, 101-
918 110 (2014).
919 65 Casanovas-Vilar, I. *et al.* Oldest skeleton of a fossil flying squirrel casts new
920 light on the phylogeny of the group. *Elife* **7**, e39270 (2018).
921 66 Huchon, D. *et al.* Rodent phylogeny and a timescale for the evolution of
922 glires: Evidence from an extensive taxon sampling using three nuclear genes.
923 *Molecular Biology and Evolution* **19**, 1053-1065 (2002).
924 67 Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: Species-
925 level sets of phylogenies for questions in ecology, evolution, and conservation.
926 *PLoS Biology* **17**, e3000494 (2019).
927 68 Jonsson, K. A. *et al.* A supermatrix phylogeny of corvoid passerine birds
928 (Ayes: Corvides). *Molecular Phylogenetics and Evolution* **94**, 87-94 (2016).
929 69 Prum, R. O. *et al.* A comprehensive phylogeny of birds (Aves) using targeted
930 next-generation DNA sequencing. *Nature* **526**, 569-573 (2015).
931 70 Benz, B. W., Robbins, M. B. & Peterson, A. T. Evolutionary history of
932 woodpeckers and allies (Aves : Picidae): Placing key taxa on the phylogenetic
933 tree. *Molecular Phylogenetics and Evolution* **40**, 389-399 (2006).
934 71 Lutzoni, F. *et al.* Contemporaneous radiations of fungi and plants linked to
935 symbiosis. *Nature Communications* **9**, 5451 (2018).
936 72 Bonfante, P. & Genre, A. Mechanisms underlying beneficial plant-fungus
937 interactions in mycorrhizal symbiosis. *Nature Communications* **1**, 48 (2010).
938 73 Miyauchi, S. *et al.* Large-scale genome sequencing of mycorrhizal fungi
939 provides insights into the early evolution of symbiotic traits. *Nature*
940 *Communications* **11**, 5125 (2020).
941 74 Varga, T. *et al.* Megaphylogeny resolves global patterns of mushroom
942 evolution. *Nature Ecology & Evolution* **3**, 668-678 (2019).
943 75 Yang, Y. Y., Qu, X. J., Zhang, R., Stull, G. W. & Yi, T. S. Plastid
944 phylogenomic analyses of Fagales reveal signatures of conflict and ancient

chloroplast capture. *Molecular phylogenetics and evolution* **163**, 107232 (2021).

76 Whitmore, L. T. E. & Schaal, B. A. Interspecific gene flow in sympatric oaks. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 2540-2544 (1991).

77 Kremer, A. & Hipp, A. L. Oaks: an evolutionary success story. *New Phytologist* **226**, 987-1011 (2020).

78 Petit, R. *et al.* Chloroplast DNA variation in European white oaks phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management* **176**, 595-599 (2003).

79 Petit, R. *et al.* Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9996-10001 (1997).

80 Petit, R. J. & Excoffier, L. Gene flow and species delimitation. *Trends in Ecology & Evolution* **24**, 386-393 (2009).

81 Premoli, A. C., Mathiasen, P., Cristina Acosta, M. & Ramos, V. A. Phylogeographically concordant chloroplast DNA divergence in sympatric *Nothofagus* s.s. How deep can it be? *New Phytologist* **193**, 261-275 (2012).

82 Tsuda, Y., Semerikov, V., Sebastiani, F., Vendramin, G. G. & Lascoux, M. Multispecies genetic structure and hybridization in the *Betula* genus across Eurasia. *Molecular Ecology* **26**, 589-605 (2017).

83 Zhang, B. W. *et al.* Phylogenomics reveals an ancient hybrid origin of the persian walnut. *Molecular Biology and Evolution* **36**, 2451-2461 (2019).

84 Bock, R. Witnessing genome evolution: Experimental reconstruction of endosymbiotic and horizontal gene transfer. *Annual Review of Genetics* **51**, 1-22 (2017).

85 Hill, W. G. Disequilibrium among several linked neutral genes in finite population. II. Variances and covariances of disequilibria. *Theoretical Population Biology* **6**, 184-198 (1974).

86 Huerta-Sanchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197 (2014).

87 Huang, Y. *et al.* Megabase-scale presence-absence variation with *Tripsacum* origin was under selection during maize domestication and adaptation. *Genome Biology* **22**, 237 (2021).

88 Zhang, X. *et al.* The history and evolution of the Denisovan-*EPAS1* haplotype in Tibetans. *Proceedings of the National Academy of Sciences of the United States of America* **118**, e2020803118 (2021).

89 Cavender-Bares, J., Gonzalez-Rodriguez, A., Pahlich, A., Koehler, K. & Deacon, N. Phylogeography and climatic niche evolution in live oaks (*Quercus* series *Virentes*) from the tropics to the temperate zone. *Journal of Biogeography* **38**, 962-981 (2011).

90 Chen, D. *et al.* Phylogeography of *Quercus variabilis* based on chloroplast dna sequence in East Asia: Multiple glacial refugia and mainland-migrated island populations. *PLoS One* **7**, e47268 (2012).

91 Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nature Plants* **4**, 440-452 (2018).

92 Leroy, T. *et al.* Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytologist* **226**, 1171-1182 (2020).

93 Maxwell, L. M., Walsh, J., Olsen, B. J. & Kovach, A. I. Patterns of introgression vary within an avian hybrid zone. *BMC Ecology and Evolution* **21**, 14 (2021).

94 Hewitt, G. M. Hybrid zones - natural laboratories for evolutionary studies. *Trends in Ecology & Evolution* **3**, 158-167 (1988).

95 Gernandt, D. S., Resendiz Arias, C., Terrazas, T., Aguirre Dugua, X. & Willyard, A. Incorporating fossils into the Pinaceae tree of life. *American Journal of Botany* **105**, 1329-1344 (2018).

96 Rose, J. P., Toledo, C. A. P., Lemmon, E. M., Lemmon, A. R. & Sytsma, K. J. 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. *Systematic Biology* **70**, 162–180 (2021).

97 Truffaut, L. *et al.* Fine-scale species distribution changes in a mixed oak stand over two successive generations. *New Phytologist* **215**, 126-139 (2017).

98 Petit, R. J., Bodénès, C., Ducousso, A., Roussel, G. & Kremer, A. Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164 (2003).

99 Sork, V. L. *et al.* Phylogeny and introgression of California scrub white oaks (*Quercus* section *Quercus*). *International Oaks Journal* **27**, 61-74 (2016).

100 Quang, N. D., Ikeda, S. & Harada, K. Nucleotide variation in *Quercus crispula* Blume. *Heredity* **101**, 166-174 (2008).
101 Graham, A. The role of land bridges, ancient environments, and migrations in the assembly of the North American flora. *Journal of Systematics and Evolution* **56**, 405-429 (2018).
102 Suarez-Gonzalez, A., Lexer, C. & Cronk, Q. C. B. Adaptive introgression: A plant perspective. *Biology Letters* **14** (2018).
103 Abbott, R. J., Barton, N. H. & Good, J. M. Genomics of hybridization and its evolutionary consequences. *Molecular Ecology* **25**, 2325-2332 (2016).
104 Goulet, B. E., Roda, F. & Hopkins, R. Hybridization in plants: Old ideas, new techniques. *Plant Physiology* **173**, 65-78 (2017).
105 Mitchell, N. *et al.* Correlates of hybridization in plants. *Evolution Letters* **3**, 570–585 (2019).
106 Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Molecular Ecology* **25**, 2337-2360 (2016).
107 Bodenes, C. *et al.* Comparative mapping in the Fagaceae and beyond with EST-SSRs. *BMC Plant Biology* **12** (2012).
108 Cannon, C. H. & Petit, R. J. The oak syngameon: more than the sum of its parts. *New Phytologist* **226**, 978-983 (2020).
109 Chen, S. C., Cannon, C. H., Kua, C. S., Liu, J. J. & Galbraith, D. W. Genome size variation in the Fagaceae and its implications for trees. *Tree Genetics & Genomes* **10**, 977-988 (2014).
110 Kremer, A. *et al.* Genomics of Fagaceae. *Tree Genetics & Genomes* **8**, 583-610 (2012).
111 Neale, D. B., Martinez-Garcia, P. J., De La Torre, A. R., Montanari, S. & Wei, X. X. Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology* **68**, 457-483 (2017).
112 Staton, M. *et al.* Substantial genome synteny preservation among woody angiosperm species: comparative genomics of Chinese chestnut (*Castanea mollissima*) and plant reference genomes. *BMC Genomics* **16**, 744 (2015).
113 Manos, P. S. & Stanford, A. M. The historical biogeography of Fagaceae: Tracking the tertiary history of temperate and subtropical forests of the Northern Hemisphere. *International Journal of Plant Sciences* **162**, S77-S93 (2001).

114 Salojarvi, J. *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **49**, 904–912 (2017).

115 Mishra, B. *et al.* A reference genome of the European beech (*Fagus sylvatica* L.). *Gigascience* **7**, giy063 (2018).

116 Xing, Y. *et al.* Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). *Gigascience* **8**, giz112 (2019).

117 Sork, V. L. *et al.* High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. Preprint at bioRxiv http://10.1101/2021.1104.1109.439191 (2021).

118 Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).

119 Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 42 (2009).

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

121 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint at* http://arxiv.org/abs/1303.3997v2 (2013).

122 McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).

123 Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B. & van Nimwegen, E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution* **31**, 1077-1088 (2014).

124 Haudry, A. *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics* **45**, 891-898 (2013).

125 Hupalo, D. & Kern, A. D. Conservation and functional element discovery in 20 angiosperm plant genomes. *Molecular Biology and Evolution* **30**, 1729-1744 (2013).

126 Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* **45**, e18 (2017).

127 Qu, X. J., Moore, M. J., Li, D. Z. & Yi, T. S. PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* **15**, 50 (2019).

128 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

129 Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).

130 Ronquist, F. *et al.* Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**, 539-542 (2012).

131 Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* **34**, 772-773 (2017).

132 Akaike, H. New look at statistical-model identification. *IEEE Transactions on Automatic Control* **19**, 716-723 (1974).

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

134 Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317-3324 (2014).

135 PAUP*. Phylogenetic analysis using parsimony (* and other methods). v. Version 4. (Sinauer Associates, Sunderland, 2003).

136 Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).

137 dos Reis, M. & Yang, Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution* **28**, 2161-2172 (2011).

138 Rabosky, D. L. *et al.* BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution* **5**, 701-707 (2014).

139 Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* **31**, 1261-1271 (2014).

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

141 Xia, X. H., Xie, Z., Salemi, M., Chen, L. & Wang, Y. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution* **26**, 1-7 (2003).

142 Xia, X. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Molecular Biology and Evolution* **35**, 1550-1552 (2018).

143 McInerney, J. O. GCUA: General codon usage analysis. *Bioinformatics* **14**, 372-373 (1998).

144 Folk, R. A., Mandel, J. R. & Freudenstein, J. V. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology* **66**, 320-337 (2017).

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

146 Olave, M., Avila, L. J., Sites, J. W., Morando, M. & Freckleton, R. Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. *Methods in Ecology and Evolution* **9**, 121-133 (2017).

147 Than, C. & Nakhleh, L. Species tree inference by minimizing deep coalescences. *PLoS Computational Biology* **5**, e1000501 (2009).

148 Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast *D*-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* **21**, 584-595 (2021).

149 Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239-2252 (2011).

150 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).

151 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).

152 Solis-Lemus, C. & Ane, C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* **12**, e1005896 (2016).

153 Solis-Lemus, C., Bastide, P. & Ane, C. Phylonetworks: A package for phylogenetic networks. *Molecular Biology and Evolution* **34**, 3292-3298 (2017).

154 Hejase, H. A. & Liu, K. J. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics* **17**, 422 (2016).

155 Shen, X. X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**, 126 (2017).

156 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).

157 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

Acknowledgements

We thank Al Keuter for assistance with the layout design and select photos of cupules used in in Figure 1. This work was supported by the Guangdong Natural Science Funds for Distinguished Young Scholar (grant no. 2018B030306040 to B.W.), National Natural Science Foundation of China (grant no. NSFC 31971673 to B.W.), and US National Science Foundation (grant no. 1146102 to P.S.M.)

Data availability statement

Short reads of whole genome sequencing data have been submitted to NCBI (BioProject number xxx). Alignments of nuclear genes and plastomes are archived in the Dryad digital data repository (xx).

Author contributions

B.-F.Z., S.Y., M.K., P.S.M. and B.W. designed the project. B.-F.Z., S.Y., Y.-Y.L., Y.S., X.-Y.C. and Q.-Q.A. collected data. B.-F.Z., S.Y., A.A.C., Y.-Y.L., P.S.M. and B.W. analyzed data. B.-F.Z., S.Y., A.A.C., P.S.M. and B.W. wrote the paper. All authors read and approved the paper.

Competing interests statement

The authors declare no competing interests.

Fig. 1

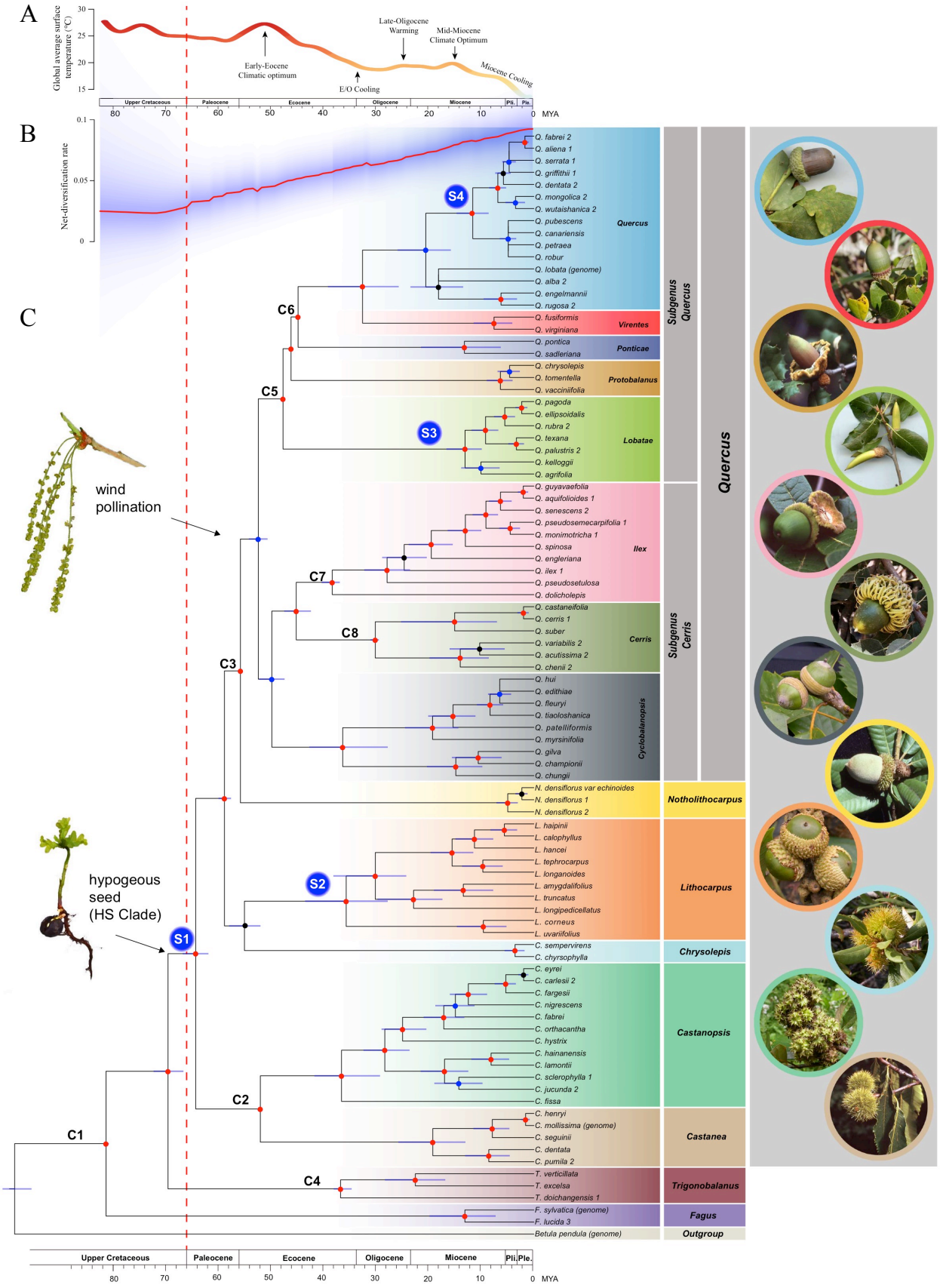
A
Global average surface temperature (°C)
Late-Oligocene Warming
Mid-Miocene Climate Optimum
Early-Eocene Climatic optimum
E/O Cooling
Miocene Cooling
Upper Cretaceous
Paleocene
Eocene
Oligocene
Miocene
MYA
B
Net-diversification rate
C
S4
S3
S2
S1
C6
C5
C7
C8
C3
C2
C1
C4
wind pollination
hypogeous seed (HS Clade)
Q. fabrei 2
Q. aliena 1
Q. serrata 1
Q. griffithii 1
Q. dentata 2
Q. mongolica 2
Q. wutaishanica 2
Q. pubescens
Q. canariensis
Q. petraea
Q. robur
Q. lobata (genome)
Q. alba 2
Q. engelmannii
Q. rugosa 2
Quercus
Q. fusiformis
Q. virginiana
Virentes
Q. pontica
Q. sadleriana
Ponticae
Q. chrysolepis
Q. tomentella
Q. vacciniifolia
Protobalanus
Q. pagoda
Q. ellipsoidalis
Q. rubra 2
Q. texana
Q. palustris 2
Q. kelloggii
Q. agrifolia
Lobatae
Subgenus Quercus
Q. guyavaefolia
Q. aquifolioides 1
Q. senescens 2
Q. pseudosemecarpifolia 1
Q. monimotricha 1
Q. spinosa
Q. engleriana
Q. ilex 1
Q. pseudosetulosa
Q. dolicholepis
Ilex
Q. castaneifolia
Q. cerris 1
Q. suber
Q. variabilis 2
Q. acutissima 2
Q. chenii 2
Cerris
Q. hui
Q. edithiae
Q. fleuryi
Q. tiaoloshanica
Q. patelliformis
Q. myrsinifolia
Q. gilva
Q. championii
Q. chungii
Cyclobalanopsis
Subgenus Cerris
Quercus
N. densiflorus var echinoides
N. densiflorus 1
N. densiflorus 2
Notholithocarpus
L. haipinii
L. calophyllus
L. hancei
L. tephrocarpus
L. longanoides
L. amygdalifolius
L. truncatus
L. longipedicellatus
L. corneus
L. uvariifolius
Lithocarpus
C. sempervirens
C. chrysophylla
Chrysolepis
C. eyrei
C. carlesii 2
C. fargesii
C. nigrescens
C. fabrei
C. orthacantha
C. hystrix
C. hainanensis
C. lamontii
C. sclerophylla 1
C. jucunda 2
C. fissa
Castanopsis
C. henryi
C. mollissima (genome)
C. seguinii
C. dentata
C. pumila 2
Castanea
T. verticillata
T. excelsa
T. doichangensis 1
Trigonobalanus
F. sylvatica (genome)
F. lucida 3
Fagus
Betula pendula (genome)
Outgroup
Upper Cretaceous
Paleocene
Ecocene
Oligocene
Miocene
Pli.
Ple.
MYA

Fig. 1 Phylogenetic relationships and divergence time estimation of Fagaceae inferred from analyses of 2124 nuclear genes. (A) The global climate curve during the last 82 million years (modified from Tierney et al. 2020). Major climate events were indicated. (B) Rate-through-time plot showing the net diversification rate (species/ million years) of Fagaceae. Red line is the median and the blue shadow represents the 95% confidence interval. (C) Chronogram derived from ASTRAL-III tree based on concatenated nuclear data. Nodes showing consistent relationships between ASTRAL-III, SVDquartets, maximum likelihood (ML), and MrBayes are marked with red (phylogenetic support ≥ 95 in all four analyses) and blue (support < 95% in any one of the four analyses). Nodes showing conflicting relationships among analyses are marked with black dots. Light blue bars on nodes represent 95% confidence intervals of divergence time estimates and dashed vertical red line represents the age of the Cretaceous-Paleogene boundary (66 million years ago). Geological timescale is shown at bottom. Fossil calibration nodes are indicated with C1- C8 (stem calibration node; Table S3). S1- S4 indicate four nodes where shifts in diversification rate were identified. Taxonomic labels of genera, subgenera and sections follow Manos et al. 2001, Manos et al. 2008; and Denk et al. 2017. Illustrations: lax catkins indicate the placement of the change from insect-pollination to wind-pollination that diagnoses the genus *Quercus*; hypogeous seed and seedling marks the origin of the HS clade. Images: representative cupule types are shown on the right. A consistent color scheme was used for taxonomic labels and image borders.

Fig. 2

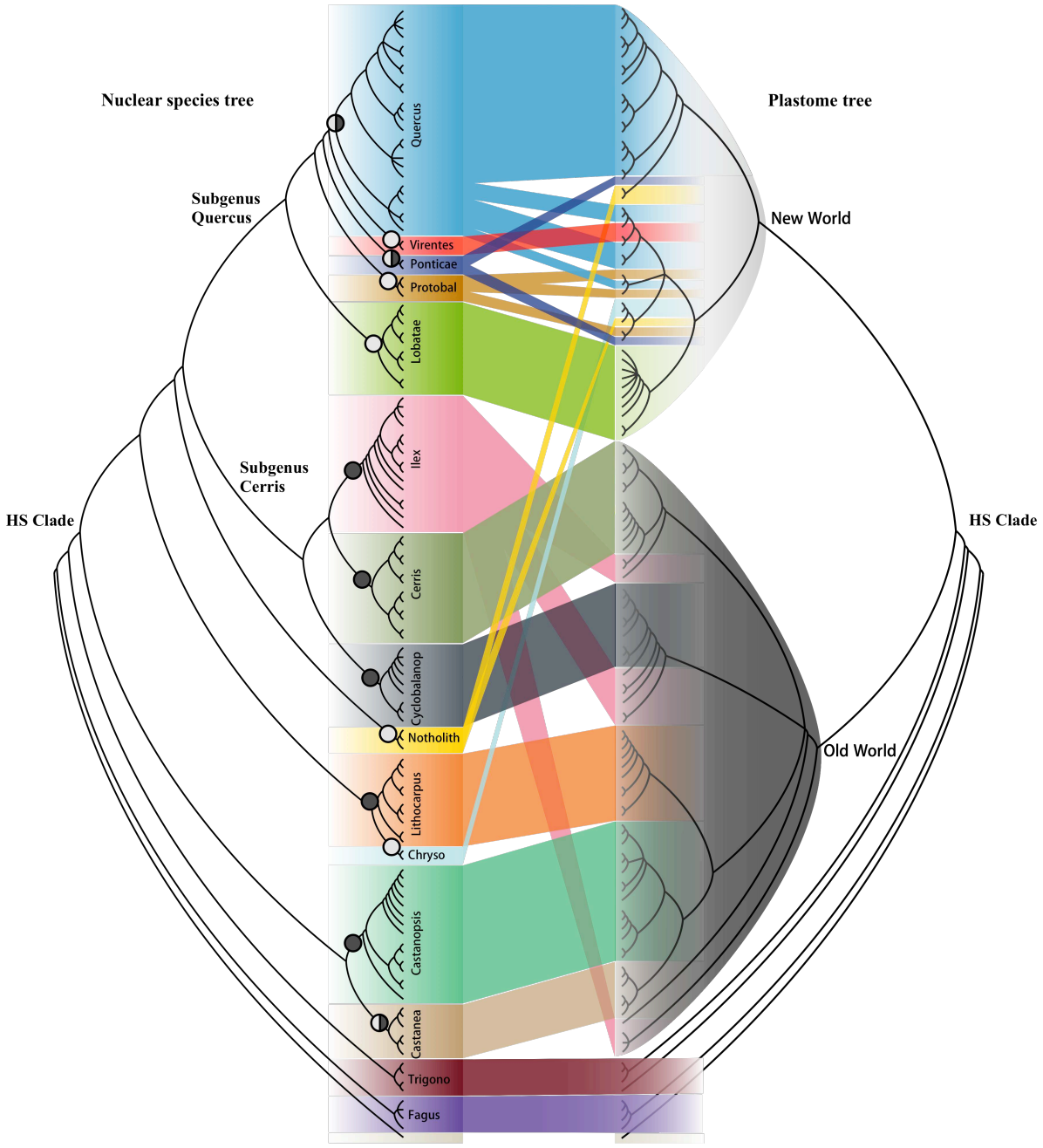

Fig. 2 Conflicts between nuclear (left) and plastome (right) species trees. Pie charts on nodes indicate the geographic distribution of the clade (black= Old World, white=New World). The HS clade consists of six genera divided into two major plastome clades: New World (light grey) and Old World (dark grey). Lineage colors are consistent with the color scheme in Fig. 1. Abbreviations: Protobal, section *Protobalanus*; Cyclobalanop, section *Cyclobalanopsis*; Notholith, *Notholithocarpus*; Chryso, *Chrysolepis*; Trigono, *Trigonobalanus*.

Fig. 3

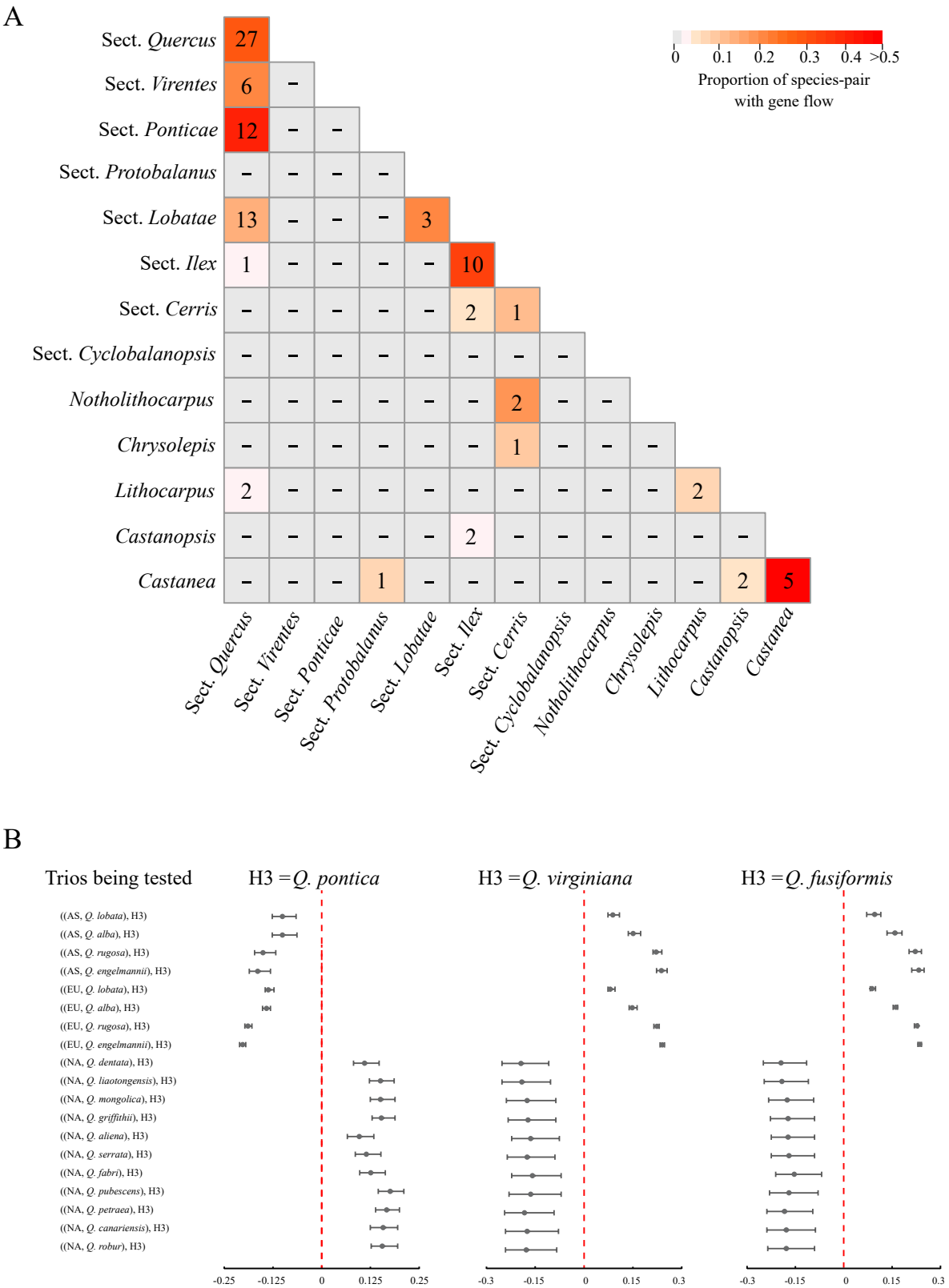
A
Sect. Quercus
Sect. Virentes
Sect. Ponticae
Sect. Protobalanus
Sect. Lobatae
Sect. Ilex
Sect. Cerris
Sect. Cyclobalanopsis
Notholithocarpus
Chrysolepis
Lithocarpus
Castanopsis
Castanea
0 0.1 0.2 0.3 0.4 >0.5
Proportion of species-pair with gene flow
B
Trios being tested
H3 = Q. pontica
H3 = Q. virginiana
H3 = Q. fusiformis
((AS, Q. lobata), H3)
((AS, Q. alba), H3)
((AS, Q. rugosa), H3)
((AS, Q. engelmannii), H3)
((EU, Q. lobata), H3)
((EU, Q. alba), H3)
((EU, Q. rugosa), H3)
((EU, Q. engelmannii), H3)
((NA, Q. dentata), H3)
((NA, Q. liaotongensis), H3)
((NA, Q. mongolica), H3)
((NA, Q. griffithii), H3)
((NA, Q. aliena), H3)
((NA, Q. serrata), H3)
((NA, Q. fabri), H3)
((NA, Q. pubescens), H3)
((NA, Q. petraea), H3)
((NA, Q. canariensis), H3)
((NA, Q. robur), H3)
-0.25 -0.125 0 0.125 0.25
-0.3 -0.15 0 0.15 0.3
-0.3 -0.15 0 0.15 0.3

Fig. 3 Gene flow between Fagaceae species revealed by *D*-statistic test. (A) Number of species-pairs with significant *D* value ($P < 0.01$ after Bonferroni correction) between sections of *Quercus* and other genera. Numbers on diagonal line indicate gene flow within each section or genus. Cells are colored based on the ratio of species-pairs with gene flow, with warmer colors indicating a higher proportion of species-pairs showing gene flow. For example, significant gene flow was detected for 12 species-pairs between sections *Quercus* (white oak) and *Ponticae*, representing 41% of tested species-pairs between these two sections. (B) Distribution of *D* values for white oaks vs. *Q. pontica* (left), and two species of section *Virentes*, *Q. virginiana* (middle) and *Q. fusformis* (right). Each line summarizes a set of *D*-statistic tests performed on trios in the format ((H1,H2),H3) with different H1 species and fixed H2 and H3 species (one of the three species above). Both H1 and H2 were white oaks, but represent different lineages. For example, if H2 was a North American white oak, then H1 was sampled from European or Asian white oaks. In each panel, points represent mean *D* values and error bars represent minimum and maximum *D* values across multiple tests. EU = European white oak; AS = Asian white oak; NA = North American white oak. A negative *D* value indicates gene flow between H1 and H3 while a positive *D* value indicates gene flow between the H2 taxon and H3. *Quercus pontica* shows a clear pattern of gene flow with EU and AS white oaks but not with NA white oaks while the opposite pattern is recovered for *Q. virginiana* and *Q. fusiformis*.

Fig. 4

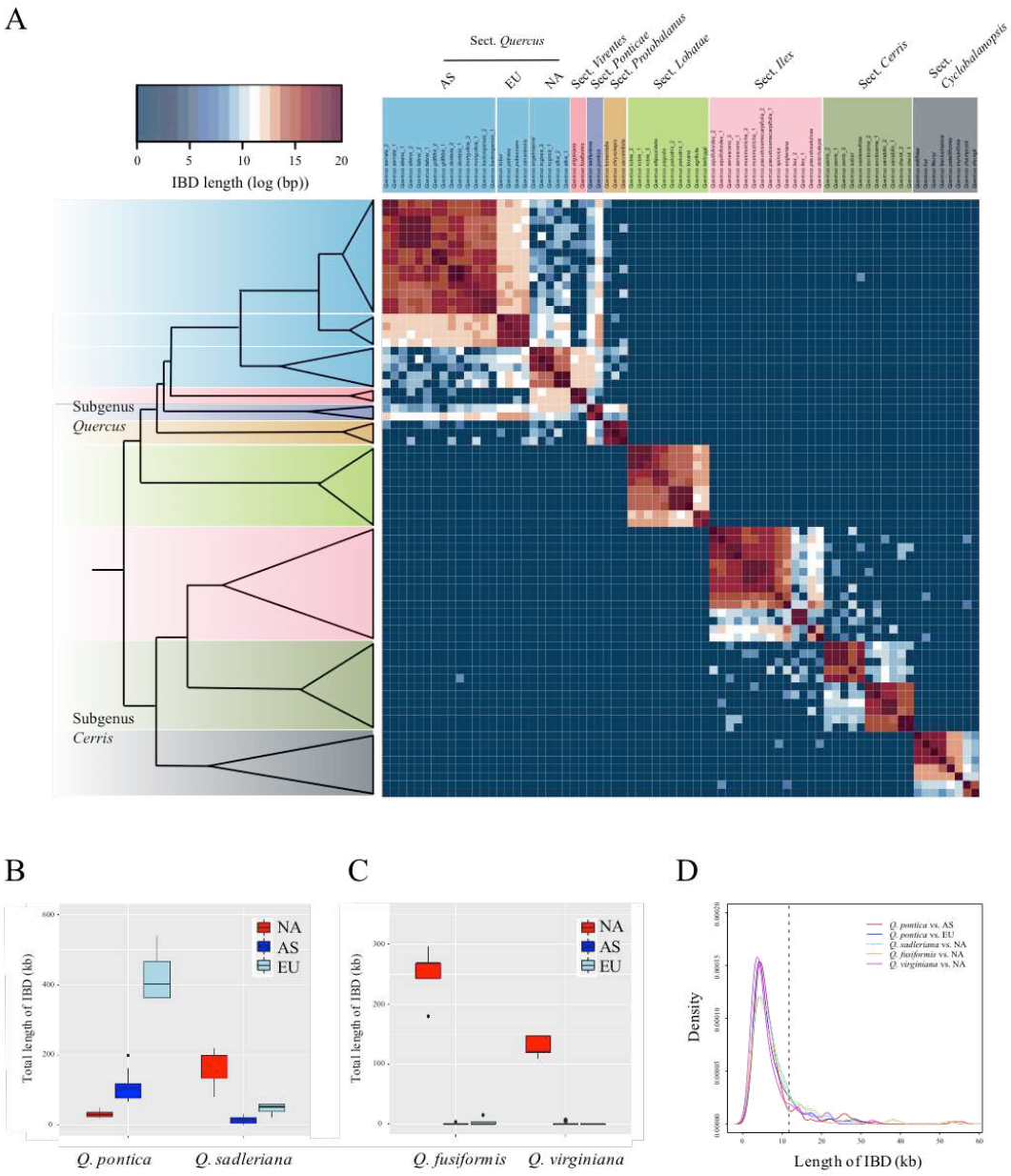

Fig. 4 Shared IBD blocks between *Quercus* species. (A) Heatmap indicating the total length of IBD blocks for each pair of comparisons. (B) and (C) box plots show shared total length of IBDs between sections *Ponticae* and *Quercus*, and between sections *Virentes* and *Quercus*. NA = North American white oak; EU = European white oak; AS = Asian white oak. (D) Kernel distribution of the length of shared IBD blocks between sections. Vertical black line (at 11724 bp) indicates the length at which IBD blocks are significantly longer than the expectation for selectively neutral introgressed fragments maintained in a population under a constant recombination rate of 10^{-8} per site per year, assuming an average divergence time of 3 million years ($P < 0.05$ after Bonferroni correction; see details in Methods).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf