Yang *et al. BMC Plant Biology* (2022) 22:611
https://doi.org/10.1186/s12870-022-03991-1

BMC Plant Biology

RESEARCH

Open Access

Plastome phylogenomics provide new perspective into the phylogeny and evolution of Betulaceae (Fagales)

Zhen Yang[1,2,3,4], Wenxu Ma[2,5], Xiaohong Yang[6], Lujun Wang[7], Tiantian Zhao[1,2,3,4], Lisong Liang[1,2,3,4], Guixi Wang[1,2,3,4] and Qinghua Ma[1,2,3,4*]

Abstract

Background: Betulaceae is a relatively small but morphologically diverse family, with many species having important economic and ecological values. Although plastome structure of Betulaceae has been reported sporadically, a comprehensive exploration for plastome evolution is still lacking. Besides, previous phylogenies had been constructed based on limited gene fragments, generating unrobust phylogenetic framework and hindering further studies on divergence ages, biogeography and character evolution. Here, 109 plastomes (sixteen newly assembled and 93 previously published) were subject to comparative genomic and phylogenomic analyses to reconstruct a robust phylogeny and trace the diversification history of Betulaceae.

Results: All Betulaceae plastomes were highly conserved in genome size, gene order, and structure, although specific variations such as gene loss and IR boundary shifts were revealed. Ten divergent hotspots, including five coding regions ($P_i > 0.02$) and five noncoding regions ($P_i > 0.035$), were identified as candidate DNA barcodes for phylogenetic analysis and species delimitation. Phylogenomic analyses yielded high-resolution topology that supported reciprocal monophyly between *Betula* and *Alnus* within Betuloideae, and successive divergence of *Corylus*, *Ostryopsis*, and *Carpinus-Ostrya* within Coryloideae. Incomplete lineage sorting and hybridization may be responsible for the mutual paraphyly between *Ostrya* and *Carpinus*. Betulaceae ancestors originated from East Asia during the upper Cretaceous; dispersals and subsequent vicariance accompanied by historical environment changes contributed to its diversification and intercontinental disjunction. Ancestral state reconstruction indicated the acquisition of many taxonomic characters was actually the results of parallel or reversal evolution.

Conclusions: Our research represents the most comprehensive taxon-sampled and plastome-level phylogenetic inference for Betulaceae to date. The results clearly document global patterns of plastome structural evolution, and established a well-supported phylogeny of Betulaceae. The robust phylogenetic framework not only provides new insights into the intergeneric relationships, but also contributes to a perspective on the diversification history and evolution of the family.

Keywords: Betulaceae, Plastome structure, Molecular dating, Biogeographic history, Character evolution

Background

Betulaceae (Fagales) is a relatively small but morphologically diverse family, comprising six extant genera and approximately 160 shrub or tree species [1]. Betulaceae species are mainly distributed in the northern temperate zone, and a few extend to subtropical highlands of

*Correspondence: mqhmary@sina.com

[1] State Key Laboratory of Tree Genetics and Breeding, Beijing 100091, China
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Central and South America [2, 3]. Morphologically, the family is characterized by typical synapomorphies, such as doubly-serrate leaves, compound catkins, and bract wrapped nuts [4], while each genus derives highly specialized traits which may play key roles in lineage-specific adaptive radiation [2]. As a geographically widespread and morphologically diverse group, Betulaceae has served as a model system for exploring taxonomic, systematic, and biogeographic issues [5–7]. To understand the diversification history and evolution of key traits, however, a robust phylogeny is required.

The generic delimitation and infra-familial relationships within Betulaceae have been examined through a series of approaches, including morphological characters [8, 9], fossil evidence [8, 10, 11], and molecular analyses [2, 4, 12]. It is now well established that Betulaceae is resolved into two subfamilies: Betuloideae (*Betula* L. and *Alnus* Mill.) and Coryloideae (*Corylus* L., *Ostryopsis* Decne., *Ostrya* Scop., and *Carpinus* L.). Nevertheless, different viewpoints on the phylogenetic relationships and morphological evolution among genera have been frequently proposed, with most controversies fastening on the divergence order among genera [9, 13], paraphyletic or sister relationships between *Alnus* and *Betula* [2, 14], phylogenetic status of *Ostryopsis* and *Corylus* [8, 15], and whether *Ostrya* and *Carpinus* were reciprocal monophyly or mutually nested [2, 16]. Throughout the above studies, the controversies can be attributed to morphologically parallel or convergent evolution, incomplete taxa sampling, and limited sequence variation (e.g., single or combined ITS, *Nia*, *mat*K, and *rbc*L). Thus, detailed relationships within Betulaceae need to be explored based on extensive taxon sampling and utilizing genome-level molecular sequences.

The relatively rich fossils for Betulaceae have promoted molecular-clock studies that inferred divergence ages among or within genera [4, 8, 12]. Nevertheless, those results varied greatly with different dating strategies and datasets. Assuming a split at either 45 or 80 Ma between *Alnus* and *Betula*, Bousquet et al. estimated the substitution rates of *rbc*L gene as 0.37 or 0.67×10^{-4} per site per million years [8]. Based on nuclear ribosomal ITS and 5S spacer sequences, Forest et al. inferred the median ages of crown lineages of Betulaceae, Betuloideae, and Coryloideae as 119.0 Ma, 109.3 Ma, and 70.2 Ma, respectively [12]. Using both chloroplast and nuclear DNA sequences, Grimm and Renner revealed the stem groups of Betulaceae could date from the upper Cretaceous, two subfamilies from the Paleocene, the crown group of six extant genera from the middle Miocene [4]. Although previous analyses have provided some insights into evolutionary history of Betulaceae, these estimates were inferred from unrobust phylogenetic framework constructed by a small number of DNA fragments, casting doubt on the inferred ages and hindering the development of a comprehensive understanding of origin and diversification.

Due to the highly conserved structure, uniparental inheritance, and composition of large numbers of single copy genes, plastome phylogenomics has been widely used in resolving problematic relationships within angiosperms [17–19]. Comparative genomics also provide a new perspective into plastome evolution, such as structural rearrangements, gene loss, and divergence hotspots. Particularly, the contraction or expansion of IRs has significant influence on the evolutionary rate of plastome [20–22]. Those events may function as effective phylogenetic signals. In Betulaceae, although some representative plastomes of each genus have been sporadically released, most studies mainly centered on describing the plastome characteristics of single species, and/or performing comparative and phylogenetic analyses based on a small number of plastomes (one or a few plastomes per genus) [23–26]. These studies may have contributed to the development of Betulaceae plastome resources, but are not sufficient to elucidate the overall structural variation and phylogenetic discordance, especially the non-monophyly of genera. So far, a pan-plastome study has not been conducted due to the imbalance between plastome number and species number in each genus. As the accumulation of Betulaceae plastomes, it is sufficient and necessary to conduct a comprehensive pan-plastome research to better understand the phylogeny, diversification, and evolution of the family.

In this study, we verify the power of whole plastomes to resolve the phylogenetic and evolutionary questions of Betulaceae by analyzing extensive plastomes of 109 accessions. All these plastomes were newly assembled or obtained from GenBank, representing six extant genera of Betulaceae and outgroups. Our objectives are as to: (1) elucidate plastome structural evolution of Betulaceae and reconstruct phylogenetic relationships among extant genera; (2) infer the origin and diversification history of Betulaceae; (3) trace the evolution of taxonomically important morphological characters across the family.

Results

Characteristics of Betulaceae plastomes

The plastomes of Betulaceae species varied little in genome size, ranging from 158,647 bp (*Betula pubescens*, MG386370) to 161,667 bp (*Corylus avellana*, MN082371). The average genome length decreased successively from *Alnus* to *Ostrya*: *Alnus* (160,690 bp), *Betula* (160,529 bp), *Corylus* (160,094 bp), *Ostryopsis* (159,630 bp), *Carpinus* (159,338 bp), and *Ostrya* (159,230 bp). All plastomes exhibited the typical

quadripartite structure of angiosperm plants, comprising two inverted repeat regions (IRa and IRb) (25,929–27,567 bp) separated by a small single copy region (SSC) (17,167–19,535 bp) and a large single copy region (LSC) (87,808–90,272 bp). The total GC content of these plastomes was highly similar (35.9–36.5%). In addition, a total of 121–136 genes were encoded, of which 14–23 genes were duplicated in the IR regions and 107–114 were single copy genes. Among the unique genes, 78–80 were protein-coding genes, four were rRNA genes, and 24–32 were tRNA genes. Of these protein-coding genes, 78 genes were commonly shared by all Betulaceae plastomes and two genes (*inf*A and *ycf*15) were lost in most of the plastomes (Tables 1, S1 and S2).

Comparative genomics and divergence hotspots

The global plastome divergence of major lineages within Betulaceae was visualized using mVISTA. The plastome-wide alignment revealed globally high similarity (Fig. S1). For the entire plastomes, the SSC and LSC regions displayed marked divergence than the IR regions. The variable proportion of non-coding regions was greater than that of protein-coding regions, and the divergence hotspots were mainly located in the intergenic spacer regions (Fig. 1). For the 78 coding regions, the P_i value for each locus varied from 0.0004 (*pet*G) to 0.0440 (*psa*I), with five loci (*psa*I, *ycf*1, *rpl*22, *psa*J, and *cem*A) over 0.02 (Table S3). Among the five coding hotspots, *rpl*22 gene showed moderate variation level ($P_i = 0.0249$) and appropriate nucleotide length (531 bp), making it an excellent potential DNA barcode. Its amino acid alignment across all plastomes was shown in Fig. 3B. For the 65 non-coding regions, nucleotide variability ranged from 0.0005 (*ycf*15-*trn*L_CAA) to 0.1100 (*trn*T_GGT-*psb*D), with the top five noncoding hotspots ($P_i > 0.035$) selected as *trn*T_GGT-*psb*D, *trn*E_TTC-*trn*T_GGT, *ndh*C-*trn*V_UAC, *trn*H-GTG_*psb*A, and *ycf*4_*cem*A (Table S3). These divergence hotspots can be used as potential DNA barcodes for phylogenetic analyses and species delimitation.

Boundaries between IR and SC regions

Betulaceae plastomes were relatively conserved, however, some structural variations were still identified, especially at the boundaries between IR and SC regions (Fig. 2). Junctions of the IRb/LSC region were located between the *rps*19 and *rpl*2 genes in 13 species, and within the *rps*19 gene in 17 species, with only *Corylus avellana* (MN082371) locating between *rps*19 and *rpl*22 genes. All Betulaceae plastomes had the IRb/SSC junction within the pseudogene (ψ) *ycf*1, ranging from 1 to 58 bp from the boundary. The *ndh*F gene overlapped 7 nucleotides with the ψ *ycf*1 in *Alnus cremastogyne*, 3 in *Ostryopsis davidiana* (MF375337), and 29 in *Alnus alnobetula* (MF136498), *Ostryopsis intermedia* (MG386376), and *Ostryopsis nobilis* (MG386378). The gene *ycf*1 spanned the IRa/SSC border in all plastomes, with the length of the *ycf*1 gene locating in the IRa region varying from 1,158 bp to 2,731 bp. The IRa/LSC boundary of 30 plastomes lay between the *rpl*2 and *trn*H genes, while that

Table 1 Sampling information, accession numbers, herbarium vouchers and structural features of 16 newly sequenced genomes

Species	Sampling locality	Length	LSC	SSC	IR	Gene	GC%	Accession Numbers	Herbarium vouchers
Ostryopsis davidiana	Neimeng, China	159,286	88,568	18,588	26,065	129	36.5	MH628451	Os-202104-dav
Alnus cremastogyne	Sichuan, China	160,538	89,074	19,094	26,185	129	36.4	MH628453	Al-202104-cre
Corylus mandshurica	Jilin, China	159,828	88,850	18,782	26,098	128	36.4	MH628448	Cor-202104-man
Corylus sieboldiana	Saitama, Japan	159,870	88,865	18,795	26,105	128	36.4	MH628456	Cor-202104-sie
Corylus wangii	Yunnan, China	159,825	88,743	18,870	26,106	129	36.5	MH628454	Cor-202104-wan
Corylus yunnanensis	Yunnan, China	160,085	89,043	18,840	26,101	127	36.4	MH628461	Cor-202104-yun
Corylus heterophylla var. *sutchuenensis*	Anhui, China	159,887	88,886	18,805	26,098	128	36.4	MH628447	Cor-202104-sut
Corylus jacquemontii	Sindh, Pakistan	159,862	88,846	18,818	26,099	128	36.4	MH628458	Cor-202105-jac
Corylus kweichowensis var. *brevipes*	Jiangxi, China	159,822	88,850	18,776	26,098	128	36.4	MH628452	Cor-202105-kwe
Corylus americana	Michigan, USA	159,912	88,895	18,813	26,102	128	36.4	MH628446	Cor-202105-ame
Corylus colurna	Tbilisi, Georgia	159,842	88,873	18,773	26,098	128	36.4	MH628457	Cor-202106-col
Corylus cornuta	Minnesota,USA	159,857	88,877	18,796	26,092	128	36.4	MH628460	Cor-202106-cor
Corylus cornuta var. *californica*	California, USA	159,915	88,867	18,844	26,102	128	36.4	MH628459	Cor-202106-cal
Corylus fargesii	Gansu, China	159,838	88,879	18,777	26,091	127	36.4	MH628462	Cor-202107-far
Corylus ferox	Yunnan, China	159,825	88,742	18,869	26,107	128	36.5	MH628455	Cor-202107-fer
Corylus ferox var. *thibetica*	Shanxi, China	160,358	88,357	18,769	26,616	128	36.4	MH628450	Cor-202107-thi

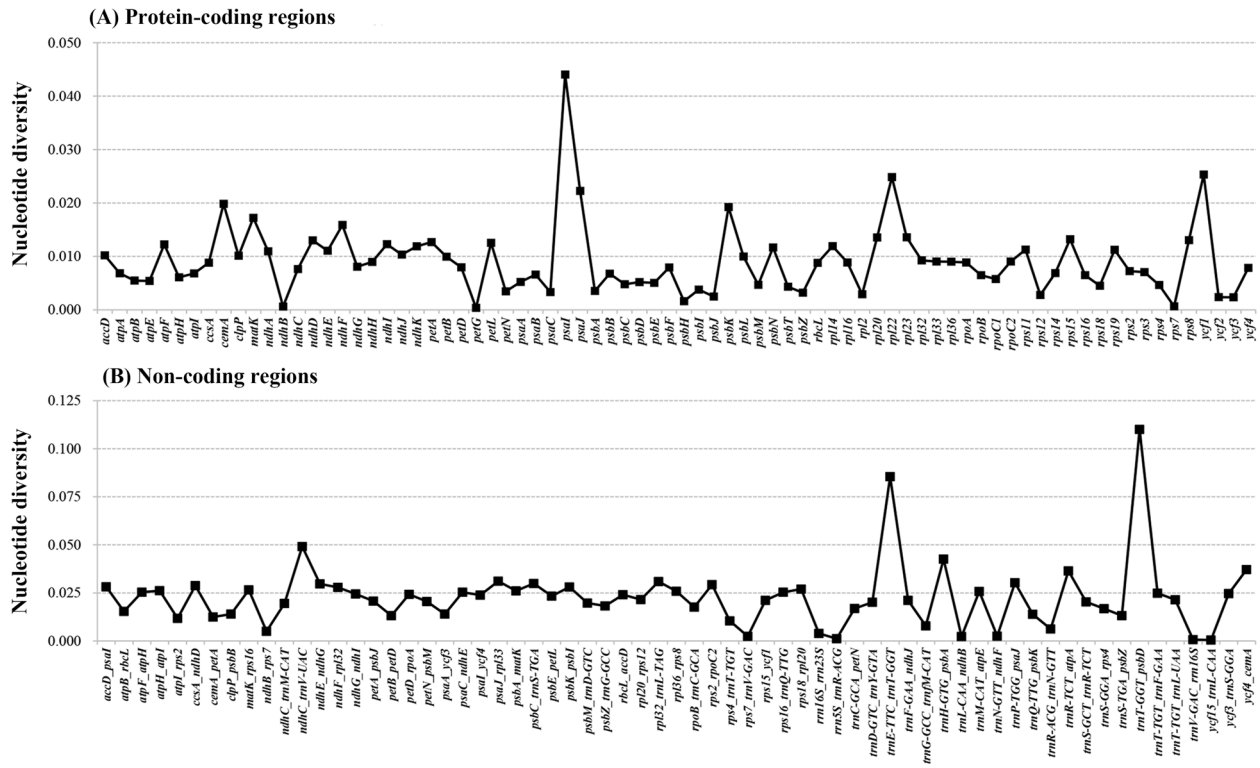

Fig. 1 Comparison of the nucleotide diversity (P_i) values across 31 Betulaceae plastomes (covering major lineages within Betulaceae). **A** Protein-coding regions. **B** Non-coding regions

of *Corylus avellana* (MN082371) was uniquely situated between *trn*H and *rps*19 genes.

Phylogenetic analyses

Four datasets, i.e., protein-coding regions (CDS), non-coding regions (CNS), whole plastomes (WP), and divergence hotspots (DH), were subjected to phylogenetic analyses to test whether different data characters were responsible for any changes observed in support and resolution of the integrated phylogenies. Data characteristics and best-fit models for Maximum likelihood (ML) and Bayesian inference (BI) analyses were presented in Table S4. ML and BI analyses of each dataset generated almost congruent topologies with generally high bootstrap support (BS) and posterior probability (PP) (Figs. 3A and S2, S3, S4). In the CDS phylogeny, monophyly of each genus was highly supported, of which *Alnus* and *Betula* were resolved as sister groups (BS/PP = 85/0.89) and formed the subfamily Betuloideae. *Corylus*, *Ostryopsis*, *Carpinus*, and *Ostrya* were included in the subfamily Coryloideae, with *Corylus* located at the stem of the phylogeny to form the tribe Coryleae (BS/PP = 100/1) and the other three genera formed the tribe Carpineae (BS/PP = 97/0.98). Within Carpineae, *Ostryopsis* was placed at the basal position and constituted sister group to *Carpinus*-*Ostrya*, while the latter two showed a well-supported sister relationship (BS/PP = 100/1) (Fig. 3A). The phylogenies of the CNS and WP datasets displayed highly identical topologies that supported the monophyly of four genera (*Alnus*, *Betula*, *Corylus*, and *Ostryopsis*) and the paraphyly between *Carpinus* and *Ostrya*. *Ostrya trichocarpa* was found to locate at the stem of *Carpinus*, whereas *Carpinus hebestroma*, *C. oblongifolia*, *C. purpurinervis*, and *C. cordata* were situated in the basal of *Ostrya* (Figs. S2 and S3). The DH phylogeny showed similar intergeneric relationships to those of CNS and WP datasets with the exception that *Ostrya trichocarpa* was located at the basal of the above four *Carpinus* species (Fig. S4).

Divergence time estimation

Molecular dating analysis based on CDS phylogeny indicated that the stem group of Betulaceae occurred during the upper Cretaceous (~ 89.15 Ma, 95% HPD = 73.03–112.03 Ma) (Table 2; Fig. 4). The crown age of Betulaceae and the split of Betuloideae and Coryloideae dated back to the Cretaceous-Paleogene boundary (~ 70.12 Ma, 95% HPD = 64.35–76.12 Ma). Within Betuloideae, *Alnus* and *Betula* diverged from each other shortly after the formation of the subfamily, approximately at 58.36 Ma (95% HPD = 37.23–72.96 Ma) during the Selandian age of

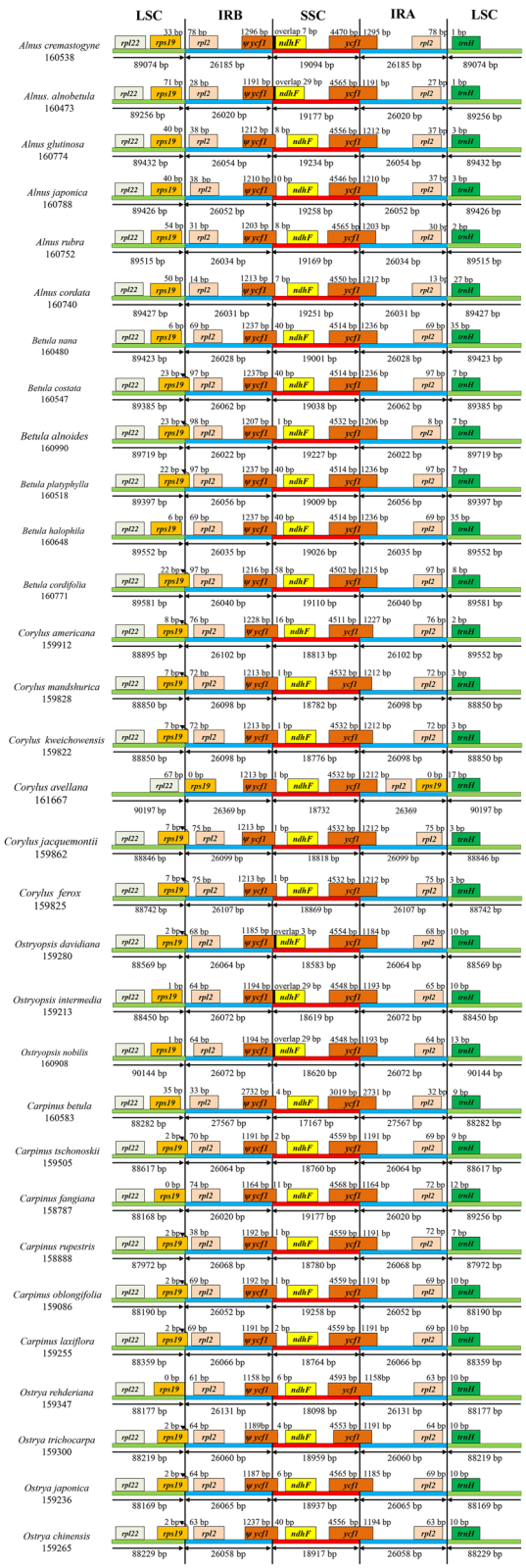

Fig. 2 Comparison of the IR/SC junctions among 31 Betulaceae plastomes (covering major lineages within Betulaceae)

the Paleocene. Within Coryloideae, *Corylus* and *Ostryopsis* successively diverged from the ancestral group in the middle (~43.50 Ma) and late Eocene (~36.97 Ma), respectively, while the sister genera *Carpinus* and *Ostrya* diverged from each other in the early Miocene (~20.48 Ma, 95% HPD = 13.56–27.75 Ma). The internal divergence within each genus occurred from the early Oligocene (~29.22 Ma) to the middle Miocene (~11.98 Ma), with *Corylus* diversifying the earliest and *Ostrya* the latest.

Ancestral area reconstruction

The likelihood implementation of Bayesian inference for discrete areas (BAYAREALIKE) reconstructed East Asia (ABC) as the ancestral area for the most recent common ancestor (MRCA) of Betulaceae (Fig. 5), although the exact subareas were not specified. In situ diversification of Betulaceae ancestors led to the formation of Coryloideae crown groups in southwestern East Asia (A), while westward dispersal resulted in the occurrence of Betuloideae crown group in southern Europe and the Mediterranean coast (E). Within Coryloideae, long-distance dispersals from A to E and North America (G, H), and subsequent vicariance events contributed to the intercontinental disjunction at the genus level. However, this pattern was not evident in *Carpinus* and *Ostrya* due to the limited sampling of non-Asian representatives. Within Betuloideae, E and eastern North America (H) were revealed as ancestral areas for *Alnus* and *Betula*, respectively. Likewise, long-distance dispersals from original centers to other parts were also observed, e.g., E to Central Asia (D) and then to A and central and eastern China (B); H to northern Europe (F) and then to E and D; as well as mutual exchanges between E and H. Overall, the connection/fracture of three important paths, i.e., North Atlantic Land Bridge, Beringian Land Bridge, and Mediterranean-eastern Himalayas/western China corridor, have played important roles in the intercontinental disjunction of Betulaceae.

Morphological characters evolution

Ancestral states for 14 morphological characters are summarized in Fig. 6 and Figs. S5, S6, S7, including five flower characters (1–5), three anatomical characters (6–8), three leaf characters (9–11), and three fruit characters (12–14). For flower characters, it is unambiguously that the ancestor of Betulaceae had bisexual inflorescence, raceme infructescence, staminate perianth or pistilloide in male floret present, separated thecae and partly divided filaments. The aggregated infructescence seems to have evolved from raceme infructescence and then reversed in *Corylus* and *Ostryopsis* (Fig. S5), while the other characters have altered

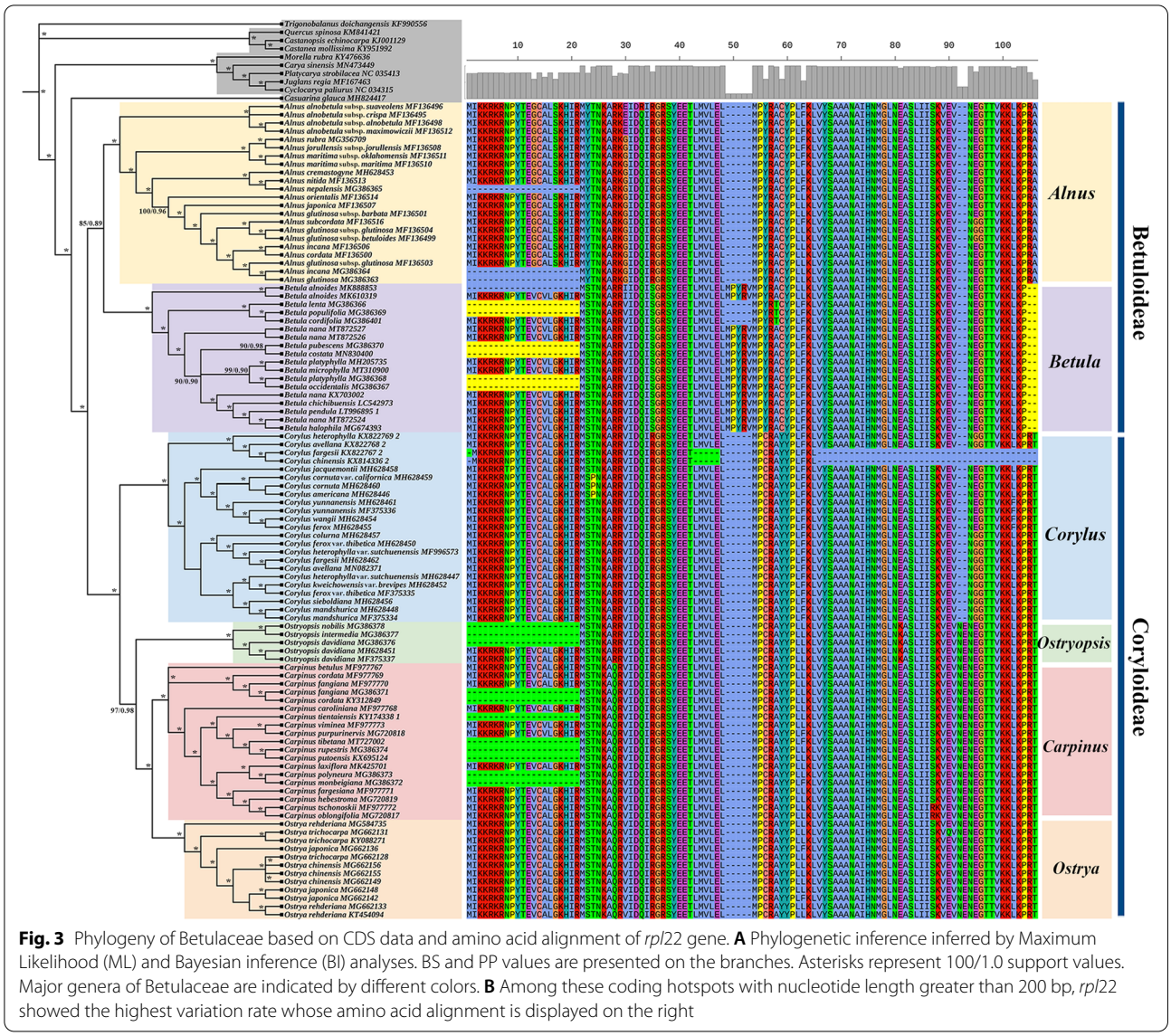

Fig. 3 Phylogeny of Betulaceae based on CDS data and amino acid alignment of *rpl*22 gene. **A** Phylogenetic inference inferred by Maximum Likelihood (ML) and Bayesian inference (BI) analyses. BS and PP values are presented on the branches. Asterisks represent 100/1.0 support values. Major genera of Betulaceae are indicated by different colors. **B** Among these coding hotspots with nucleotide length greater than 200 bp, *rpl*22 showed the highest variation rate whose amino acid alignment is displayed on the right

Table 2 Estimated divergence times for main clades within Betulaceae

Clades	Stem age (Ma)	95% HPD	Crown age (Ma)	95% HPD
Betulaceae	89.15	112.03–73.03	70.12	76.12–64.35
Subfam. Betuloideae	70.12	76.12–64.35	58.36	72.96–37.23
Alnus	58.36	72.96–37.23	27.43	42.38–16.34
Betula	58.36	72.96–37.23	20.94	37.35–8.23
Subfam. Coryloideae	70.12	76.12–64.35	43.50	49.46–38.03
Corylus	43.50	49.46–38.03	29.22	40.95–20.10
Ostryopsis	36.97	44.83–26.91	17.67	25.90–6.39
Carpinus	20.48	27.75–13.56	15.52	22.29–10.72
Ostrya	20.48	27.75–13.56	11.98	18.29–5.41

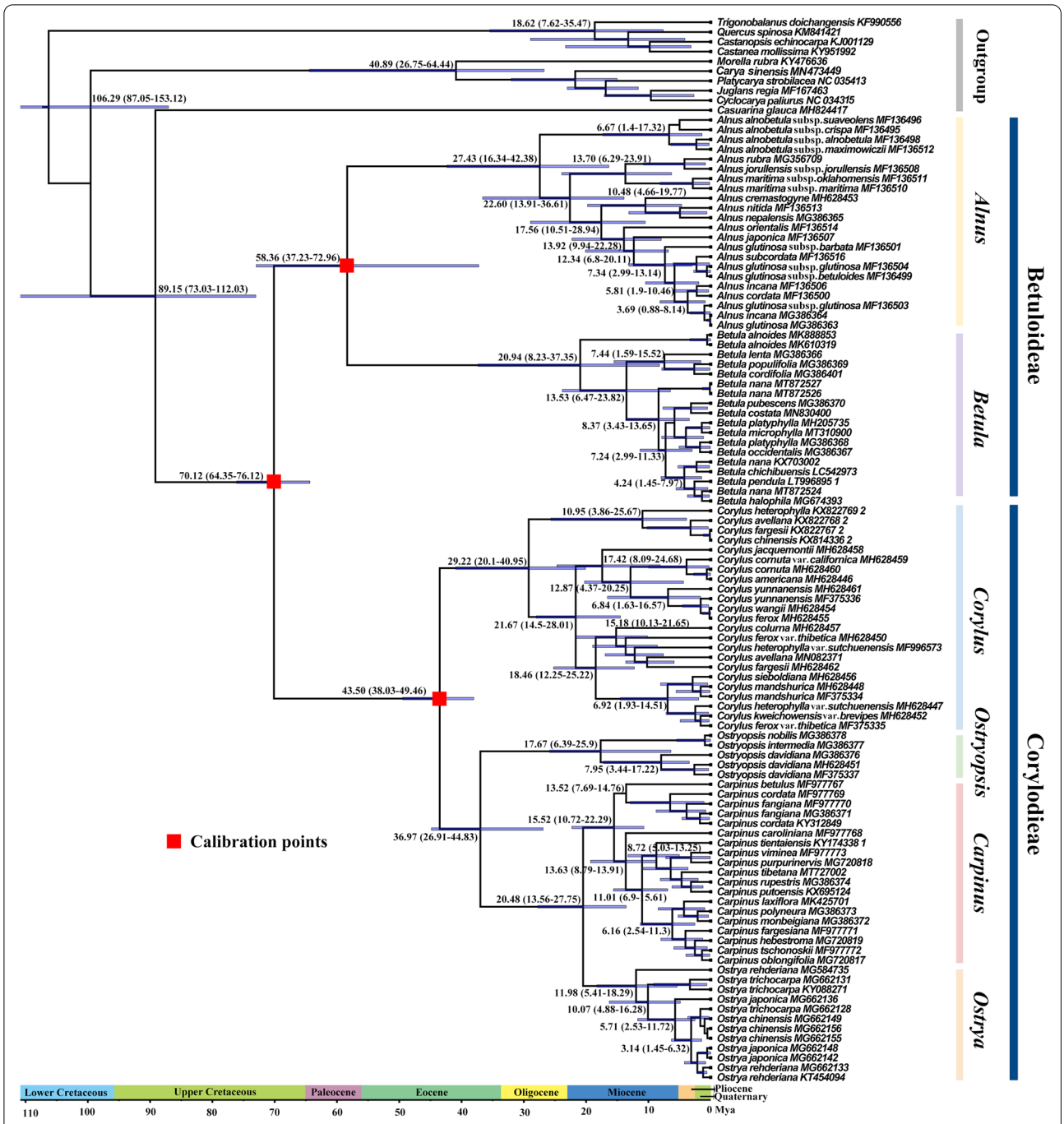

Fig. 4 BEAST chronogram of divergence times for Betulaceae based on CDS data. Fossil calibrations are indicated by red boxes. Numbers above the tree branches represent median divergent ages and 95% HPD intervals. The blue bars represent the 95% highest posterior density of node ages

their states in parallel in different lineages. For anatomical characters, their ancestral states (e.g., scalariform vessel perforation, present tracheids, and absent tyloses) have evolved independently in different genera (Fig. S6). Regarding leaf characters, the presence of stomatal apparatus and embedded glands on leaves, and compound teeth were inferred as ancestral states, with embedded glands absent in subfamily Coryloideae and then inverted in *Ostryopsis* (Fig. S7). As for fruit characters, the ancestral states of winged diaspore and epigeal seed germination retained in all the extant genera except *Corylus*, whereas the shape of fruit bracts have

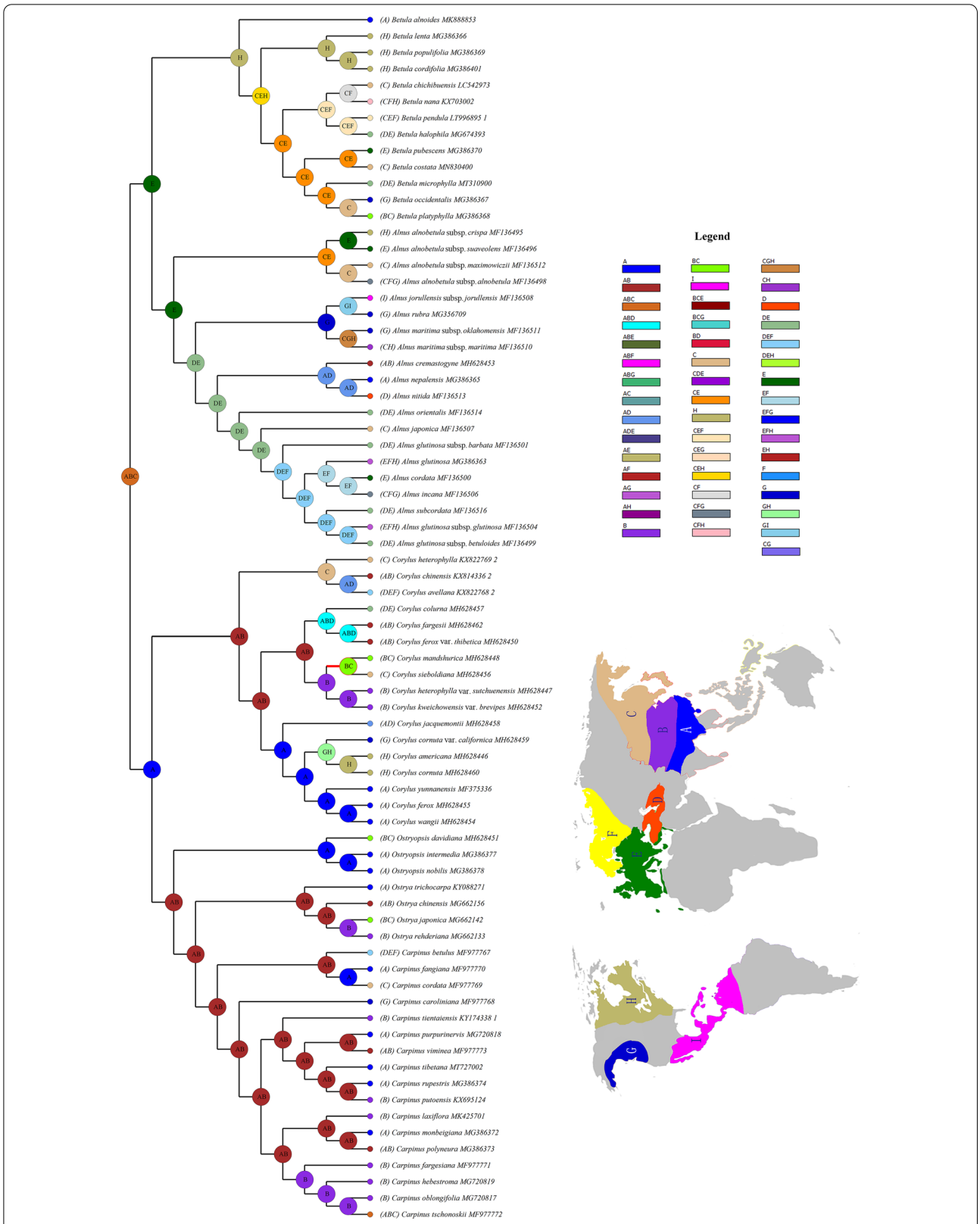

Fig. 5 Ancestral area reconstruction based on the likelihood implementation of Bayesian inference for discrete Areas (BAYAREALIKE). Current distributions are indicated before the species names. The inserted map shows the contemporary distribution of Betulaceae species, covering nine major floristic divisions (**A-I**). Numbers and colors in the legend refer to extant and possible ancestral areas, and combinations of these

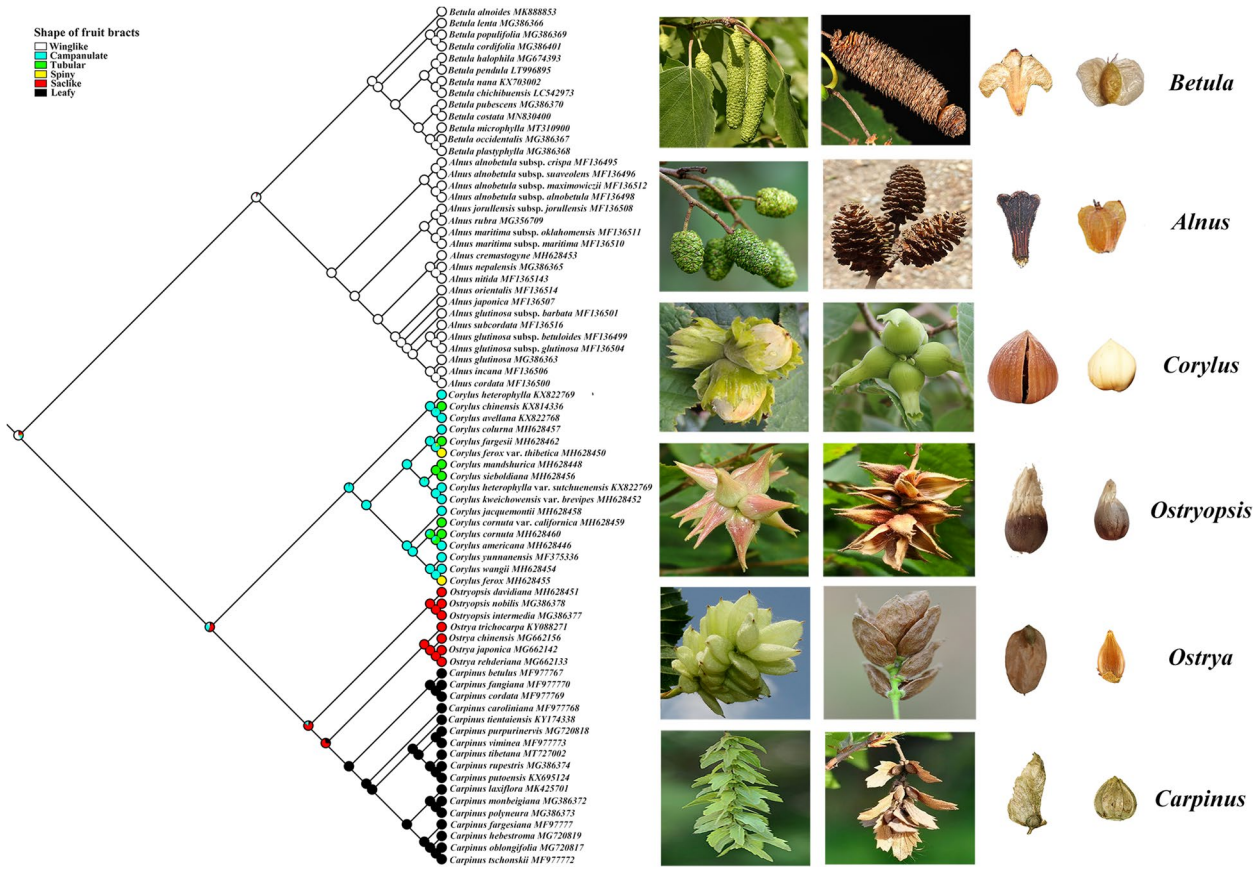

Fig. 6 Ancestral state reconstruction of shape of fruit bracts. The images on the right show the typical characteristics of infructescence, fruits, and bracts for each genus, respectively

evolved from winglike into multiple forms, especially in *Corylus* (Fig. 6).

Discussion

Plastome structural evolution of Betulaceae

Previous researches have suggested that the plastome size of angiosperms ranges from 107 kb in Pinaceae to 218 kb in Geraniaceae, and the size of the IR region is 20–30 kb [27, 28]. Our results showed that the plastomes of Betulaceae were located at the larger end of the angiosperm organelle genome, with Betuloideae having relatively larger genome (*Alnus*: ~ 160,690 bp; *Betula*: ~ 160,529 bp). By contrast, Coryloideae especially tribe Carpineae evolved relatively smaller genome (*Ostryopsis*: ~ 159,630 bp; *Carpinus*: ~ 159,338 bp; *Ostrya*: ~ 159,230 bp), whereas the transitional *Corylus* owned medium plastome size (~ 160,094 bp). Contraction and expansion of IR regions are very common in the process of evolution, and has been proved to be an important source of plastome size variation [20, 22, 29]. In our research, however, insignificant length variation for IR was detected among Betulaceae plastomes, with species of *Corylus* and *Carpinus* expanded slightly (Fig. 2; Table S1). This is normal because the high conservation of the IR region is also crucial for the stability of plastome structure [30]. Correspondingly, the length variation of LSC or SSC regions could have contributed to the differences in genome size. For example, *Alnus* and *Betula* had larger LSC and SSC regions than other genera, and although *Ostrya* possessed the largest SSC region in the Coryloideae, its LSC region was the smallest (Table 1, Table S1). Comparable results were also discovered in other taxa, such as Apiales [20], *Eucalyptus* [31], *Ampelopsis* [32], and those early diverging eudicots [33]. Gene loss occurs frequently in plastomes, for instance, the genes *rpl*22, *rps*16, *rpl*23, *acc*D, *ycf*1, and *inf*A were utterly or partially lost in the plastomes of legumes [34], and *acc*D, *ycf*1 were entirely missed in Poaceae [35]. In Betulaceae, the gene content varied slightly among species (121–134), with major differences lying in the numbers of tRNA (24–32) and protein-coding genes (78–80) (Tables 1, S1 and S2). Particularly, we discovered that

two genes (*inf*A and *ycf*15) were lost in most of the Betulaceae plastomes (Table S2). Despite the changes in gene content, Betulaceae plastomes were highly conserved in terms of genome structure with only trivial IR expansion detected (Figs. 2 and S1). In addition, synteny analysis demonstrated that IR regions were more conservative than two SC regions, which is accordance with the conclusion that the accumulation of point mutations in the IR region is slower than the SC regions [36].

Plastome comparative genomics has been confirmed to facilitate the development of divergence hotspots which can be used for species delimitation and phylogenetic research of different levels [37, 38]. Relevant researches have shown that some coding genes of plastomes were efficient in resolving complex phylogenetic relationships of special plant taxa, for instance, *ndh*K, *psa*I and *rpl*22 in *Allium* [17], *psa*I, *pet*B, and *rps*16 in *Notopterygium* [39]. Furthermore, more studies revealed that non-coding regions were more variable than coding regions, and had higher resolution in species identification of related groups, e.g., *trn*T-*trn*L, *pet*D-*rpo*A and *ycf*4-*cem*A displayed apparent divergence in Veroniceae species [40], while *rpl*32-*trn*L, *rpo*B-*trn*C, *psa*C-*ndh*E, and *clp*P-*psb*B were highly variable in *Phalaenopsis* species [41]. Two widely used plastome markers, *rbc*L and *mat*K have been revealed to have limited resolution in previous phylogenetic studies of Betulaceae [2, 15, 42], conforming to their low nucleotide variation as shown in our research ($P_{rbcL}=0.0088$, $P_{matK}=0.0092$) (Table S3). In the present study, both nucleotide diversity and mVISTA analysis revealed that the variation level of non-coding regions was significantly higher than coding regions (Figs. 1 and S1), which is consistent with previous results of most angiosperms [17, 37]. Correspondingly, we identified five non-coding hotspots ($P_i>0.035$) that have not been reported in previous studies, i.e., *trn*T_GGT-*psb*D, *trn*E_TTC-*trn*T_GGT, *ndh*C-*trn*V_UAC, *trn*H-GTG_*psb*A, and *ycf*4_*cem*A; simultaneously, five coding genes, i.e., *psa*I, *ycf*1, *rpl*22, *psa*J, and *cem*A, exhibited higher diversity ($P_i>0.02$) than other genes (Fig. 1; Table S3). Ten divergence hotspots can serve as candidate DNA barcodes for inferring phylogenetic relationships and intergeneric divergence of Betulaceae.

Phylogenetic implications of plastome-scale dataset

Previous studies based on molecular and morphological data have contributed to an improved grasp on the taxonomy and intergeneric relationships of Betulaceae. Nevertheless, wide controversies still existed in terms of genera delimitation and their evolutionary relationships. Within Betuloideae, *Betula* and *Alnus* were either treated as sister monophyly [2, 4, 12] or assigned as paraphyly in which *Alnus* lay in the basal position [14, 43]. The situation was more complicated within Coryloideae, especially the phylogenetic placement of *Ostryopsis* as well as the generic relationships between *Carpinus* and *Ostrya*. *Ostryopsis* was located at the basal of the *Carpinus-Ostrya* clade based on nuclear ITS [2, 12, 16], chloroplast *rbc*L [2] and morphology [2, 8] data, while the genus formed sister clade with *Corylus* in the multi-gene (ITS, *mat*R, *rbc*L, *trn*L) partitioned phylogeny [44] and the *mat*K tree [15]. Inferred from ITS phylogeny, *Carpinus* and *Ostrya* were either mutual monophyly [2], or *Carpinus* was treated as paraphyly with *Ostrya* inserted within it [16]. Besides, the combined chloroplast fragments (*psb*A-*trn*H, *trn*L-*trn*F, and *mat*K) inferred *Ostrya* as paraphyly [42]. Such kind of phylogenetic controversies can result from a variety of factors, including incomplete taxa sampling, limited sequence variation, and heterogeneous evolutionary rate among genes [45, 46]. In this research, we obtained the most extensively sampled and well-resolved phylogeny for Betulaceae based on pan-plastome data. Our trees presented an essential improvement in internode resolution compared with previous phylogenetic inferences. With the results, some long-standing controversies have been clarified. The phylogenomic backbone inferred from all datasets (CDS, CNS, WP, and DH) highly supported the division of Betulaceae into Betuloideae and Coryloideae, and reciprocal monophyly between *Alnus* and *Betula* (Figs. 3A and S2, S3, S4). Particularly, the phylogenetic position of *Ostryopsis* was well established, with *Corylus*, *Ostryopsis*, *Carpinus-Ostrya* forming successive sister lineages in Coryloideae. The evolutionary relationships between *Carpinus* and *Ostrya* were also illuminated. The two genera formed mutual paraphyly in three phylogenies (CNS, WP, and DH), but were well supported as reciprocal monophyly in the CDS tree. This phylogenetic conflict actually reflects the heterogeneity of evolutionary rate in different genes and structural regions. Non-coding regions, especially intergenic spacers, have higher evolutionary rate than protein-coding regions, while divergence hotspots represent the most variable level in the whole plastome. Hence, despite high-resolution phylogenies inferred from these concatenated super matrices, various evolutionary rates in different regions could inevitably lead to rampant phylogenetic discordance at all levels of angiosperm phylogeny [46].

Just as the paraphyletic relationships between *Carpinus* and *Ostrya*, non-monophyly is relatively common when multiple accessions of each taxon are applied in phylogenetic studies [47]. Such non-monophyly usually reflects two genealogical processes, i.e., incomplete lineage sorting and introgressive hybridization, which are hard to distinguish due to their similar phylogenetic signature [48]. The relatively late differentiation implies that

incomplete sorting of ancestral polymorphisms may play an important role in maintaining the paraphyletic status of the two genera. In turn, incomplete lineage sorting could result from rapid radiation during early diversification. Accordingly, it is not surprising that previous single-gene studies or even our phylogenomic research have placed *Carpinus* and *Ostrya* into different phylogenetic arrangements. Hybridization/introgression can result in horizontal transfer of maternally inherited plastomes into introgressive species when closely related species are sympatric distribution and reproductive compatibility [49]. Introgression-induced chloroplast capture has been proved as a particular mechanism to generate distorted phylogenetic relationships, in which introgressive taxa always present typical geographic clustering [50]. In Betulaceae, natural hybridization is frequently found to occur within a genus [50–52], but few has been reported between two different genera. Recently, homploid hybrid speciation between ancestors of *Carpinus* and *Ostrya* was revealed by genetic evidence [53], indicating that intergeneric hybridization could be achieved during the initial stages of differentiation. Similar cases have been reported in other plant taxa, e.g., a common ornamental plant known as "×*Heucherella*" in the nursery trade stems from the hybridization between *Heuchera* and *Tiarella* [54]. In our plastome phylogenetics, the nonrandom phylogenetic clustering (reciprocal basal lineages of each other) among *Carpinus* and *Ostrya* species (Figs. S2 and S3) may further corroborate the hypothesis of chloroplast capture caused by intergeneric introgressive hybridization. However, the geographic clustering of intergeneric species is not obvious probably because ancient chloroplast capture between genera may be blurred by frequent plastome transfer among intrageneric species. In total, our phylogenies are relatively robust and can well reflect the actual evolutionary relationships of the family.

Diversification and biogeographic history

A well-resolved phylogeny, extensive taxon sampling, and reliable fossil calibrations have been shown to be crucial for estimating divergence times and biogeographic history [55–57]. Although relatively rich fossils for Betulaceae have encouraged molecular dating studies, no agreement has been reached, mainly due to different sampling scales, molecular markers, and fossil calibrations [4, 12, 44, 58]. In this research, utilizing an extensively sampled, well-resolved phylogeny, and reliable fossil calibrations, we re-evaluated the divergence history of major lineages within Betulaceae (Fig. 4; Table 2). Our estimated ages were much younger than those of Forest et al. in many clades, especially the crown group of Betulaceae (~119.3 Ma), Betuloideae (~109.0 Ma), and Coryloideae (~70.2 Ma) [12]. By contrast, estimates for most basal nodes were nearly ten million years older than those of Grimm and Renner when they calibrated the stem lineage of Betulaceae with a 71 Ma-old flower fossil [4]. Comparably, our results were compatible with a recent study which also applied whole plastomes to infer the divergence time within Betulaceae [58]. However, their sampling only involved a few species in each genus and their basal lineages were largely missed, which may have led to the under-estimation of the crown ages of extant genera. Overall, our estimates are relatively robust and reliable. On one hand, the phylogenetic frame was constructed based on pan-plastome data that was more informative than single or several fragments so that less bias was introduced. On the other hand, the estimated ages for major nodes were congruent with reliable fossil evidences. For instance, the earliest pollen fossils belonging to the complex groups of *Betula* and *Alnus* were confirmed to have occurred in the Santonian/Campanian of Japan and North America [59], which were coetaneous with the 71 Ma-old flower fossils used in our calibration. They jointly indicated that the crown of Betulaceae occurred during the upper Cretaceous, as inferred by Xiang et al. (~74.99 Ma) and the present study (~70.12 Ma) [60]. Likewise, fruit fossils from the Republic flora of northeastern Washington were the oldest known occurrences of two extant genera (*Carpinus* and *Corylus*) [10], indicating a middle-Eocene origin of Coryloideae. Coincidentally, our estimates revealed an almost identical age for the crown group of Coryloideae (43.50 Ma). Therefore, our results present a more realistic scenario for the divergence history of Betulaceae.

Betulaceae is believed to have originated in central/western China in the late Cretaceous (~70 Ma) [61]. The assumption is supported by extensive fossil records and the fact that all six extant genera and almost one-third of the species of Betulaceae are native to this region. This biogeographic origin is further evidenced by ancestral area reconstructions and divergence age estimates conducted in the present study, that is, the MRCA of Betulaceae occurred in East Asia at the Cretaceous-Paleocene boundary (~70.12 Ma) (Figs. 4 and 5). During this period, a multitude of species have experienced rapid diversification on account of global cooling and the emergence of new habitats [62, 63]. Hence, it is very likely that adaptive radiation triggered by environmental changes has contributed to the initial differentiation of Betulaceae. Despite the East Asian origin of Betulaceae, both in situ and allopatric diversification were detected, with A and E inferred as the original center for the crown groups of Coryloideae and Betuloideae, respectively. Apparently, the preexisting corridor between eastern Himalayas/western China and European-Mediterranean region has facilitated ancient intercontinental dispersal, thus

generating the European ancestors of Betuloideae. The strong floristic connection between East Asia and Europe has also been demonstrated in previous biogeographic studies [50, 64]. The intergeneric divergence within Betuloideae (~58.36 Ma) was relatively earlier than that of Coryloideae, which was probably caused by contrasting intercontinental habitat differences between *Alnus* and *Betula*, i.e., E and H. The favorable environments (warm and humid climate) during the middle Paleocene may have further contributed to their distribution around the northern hemisphere [65]. Particularly, rich macrofossil records of *Alnus*-like species indicated that the ancestors of this genus have dominated the flora of North America since the Paleocene [66]. By comparison, the ancestors of the four genera (*Corylus*, *Ostryopsis*, *Carpinus*, and *Ostrya*) within Coryloideae were sympatric distribution in East Asia, and the relatively homogeneous habitats could have delayed their intergeneric divergence, roughly in the middle Eocene (~43.50 Ma). With the global cooling at that time, adaptive radiation should have played an essential role in promoting this divergence processes. From the late Eocene to the Pliocene, a series of geoclimatic events, such as the uplift of the Qinghai-Tibet Plateau, the formation of the Asia monsoon, as well as the Quaternary glaciation cycles [67–69], may have driven the lineage diversification and radiative speciation within each genus. Notably, the connection/fracture of three important paths, i.e., North Atlantic Land Bridge, Beringian Land Bridge, and Mediterranean-eastern Himalayas/western China corridor, have played important roles in the intercontinental disjunction of Betulaceae.

Evolution of key characters of Betulaceae

The acquisition of new morphological traits has influenced the diversification of various plant groups [70, 71]. The six genera within Betulaceae are typical representatives to increase the number of such cases. State reconstructions revealed that bisexual and raceme infructescence with staminate perianth and pistilloide in male floret were the most likely ancestral traits for Betulaceae (Fig. S5). Most of these flower characters have changed their states only once during evolution in different genera except that raceme infructescence evolved into aggregated infructescence and then reversed in *Corylus* and *Ostryopsis*. Generally, flower characters for the six genera have a tendency towards simplification. In recent studies, the hypothesis that the evolution of floral variation is driven by pollinator transfer has been confirmed by phylogenetic evidences [22, 72, 73]. In Betulaceae, two types of pollinators, namely, anemophilous and entomophilous pollination, were recognized. Although both ways were equally important for pollination of Betulaceae in the early Tertiary, enhancement of fecundity and pollination efficiency gradually promoted insect-pollination as the dominant mode [2]. Accordingly, each genus has evolved corresponding floral features independently. Currently, few studies have been carried out on the interaction between Betulaceae plants and pollinators. In addition to overall survey on plant-pollinator interaction, relevant ecological factors should also be considered to better clarify the evolution of flower traits of Betulaceae plants.

Anatomical and leaf characters were important traits for taxonomic and systematic consideration within Betulaceae. And, it becomes clear that some key characters well reflect the evolutionary order among the six genera (Figs. S6 and S7). For example, *Carpinus* and *Ostrya*, the two youngest sister genera in the family, coevolved simple vessel perforations and tyloses in anatomical structure and degraded glands on leaves, of which *Carpinus* species further obtained typical stomatal apparatus in their leaves. By comparison, two primitive genera *Alnus* and *Betula* retained most of the ancestral states of Betulaceae, including scalariform perforations, distinct tracheid, absent tyloses and stomatal apparatus, and embedded glands on leaves. *Corylus* and *Ostryopsis* seem in some aspects to be transitional between Coryloideae and Betuloideae. On one hand, the two share some characters with *Alnus* or *Betula*, such as scalariform perforations, absent tyloses and stomatal apparatus. On the other hand, they display characters in common with *Ostrya* or *Carpinus*, including degraded tracheid, and present pistilloide in male floret. Notably, we reveal that some characters are parallelly evolved even between sister genera, including the cases of pistilloide in male floret, thecae and filaments, stomatal apparatus, and leaf teeth between *Carpinus* and *Ostrya*, *Alnus* and *Betula*. This indicates that divergent evolution of multiple characters has involved in the divergence among sister lineages.

Evolution of fruit types and their dispersal modes is recognized as important drivers of angiosperms diversification [74, 75]. Despite diverse fruit and diaspore types in the plant kingdom we observe today, it has taken a long time to finish this evolutionary process. In the Cretaceous, angiosperms were still dominated by small fruits, and abiotic dispersal was the mainstream. During the Eocene, average fruit size increased sharply, and biotic dispersal by vertebrates has become much more prevalent. In the late Tertiary, average fruit size tended to decrease as the climate cooled and vegetation opened [76]. In the present study, divergence time estimation and ancestral character reconstruction reveal consistent trends between fruit character evolution and historical geo-climatic changes (Figs. 4 and 5). First, Betulaceae ancestors were inferred to have small winglike fruits and winged diaspores, corresponding to the ancestral states

of most angiosperms in the Cretaceous. Thereafter, the two primitive genera *Alnus* and *Betula* retained these ancestral states in subsequent evolution. Second, *Corylus* exclusively evolved prominent large nuts in the Eocene, conforming with the conspicuous increase in seed and fruit sizes in the early Tertiary. Especially, a variety of bract shapes have occurred in *Corylus*, with most species owning campanulate or tubular bracts and a few species having spiny bract. Finally, *Ostryopsis* and *Ostrya* parallelly retrieved small nuts enclosed by saclike bracts, while *Carpinus* further obtained small leafy nuts independently, coinciding with the reversal of fruit size in the late Tertiary. Hence, adaptive differences among genera at different historical stages may be the primary driving force for the evolution of fruit size and dispersal mode.

Conclusion

Our research documented a comprehensive plastome feature for Betulaceae at species level, and represented the most robust phylogenetic inference to date. Comparative genomics showed that Betulaceae plastomes were highly conserved in genome size, gene order, and structure, although specific variations such as gene loss and IR boundary shifts were revealed. Six coding regions ($P_i > 0.02$) and five non-coding regions ($P_i > 0.035$) were identified as candidate DNA barcodes for phylogenetic analyses and taxonomic research. Furthermore, our phylogenomic studies clarified some unsolved phylogenetic issues, e.g., reciprocal monophyly between *Betula* and *Alnus*, successive sister relationships among *Corylus*, *Ostryopsis*, and *Carpinus-Ostrya*, and mutual paraphyly between *Ostrya* and *Carpinus*. Based on robust phylogenetic framework, we inferred that Betulaceae ancestors originated from East Asia during the upper Cretaceous; dispersals and subsequent vicariance accompanied by historical environment changes contributed to its diversification and intercontinental disjunction. Ancestral state reconstruction indicated the acquisition of many taxonomic characters was actually the results of parallel or reversal evolution. Overall, the results provide new insights into the plastome structural variation, phylogenetic relationships among major lineages, and simultaneously help to elucidate the diversification history and evolution of Betulaceae.

Methods

Taxon sampling and DNA isolation

A total of 109 plastomes representing Betulaceae and allied families were included in this study. Ninety-nine plastomes from the six genera of Betulaceae were selected, including *Alnus* (22, ca. 30 spp.), *Betula* (18, ca. 50 spp.), *Corylus* (23, ca. 20 spp.), *Ostryopsis* (5, 3 spp.), *Carpinus* (19, ca. 30 spp.), and *Ostrya* (12, ca. 8 spp.). Ten plastomes of allied families in Fagales were chosen as outgroup in the phylogenetic analysis. Among these 109 plastomes, sixteen were sequenced and assembled by our laboratory and the others were obtained from GenBank (Table S1). Leaves used in this study were either collected from natural populations in China or friendly provided as herbarium specimens by cooperative institutions. The formal identification of these plant materials was undertaken by Prof. Guixi Wang (Chinese Academy of Forestry). Voucher specimens were deposited in the non-wood forest laboratory of Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China. Sampling information and herbarium vouchers were offered in Table 1. Genomic DNA of each sample was extracted from leaves (silica gel-dried or fresh materials) using DNeasy Plant Mini Kit (Qiagen, Beijing, China). The DNA quality and purity were evaluated using the Qubit Fluorometric Quantitation (Thermo Fisher, Scientific, USA) instrument and agarose gel electrophoresis.

Plastome assembly and annotation

Library construction and paired-end sequencing were performed with Illumina HiSeq 2500-PE125 platform at Novogene (Beijing, China). The raw reads were checked using NGS QC Toolkit [77] with the following criteria: removing PCR duplicates and adapters; filtering reads with N over 10%; filtering reads with a mass value more than 40% and less than 10%. Based on these obtained high-quality reads, plastomes were assembled using MITObim v1.7 [78] with multiple Betulaceae plastomes as reference sequences. Small gaps existed in the assembled plastomes were filled and corrected by PCR-based sanger sequencing. PCR procedures and primers were provided in Table S5. These newly assembled plastomes were annotated employing both the program GeSeq [79] and DOGMA [80]. Draft annotations were further adjusted and verified through a BLAST alignment against the published Betulaceae plastomes. Besides, we renovated the annotations of previously published plastomes before using them in our study. All 16 newly obtained plastomes were submitted to GenBank with accession numbers provided in Table 1.

Comparative genomics and structural analyses

To compare the structure variation and identify arrangement events across the family, comparative genomics of major lineages (31 representing plastomes) were performed under the Shuffle-LAGAN strategy in mVISTA [81] with the annotation of *Alnus alnobetula* as reference. The junction sites of four structural regions (IRA, LSC, SSC, IRB) and adjacent genes in these plastomes were ascertained using the online program IRSCOPE [82] to investigate the expansion or contraction of IRs. To

explore variability among different regions (protein-coding regions and non-coding regions) for species identification and population genetics, nucleotide diversity (P_i) of these regions was estimated with DnaSP 5.0 [83].

Phylogenetic inference

Due to the heterogeneous evolutionary rate of different regions in plastome, we generated three datasets for phylogenetic analyses: protein-coding regions (CDS), non-coding regions (CNS), and whole plastomes (WP). The CDS and CNS sequences of all 109 plastomes were extracted with PhyloSuite v1.2.2 [84] and aligned separately using MAFFT v7.4 [85]. Then, these individual alignments were concatenated into supermatrices by PhyloSuite v1.2.2. We excluded ambiguously aligned sites from the three datasets, especially the WP dataset and the CNS dataset, using trimAl v1.2 with the automated1 model [86]. In addition, in order to evaluate the power of potential barcodes in phylogenetic inference, we also generated the fourth dataset (DH) by concatenating ten newly developed divergence hotspots obtained through comparative genomics analysis (Fig. 1; Table S3). Independent phylogenetic analyses were performed for each dataset (CDS, CNS, WP, and DH) using both Maximum likelihood (ML) and Bayesian inference (BI) strategies. The optimal nucleotide substitution models were calculated with the built-in ModelFinder program of PhyloSuite v1.2.2 under the Bayesian information criterion. The ML analysis was conducted by IQ-tree 1.63 [87] with support values evaluated by approximate likelihood-ratio test of 1,000 replicates and ultrafast bootstrapping of 5,000 replicates. The BI inference was implemented in MrBayes v3.26 [88]. Two independent chains (2×10^7 generations for each chain) were conducted with starting from random trees. The sampling frequency was set as 1,000 generations and the first 25% of trees were discarded as burn-in. Stationarity was assumed when the average standard deviation of split frequencies < 0.01. The posterior probability (*PP*) values were estimated based on the remaining trees. FigTree v1.4.2 [89] was utilized to visualize the phylogeny.

Divergence time estimation

Molecular dating analysis was performed with the Bayesian molecular-clock method in BEAST 1.84 [90] based on CDS data set. Three fossils were used to constrain the internal nodes: (1) According to the ancient flower fossils that represent an extinct lineage at the basal Betulaceae [91], 71.0 Ma was assigned as a mean age for the crown group of the family, and sigma of 3.0 was set for the normal prior distribution. (2) Based on the earliest pollen fossils for *Alnus* [92], we set a minimum age as 58.0 Ma to calibrate the split between *Betula* and *Alnus*. The prior was set as a log-normal distribution, with the offset 58.0, mean 1.0, and sigma 0.5. (3) From reports on the oldest fossils for *Palaeocarpinus*, *Carpinus*, and *Corylus* [10, 12], the most recent common ancestor of subfamily Coryloideae was constrained with an age between 37.0 Ma and 49.0 Ma. The prior was treated as normal distribution and sigma 3.0. BEAST was implemented using the uncorrelated log-normal relaxed clock and the GTR + G substitution model that was calculated by ModelFinder (Table S4). Yule process was selected as tree prior. Two independent MCMC simulations were conducted with each running 4.0×10^7 generations and sampling every 1,000 generations. Convergence and stationarity of the results were checked by Tracer v1.7 [93]. Nodal heights and 95% highest posterior density intervals were summarized using TreeAnnotator v2.12 [94], with the first 25% trees treated as burn-in.

Biogeographical inference

Ancestral area reconstruction was performed using the BioGeoBEARS plugin in RASP v4.3 [95]. A pruned time-calibrated tree including 74 taxa (one accession per species) inferred from BEAST analysis was used as the input tree. Due to lack of information on the ancestral distribution of outgroups, we also removed the outgroups from the tree in the biogeographic analysis. The likelihood implementation of Bayesian inference for discrete areas (BAYAREALIKE) was selected as the best biogeographical model by BioGeoBEARS according to the corrected Akaike information criterion (Table S6). All terminal taxa were assigned to nine geographic areas based on the distribution of species diversity and endemicity: (A) southwestern East Asia, (B) central and eastern China, (C) Northeast Asia, (D) Central Asia, (E) southern Europe and the Mediterranean coast, (F) northern Europe, (G) western North America, (H) eastern North America, (I) Central America. The maximum range size was set as three in that it is the maximum number of geographic areas where an extant genus occurs.

Ancestral character state reconstruction

We explored character evolution for 14 morphological characters frequently used in the systematic and taxonomical studies of Betulaceae [2, 16, 44, 96], including characters of inflorescence, flowers, wood anatomy, leaves, and fruits. Information of these characters and their state scores for each species are provided in Table S7. Ancestral states were reconstructed using the one-parameter Markov k-state evolutionary model in Mesquite 3.51 [97]. For this analysis, a compiled tree (generated by the BEAST) containing 74 Betulaceae species was employed as the input tree. The corresponding characters were mapped onto the BI tree and the levels of homoplasy for each character were evaluated qualitatively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-022-03991-1.

Additional file 1: Figure S1. Sequence identity plot of 31 Betulaceae plastomes using mVISTA. Gray arrows and thick black lines above the alignment indicate genes with their orientation and the position of the IRs, respectively. Different regions (LSC, IR, and SSC) are represented by different colors. **Figure S2.** Phylogeny of Betulaceae inferred from Maximum likelihood (ML) and Bayesian inference (BI) based on 109 whole plastomes (WP). BS and PP values are presented on the branches. Asterisks represent 100/1.0 support values. Major genera of Betulaceae are indicated by different colors. **Figure S3.** Phylogeny of Betulaceae inferred from Maximum likelihood (ML) and Bayesian inference (BI) based on non-coding sequences (CNS). BS and PP values are presented on the branches. Asterisks represent 100/1.0 support values. Major genera of Betulaceae are indicated by different colors. **Figure S4.** Phylogeny of Betulaceae inferred from Maximum likelihood (ML) and Bayesian inference (BI) based on divergence hotspots (DH). BS and PP values are presented on the branches. Asterisks represent 100/1.0 support values. Major genera of Betulaceae are indicated by different colors.

Additional file 2: Table S1. Accession numbers and structural features of all 93 plastomes obtained from GenBank. **Table S2.** Summary of protein-coding genes in all Betulaceae plastomes. **Table S3.** Statistics of nucleotide diversity for 78 coding genes and 68 non-coding regions. **Table S4.** Data characteristics and best-fit models for ML and BI phylogenetic analyses. **Table S5.** Primers used for gap closure in this study. **Table S6.** Results of model test used for biogeographic inference.

Acknowledgements
We thank Drs Qing Li and Sihao Hou for providing materials in this research, and to Wenpan Dong and Chao Xu for their help with bioinformatic work.

Authors' contributions
MA Q., conceived the study and supervised the project; Yang Z., Ma W., Yang X., and Wang L., performed the experiments and analyzed the data; Zhao T., Liang L., and Wang G., collected the materials; Yang Z., wrote the draft manuscript. All authors have contributed to revisions and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (32101541), the National Key Research and Development Program of China (2022YFD2200400), and the Key Research and Development Program of Hebei Province (21326804D).

Availability of data and materials
All sequences described in this study are available in the GenBank repository under the accessions as summarized in Table 1 and Additional file 2: Table S1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]State Key Laboratory of Tree Genetics and Breeding, Beijing 100091, China. [2]Key Laboratory of Tree Breeding and Cultivation, National Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China. [3]National Innovation Alliance of Hazelnut Industry, Beijing 100091, China. [4]Hazelnut Engineering and Technical Research Center of the State Forestry and Grassland Administration, Beijing 100091, China. [5]Forest Botany and Tree Physiology, University of Goettingen, 37077 Goettingen, Germany. [6]Guizhou Academy of Forestry, Guiyang 550005, China. [7]Anhui Academy of Forestry, Hefei 230031, China.

Received: 23 May 2022 Accepted: 9 December 2022
Published online: 24 December 2022

References

1. Christenhusz MJM, Byng JW. The number of known plant species in the world and its annual increase. Phytotaxa. 2016;261(3):201–17.
2. Chen ZD, Manchester SR, Sun HY. Phylogeny and evolution of the Betulaceae as inferred from DNA sequences, morphology and paleobotany. Am J Bot. 1999;86(8):1168–81.
3. Kubitzki K. Flowering plants, dicotyledons: Celastrales, oxalidales, rosales, cornales, ericales. Berlin: Springer; 2004.
4. Grimm GW, Renner SS. Harvesting Betulaceae sequences from GenBank to generate a new chronogram for the family. Bot J Linn Soc. 2013;172(4):465–77.
5. Yang Z, Wang GX, Ma QH, Ma WX, Liang LS, Zhao TT. The complete chloroplast genomes of three Betulaceae species: implications for molecular phylogeny and historical biogeography. PeerJ. 2019;7:e6320.
6. Bina H, Yousefzadeh H, Ali SS, Esmailpour M. Phylogenetic relationships, molecular taxonomy, biogeography of *Betula*, with emphasis on phylogenetic position of Iranian populations. Tree Genet Genomes. 2016;12(5):1–17.
7. Chen ZD, Li JH. Phylogenetics and biogeography of *Alnus* (Betulaceae) inferred from sequences of nuclear ribosomal DNA ITS region. Int J Plant Sci. 2004;165(2):325–35.
8. Bousquet J, Strauss SH, Li P. Complete congruence between morphological and *rbc*L-based molecular phylogenies in birches and related species (Betulaceae). Mol Biol Evol. 1992;9(6):1076–88.
9. Furlow JJ. The genera of Betulaceae in the southeastern United States. J Arnold Arbor. 1990;71(1):1–67.
10. Pigg KB, Manchester SR, Wehr WC. *Corylus*, *Carpinus*, and *Palaeocarpinus* (Betulaceae) from the middle Eocene Klondike Mountain and Allenby formations of northwestern North America. Int J Plant Sci. 2003;164(5):807–82.
11. Crane PR. Early fossil history and evolution of the Betulaceae. In: Crane PR, Blackmore S (eds) Evolution, systematics, and fossil history of the Hamamelidae. Vol 2. "Higher Hamamelidae." Systematic Association, Clarendon Press, Oxford. 1989;40:87–116.
12. Forest F, Savolainen V, Chase MW, Lupia R, Bruneau A, Crane PR. Teasing apart molecular-versus fossil-based error estimates when dating phylogenetic trees: A case study in the birch family (Betulaceae). Syst Bot. 2005;30(1):118–33.
13. Abbe EC. Flowers and Inflorescences of the "Amentiferae." Bot Rev. 1974;40(2):159–261.
14. Li RQ, Chen ZD, Lu AM, Soltis DE, Soltis PS, Manos PS. Phylogenetic relationships in Fagales based on DNA sequences from three genomes. Int J Plant Sci. 2004;165(2):311–24.
15. Kato H, Oginuma K, Gu Z, Hammel B, Tobe H. Phylogenetic relationships of Betulaceae based on *mat*K sequences with particular reference to the position of *Ostryopsis*. Acta Phytotax Geobot. 1999;49(2):89–97.
16. Yoo KO, Wen J. Phylogeny and biogeography of Carpinus and subfamily Coryloideae (Betulaceae). Int J Plant Sci. 2002;163(4):641–50.
17. Xie DF, Tan JB, Yu Y, Gui LJ, Su DM, Zhou SD, et al. Insights into phylogeny, age and evolution of *Allium* (Amaryllidaceae) based on the whole plastome sequences. Ann Bot. 2020;125(7):1039–55.
18. Gitzendanner MA, Soltis PS, Wong GKS, Ruhfel BR, Soltis DE. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. Am J Bot. 2018;105(3):291–301.
19. Amenu SG, Wei N, Wu L, Wu L, Oyebanji O, Hu G, et al. Phylogenomic and comparative analyses of Coffeeae alliance (Rubiaceae): deep insights into phylogenetic relationships and plastome evolution. Bmc Plant Biol. 2022;22(1):1–13.
20. Downie SR, Jansen RK. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat,

mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. Syst Bot. 2015;40(1):336–51.
21. Wang HX, Liu H, Moore MJ, Landrein S, Liu B, Zhu ZX, et al. Plastid phylogenomic insights into the evolution of the Caprifoliaceae sl (Dipsacales). Mol Phylogenet Evol. 2020;142:106641.
22. Guo MY, Pang XH, Xu YQ, Jiang WJ, Liao BS, Yu JS, et al. Plastid genome data provide new insights into the phylogeny and evolution of the genus *Epimedium*. J Adv Res. 2022;36:175–85.
23. Meucci S, Schulte L, Zimmermann HH, Stoof-Leichsenring KR, Epp L, Bronken-Eidesen P, et al. Holocene chloroplast genetic variation of shrubs (Alnus alnobetula, Betula nana, Salix sp.) at the siberian tundra-taiga ecotone inferred from modern chloroplast genome assembly and sedimentary ancient DNA analyses. Ecol Evol. 2021;11(5):2173–93.
24. Hu GL, Cheng LL, Huang WG, Cao QC, Zhou L, Jia WS, et al. Chloroplast genomes of seven Coryloideae species: structures and comparative analysis. Genome. 2020;63(7):337–48.
25. Lee SI, Nkongolo K, Park D, Choi IY, Choi AY, Kim NS. Characterization of chloroplast genomes of *Alnus rubra* and *Betula cordifolia*, and their use in phylogenetic analyses in Betulaceae. Genes Genom. 2019;41(3):305–16.
26. Li Y, Yang YZ, Yu L, Du X, Ren GP. Plastomes of nine hornbeams and phylogenetic implications. Ecol Evol. 2018;8(17):8770–8.
27. Zhang TW, Fang YJ, Wang XM, Deng X, Zhang XW, Hu SN, et al. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLoS ONE. 2012;7(1):e30531.
28. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17(1):1–29.
29. Weng ML, Ruhlman TA, Jansen RK. Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. New Phytol. 2017;214(2):842–51.
30. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytol. 2010;186(2):299–317.
31. Alwadani KG, Janes JK, Andrew RL. Chloroplast genome analysis of box-ironbark Eucalyptus. Mol Phylogenet Evol. 2019;136:76–86.
32. Raman G, Park SJ. The complete chloroplast genome sequence of *Ampelopsis*: gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. Front Plant Sci. 2016;7:341.
33. Sun Y, Moore MJ, Zhang S, Soltis PS, Soltis DE, Zhao T, et al. Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. Mol Phylogenet Evol. 2016;96:93–101.
34. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, et al. Many parallel losses of *inf*A from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell. 2001;13(3):645–58.
35. Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK. Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. J Mol Evol. 2010;70(2):149–66.
36. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol. 2016;209:1747–56.
37. Downie SR, Jansen RK. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. Syst Bot. 2015;40:336–51.
38 Ahmed I, Matthews PJ, Biggs PJ, Naeem M, Mclenachan PA, Lockhart PJ. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of Colocasia esculenta (L.) Schott (Araceae) and closely related taxa. Mol Ecol Resour. 2013;13:929–37.
39. Yang J, Yue M, Niu C, Ma XF, Li ZH. Comparative analysis of the complete chloroplast genome of four endangered herbals of *Notopterygium*. Genes. 2017;8:124.
40. Choi KS, Chung MG, Park S. The complete chloroplast genome sequences of three Veroniceae species (Plantaginaceae): comparative analysis and highly divergent regions. Front plant sci. 2016;7:355.
41. Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. Am J Bot. 2014;101:1987–2004.
42. Yoo KO, Wen J. Phylogeny of *Carpinus* and subfamily Coryloideae (Betulaceae) based on chloroplast and nuclear ribosomal sequence data. Plant Syst Evol. 2007;267(1):25–35.
43. Xing Y, Onstein RE, Carter RJ, Stadler T, Peter LH. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected. Evolution. 2014;68(10):2821–32.
44. Larson-Johnson K. Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. New Phytol. 2016;209(1):418–35.
45. Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ. 2019;7:e7747.
46. Goncalves DJP, Simpson BB, Ortiz EM, Shimizu GH, Jansen RK. Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. Mol Phylogenet Evol. 2019;138:219–32.
47. Buckley TR, Cordeiro M, Marshall DC, Simon C. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada Dugdale*). Syst Biol. 2006;55:411–25.
48. Wendel JF, Doyle JJ. Phylogenetic incongruence: window into genome history and molecular evolution. Boston: Molecular systematics of plants II. Springer; 1998. p. 265–96.
49. McCauley DE, Stevens JE, Peroni PA, Raveill JA. The spatial distribution of chloroplast DNA and allozyme polymorphisms within a population of *Silene alba* (Caryophyllaceae). Am J Bot. 1996;83:727–31.
50. Zhao T, Wang G, Ma Q, Liang L, Yang Z. Multilocus data reveal deep phylogenetic relationships and intercontinental biogeography of the Eurasian-North American genus *Corylus* (Betulaceae). Mol Phylogenet Evol. 2020;142:106658.
51. Tsuda Y, Semerikov V, Sebastiani F, Vendramin GG, Lascoux M. Multispecies genetic structure and hybridization in the *Betula* genus across Eurasia. Mol Ecol. 2017;26:589–605.
52. Liu B, Abbott RJ, Lu Z, Tian B, Liu J. Diploid hybrid origin of *Ostryopsis intermedia* (Betulaceae) in the Qinghai-Tibet Plateau triggered by Quaternary climate change. Mol Ecol. 2014;23:3013–27.
53. Wang Z, Kang M, Li J, Zhang Z, Wang Y, Chen C, Liu J. Genomic evidence for homoploid hybrid speciation between ancestors of two different genera. Nat Commun. 2022;13:1–9.
54. Liu LX, Du YX, Folk RA, Wang SY, Soltis DE, Shang FD, Li P. Plastome evolution in Saxifragaceae and multiple plastid capture events involving *Heuchera* and *Tiarella*. Front Plant Sci. 2020;11:361.
55. Hauenschild F, Favre A, Michalak I, Muellner-Riehl AN. The influence of the Gondwanan breakup on the biogeographic history of the *ziziphoids* (Rhamnaceae). J Biogeogr. 2018;45(12):2669–77.
56. Muellner-Riehl AN, Weeks A, Clayton JW. Molecular phylogenetics and molecular clock dating of Sapindales based on plastid *rbc*L, *atp*B and *trn*L-*trn*F DNA sequences. Taxon. 2016;65(5):1019–36.
57. Mao KS, Liu JQ. Current 'relicts' more dynamic in history than previously thought. New Phytol. 2012;196(2):329–31.
58. Yang XY, Wang ZF, Luo WC, Guo XY, Zhang CH, Liu JQ, et al. Plastomes of Betulaceae and phylogenetic implications. J Syst Evol. 2019;57(5):508–18.
59. Takahashi K. Palynology of the Upper Aptian Tanohata Formation of the Miyako Group, northeast Japan. Pollen Spores. 1974;16(4):535–64.
60. Xiang XG, Wang W, Li RQ, Lin L, Liu Y, Zhou ZK, et al. Large-scale phylogenetic analyses reveal fagalean diversification promoted by the interplay of diaspores and environments in the Paleogene. Perspect Plant Ecol. 2014;16(3):101–10.
61. Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, et al. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 2011;98(4):704–30.
62. Schulte P, Alegret L, Arenillas I, Arz JA, Barton PJ, Bown PR, et al. The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. Science. 2010;327(5970):1214–8.
63. Zhai W, Duan XS, Zhang R, Guo CC, Li B, Xu GX, et al. Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. Mol Phylogenet Evol. 2019;135:12–21.
64. Sun H. Tethys retreat and Himalayas-Hengduanshan mountains uplift and their significance on the origin and development of the Sino-Himalayan elements and alpine flora. Acta Bot Yun. 2002;24:273–88.

65. Utescher T, Mosbrugger V. Eocene vegetation patterns reconstructed from plant diversity-a global perspective. Palaeogeogr Palaeoecol. 2007;247(3–4):243–71.
66. Manchester SR. Biogeographical relationships of North American tertiary floras. Ann Mo Bot Gard. 1999;86:472–522.
67. Graham A. The role of land bridges, ancient environments, and migrations in the assembly of the North American flora. J Syst Evol. 2018;56(5):405–29.
68. Sun XJ, Wang PX. How old is the Asian monsoon system?-Palaeobotanical records from China. Palaeogeogr Palaeoecol. 2005;222(3–4):181–222.
69. Shi YF, Tang MC, Ma YZ. Linkage between the second uplifting of the Qinghai-Xizang (Tibetan) Plateau and the initiation of the Asian monsoon system. Sci China Ser D. 1999;42(3):303–12.
70. Silvestro D, Zizka G, Schulte K. Disentangling the effects of key innovations on the diversification of Bromelioideae (Bromeliaceae). Evolution. 2014;68(1):163–75.
71. Antonelli A, Sanmartín I. Why are there so many plant species in the Neotropics? Taxon. 2011;60(2):403–14.
72. Blanco-Pastor JL, Ornosa C, Romero D, Liberal IM, Gómez JM, Vargas P. Bees explain floral variation in a recent radiation of *Linaria*. J Evolution Biol. 2015;28(4):851–63.
73. Boberg E, Alexandersson R, Jonsson M, Maad J, Agren J, Nilsson LA. Pollinator shifts and the evolution of spur length in the moth-pollinated orchid *Platanthera bifolia*. Ann Bot. 2014;113(2):267–75.
74. Marcussen T, Meseguer AS. Species-level phylogeny, fruit evolution and diversification history of *Geranium* (Geraniaceae). Mol Phylogenet Evol. 2017;110:134–49.
75. Beaulieu JM, Donoghue MJ. Fruit evolution and diversification in Campanulid angiosperms. Evolution. 2013;67(11):3132–44.
76. Eriksson O. Evolution of seed size and biotic seed dispersal in angiosperms: paleoecological and neoecological evidence. Int J Plant Sci. 2008;169(7):863–70.
77. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE. 2012;7:e30619.
78. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. Nucleic Acids Res. 2013;41(13):e129–e129.
79. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq-versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–11.
80. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20(17):3252–5.
81 Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32(suppl_2):W273–9.
82. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–1.
83. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.
84. Zhang D, Gao FL, Jakovlić I, Zou H, Zhang J, Li WX, et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20(1):348–55.
85. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
86. Capellagutiérrez S, Sillamartínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
87. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
88. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19(12):1572–4.
89. Rambaut A. FigTree v1. 4. University of Edinburgh, Edinburgh, UK. 2012. Available at: http://tree.bio.ed.ac.uk/software/figtree.
90 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73.
91. Friis EM. *Endressianthus*, a new normapolles-producing plant genus of Fagalean affinity from the late Cretaceous of Portugal. Int J Plant Sci. 2003;164(S5):S201–23.
92. Konzalova M. Paraalnipollenites Hills and Wallace 1969, in the Turonian of the upper Cretaceous of North Bohemia. Vestnik Ustredniho Geologickeho. 1971;46:39–40.
93 Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67(5):901–4.
94. Rambaut A, Drummond AJ. TreeAnnotator v2.1.2. Edinburgh: University of Edinburgh, Institute of Evolutionary Biology; 2014.
95. Yu Y, Blair C, He XJ. RASP 4: Ancestral State Reconstruction Tool for Multiple Genes and Characters. Mol Biol Evol. 2020;37(2):604–6.
96. Zhu JY, Zhang LF, Ren BQ, Chen M, Li RQ, Zhou Y, et al. Comparative flower and inflorescence organogenesis among genera of Betulaceae: implications for phylogenetic relationships. Bot Rev. 2018;84(1):79–98.
97. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. Version 3.51. 2018. Available at: http://mesquiteproject.org.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC