Title: Harvesting GenBank for Betulaceae sequences to generate a new chronogram for the family

Authors: Guido W. Grimm¹ and Susanne S. Renner²
¹Department of Palaeobotany, Swedish Museum of Natural History, P.O. Box 50007, 10405 Stockholm
²Systematic Botany and Mycology, University of Munich (LMU), Menzinger Strasse 67, 80638 Munich, Germany

Corresponding author’s e-mail: guido.grimm@nrm.se

Running head: A new chronogram for the Betulaceae

Abstract
Betulaceae, with 120-150 species in six genera, are a family of Fagales that occurs mainly in the Northern Hemisphere. Previous studies of the evolution of Alnus, Betula, Carpinus, Corylus, Ostrya, and Ostryopsis have relied on relatively small amounts of sequence data and molecular clocks with fixed-point calibrations. We exploited GenBank to construct Betulaceae matrices of up to 900 sequence accessions and 9300 nucleotides of nuclear and plastid DNA; we also computed species consensus sequences to build 46- and 29-species matrices that strike a balance between species sampling and nucleotide sampling. Trees were rooted on Ticodendraceae and Casuarinaceae, and divergence times were inferred under relaxed and strict molecular clocks, using alternative fossil constraints. The data support the traditional two subfamilies, Betuloideae (Alnus, Betula) and Coryloideae, and show that Ostryopsis is sister to Ostrya/Carpinus. The fossil record and molecular clocks calibrated with alternating fossils indicate that the stem lineage of Betulaceae dates back to the Upper Cretaceous, the two subfamilies to the Paleocene, and the most recent common ancestors of each of the living genera to the middle to late Miocene. A substitution rate shift within the Coryloideae between 25 and 15 million years ago precedes the mid-Miocene climatic optimum and may be linked to temperate niches that became available following the mid-Miocene.

Keywords: Fossil calibrations, consensus sequences, GenBank data curation, molecular clocks, substitution rate shift
INTRODUCTION

The Fagales family Betulaceae comprises some 120–150 species of trees or shrubs occurring mostly in the northern temperate zone. A few species occur as far south as the Andes and the island of Sumatra (Kubitzki, 1993). Betulaceae can be recognized by their stipulate, doubly serrate leaves, catkins, and small winged fruits or nuts associated with leafy bracts. There are six genera in two subfamilies of overlapping geographic distributions: *Alnus* Mill. (29–35) and *Betula* (42–50) make up the Betuloideae, and *Carpinus* L. (26–35), *Corylus* L. (16), *Ostrya* Scop. (5–9), and *Ostryopsis* Decne. (3) make up the Coryloideae. Several species of *Alnus*, especially *A. rubra* Bong., are used in reforestation because of their fast growth and nitrogen fixation; European hazelnut, *Corylus avellana* L., is an economically important food crop.

In spite of its relatively small size and several molecular-phylogenetic studies, relationships among the subfamilies have remained unclear (Table 1). The first plastid *rbcL* phylogeny for the Betulaceae was published 20 years ago (Bousquet, Strauss & Li, 1992), but did not include *Ostryopsis*. Subsequent studies added plastid *matK* sequences (Kato et al., 1998: one species of *Ostryopsis* sampled), nuclear internal transcribed spacer (ITS) sequences (Chen, Manchester & Sun, 1999, Yoo & Wen, 2002, Forrest et al., 2005, Li, Shoup & Chen, 2005), further plastid genes and spacers (Yoo & Wen, 2007: combined 2906 nucleotides), and nuclear nitrate reductase sequences (*NIA*, 3rd intron; Li, Shoup & Chen, 2007: 11 diploid species of *Betula*; Li, 2008: 22 species of Coryloideae). Those of these studies that included outgroups mostly placed the root between the two subfamilies of Betulaceae (Table 1). A study focusing on the entire order Fagales, however, found *Alnus* and *Betula* as successive sisters to the remaining genera (Li et al., 2004: nuclear 18S rDNA, mitochondrial *matR*, and four plastid markers; Table 1). Species of *Ostrya* and *Carpinus* were variously intermixed in nuclear and plastid trees (Yoo & Wen, 2002, Yoo & Wen, 2007, Li, 2008). No study so far has sampled both species of *Ostryopsis* (a third species, *O. intermedia* B.Tian & J.Q.Liu, was described recently Tian et al., 2010).

Molecular data show that the closest relatives of the Betulaceae are the Ticodendraceae, a monotypic family ranging from Panama to Mexico, and the Casuarinaceae, with 69 species in Australia and 27 in Malesia (Li et al., 2004). All three families have good fossil records (Crane, 1989, Scriven & Hill, 1995, Chen et al., 1999, Manchester, Pigg & Crane, 2004, Manchester, 2011). The first *Alnus*-type pollen is from the Upper Cretaceous (Coniacian, 86.3–89.8 Myr; Konzalova, 1971, Forrest et al., 2005; absolute ages from the chronostratigraphic chart of Cohen, Finney & Gibbard, 2012) and the oldest macrofossils assignable to the family are nutlets of *Palaeocarpinus* from the late Paleocene (56-59.2) of southern England and North Dakota (Crane, 1981, Crane, Manchester & Dilcher, 1990). These nutlets resemble both *Ostrya* and *Corylus*, and do not represent any extant genus (Sun & Stockey, 1992). An extinct lineage close to the root of the
The relatively good fossil record of Betulaceae has encouraged molecular-clock studies comparing the clade’s plastid substitution rates to those in other angiosperms (Bousquet et al., 1992) and studies that inferred divergence times among the genera (Forest et al., 2005). Assuming a split of *Alnus* and *Betula* at either 80 Myr or 45 Myr, Bousquet *et al.* inferred *rbcL* rates of 0.37 or $0.67 \times 10^{-4}$ substitution per site per Myr (and similar rates when they used coryloid fossils as calibrations). Forest *et al.* applied non-parametric rate smoothing (Sanderson, 1997) and five Betulaceae fossils as fixed calibrations, using one at a time, either placed at crown group nodes or at stem lineage nodes. Many of the inferred ages exceeded oldest fossil occurrences of the relevant clades. For instance, the Betulaceae crown group was estimated at 25 million years older than the first Betulaceae-type pollen from the Coniacian (86.3–89.8 Myr). However, these estimates were based on just 462 nucleotides of ITS, and the age for each node was the median of the ages obtained from 10 alternative fossil constraints, including five where ancient fossils were assigned to the crown groups of the living genera *Alnus*, *Betula*, *Carpinus*, *Corylus*, and *Ostrya*. The combination of a small amount of sequence data with so-called “consensus estimates” from 10 vastly divergent crown and stem constraints throws doubt on the inferred ages.

Taking advantage of the Betulaceae DNA sequences now available in GenBank, we compiled a matrix of up to 900 sequence accessions and 9300 nucleotides of nuclear and plastid DNA. Species consensus sequences were then computed to obtain less patchy matrices that strike a balance between species sampling and nucleotide sampling. We also added new sequences for *Ostryopsis*. By representing the genetic variation found in each genus (judged from hundreds of sequences), we hoped to better resolve genus and subfamily relationships in the Betulaceae. To infer divergence times, we used relaxed and strict clock methods with prior age probability distributions on stem lineage fossils, rather than fixed calibrations on either stem or crown groups. We also wanted to test whether plastid substitution rates inferred 20 years ago (Bousquet *et al.*, 1992) would hold up with more sophisticated clock approaches and larger taxon and gene sampling and if there were any drastic shifts in substitution rates (i.e., branch length differences).

**MATERIALS AND METHODS**

**SAMPLING OF TAXA AND GENETIC MARKERS, DNA ALIGNMENT**

Species names, GenBank accession numbers, and vouchers (where available) are listed in the nexus-formatted single-partition matrices included in the online supporting archive (OSA) hosted at [www.palaeogrimm.org/data](http://www.palaeogrimm.org/data). The sources of our new sequences (accession numbers KC412166–
KC412181) are *Ostryopsis davidiana* Decne. are Huhehaote, Inner Mongolia, 40°54'/111°34', alt. 1280 m, Liu Jianquan 152-3 (LZU) and *Ostryopsis nobilis* Balf.f. & W.W.Sm., Daju, Yunnan, 27°16'/100°13', alt. 1910 m, Liu Jianquan 53-7 (LZU).

GenBank data were harvested and processed using GBK2FAS (Göker *et al.*, 2009), and alignments were done with MAFFT v. 5 (standard settings; Katoh *et al.*, 2005), followed by visual checked for inconsistencies or erroneous sequences (details are described in the supplementary information [SI] 1). The full alignments comprised 230 sequences of *atpB-rbcL* and *rbcL*, 49 of granule-bound starch synthase (GBSSI), 502 of ITS (with many Betulaceae represented by multiple sequences), 252 of *trnH-psbA*, 75 of *rpl16*, and 395 of *trnK/matK*, 146 of the *trnL* region. There were no statistically supported (>80% ML bootstrap support) conflicting topological placements in trees generated from the individual markers. We also aligned and studied sequences of the nuclear-encoded 18S rDNA, nitrate reductase (*NIA*), mitochondrial *matR*, and the plastid *atpB* gene, but did no use them in the final analyses because they contained no genus-level signal or in the case of *NIA* were too variable to be aligned between the genera. Instead of choosing a single placeholder accession for each species, species consensus sequences were generated with G2CEF (Göker & Grimm, 2008), using the option “strict” and gaps treated as missing data. We then built a 46-taxon-9321-nucleotide matrix that includes five species of *Alnus*, nine of *Betula*, nine of *Corylus*, five of *Ostrya*, two of *Ostryopsis*, and 14 of *Carpinus* for a total of 44 Betulaceae, which strikes a balance between species sampling and nucleotide sampling: The matrix contained 32.37% empty cells (gaps or missing data). The matrix Delta Value (mDV; Holland *et al.*, 2002; calculated using DIST_STATS (Auch *et al.*, 2006) of the concatenated 46-species matrix is low (0.18), indicating a high tree-likeness of the signal contained in the matrix (mDV).

For dating, we built a reduced matrix of 29 species by selecting up to five species per genus based on their individual Delta Values (iDV; Auch *et al.*, 2006, Göker & Grimm, 2008; SI 2), computed from a pairwise model-based (HKY + Γ) distance matrix (see dating). Sequences with low iDV behave more tree-like, whereas sequences with high iDV produce topological incongruence or indecisiveness. Earlier studies have established the monophyly of *Alnus* (Chen & Li, 2004: 34 of the estimated 35 species sampled for ITS); *Corylus* (Forest & Bruneau, 2000: 15 species sampled for 5S; Erdogan & Mehlenbacher, 2000: 12 species samples for nuclear and plastid markers; Whitcher & J., 2001: 13 species sampled for ITS); and *Betula* (Järvinen *et al.*, 2004: 16 species sampled for a nuclear and a plastid locus; Li *et al.*, 2005: 34 species sampled for ITS).

**PHylogenetic Analyses**

Trees were rooted on Casuarinaceae and Ticodendraceae. The individual data partitions did not yield genus-level topological contradictions, and the markers were therefore concatenated (see
Phylogenetic trees were estimated using ML optimization in RAxML-HPC version 7.2.6 (Stamatakis, 2006b) and Bayesian optimization (BI) in BEAST version 1.74 (Drummond and Rambaut, 2012). The ML analyses used the per site rate model (originally labeled “CAT” model), an approximation of the GTR + Γ model (Stamatakis, 2006a) with 25 rate categories, independent models for each data partition, and model parameters estimated over the duration of specified runs. Final model parameters and likelihood were optimized under a GTR + Γ model.

Statistical support came from the fast implementation (Stamatakis, Hoover & Rougemont, 2008) of nonparametric bootstrapping (Felsenstein, 1985) in RAxML with the number of replicates determined by the extended majority rule consensus bootstrap criterion (Pattengale et al., 2009). Competing bootstrap support for alternate phylogenetic splits was investigated with the consensus network module implemented in SplitsTree 4 (Holland & Moulton, 2003, Huson & Bryant, 2006), with edge weights set to “COUNT” (“bipartition networks”, Grimm et al., 2006). Alternative topologies, generated using Mesquite 2.75 (Maddison & Maddison, 2011) were tested using the Shimodaira-Hasegawa test (Shimodaira & Hasegawa, 1999) implemented in RAxML.

Bayesian tree searches relied on the uncorrelated lognormal relaxed clock, the HKY + Γ substitution model with four rate categories, and a Yule tree prior. Monte Carlo Markov chains (MCMC) were run for 10 million generations, with parameters sampled every 1000 generations. Log files were then analyzed with Tracer version 1.5 (http://beast.bio.ed.ac.uk/) to assess convergence and to confirm that the effective sample sizes for all parameters were larger than 200, indicating that MCMC chains were run long enough to reach stationarity. After discarding c. 25% of the saved trees as burn-in, maximum clade credibility trees with median branch lengths based on the remaining trees were produced using TreeAnnotator (part of the BEAST package) and FigTree version 1.3.1 (http://tree.bio.ed.ac.uk).

Molecular Clock Calibrations
Three fossil constraints were used in alternative runs and under strict or relaxed clock models, in each case using gamma prior distributions, with the fossil age as the offset and a shape and scale of 1, which allowed some proportion of nodes to be 4–5 Myr older than the offset. First, 71 Myr-old flower fossils (*Endressianthus miraensis* and *E. foveocarpus*) from Portugal, which are “particularly close to members of the Betulaceae and may represent an extinct lineage at the root of the Betulaceae” (Friis et al., 2003: S201), were used as a minimum constraint on the Ticodendraceae/Betulaceae node. These fossils are of Campanian-Maastrichtian age (Friis et al., 2003). Second, we constrained the split between *Alnus* and *Betula* to minimally late Paleocene (58 Myr), based on the earliest *Alnus*-type pollen from North Bohemia (Konzalova, 1971). Fruiting
Evolution of the Betulaceae, p. 6

Alnus and Betula material is not known until the middle Eocene (48 million years ago; Chen et al., 1999). Third, the crown age of Coryloideae was constrained to minimally 56 Myr old based on the extinct genus Cranea known from late Paleocene (56–59.2 Myr) localities of the Fort Union Formation in Wyoming and from a lower Eocene (41–56 Myr) locality in the Big-Horn Basin (Manchester & Chen, 1998). Parsimony analyses of morphological characters cited in Forest et al. (2005) placed Cranea as sister to Carpinus, Ostrya, and Ostryopsis. Another extinct genus of Coryloideae is Paleocarpinus, known from Paleocene of North America, Southern England, and China (Sun & Stockey, 1992, Manchester & Guo, 1996, Manchester et al., 2004). The involucres and nuts of Paleocarpinus are similar to those of extant Ostryopsis, and the male catkins have triporate pollen so that “the reconstructed plant conforms to the subfamily Coryloideae but cannot be placed in a modern genus” (Manchester & Chen, 1998: 522).

RESULTS

RELATIONSHIPS, DIVERGENCE TIMES AND SUBSTITUTION RATES IN THE BETULACEAE

A neighbor-net shows the genetic distinctness of five of the six genera (Fig. 1). Edge lengths mirror the ML bootstrap and PP values obtained in tree inference. A long, tree-like portion separates the Betuloideae and Coryloideae (Figs. 1, 2; SI 3). Nuclear GBSSI sequences, and a ML tree from the plastid data (Fig. 3), support the genera as mutually monophyletic with a bootstrap support of 92% for Ostrya (5 species of 5–9 species sampled) and 65% for Carpinus (14 of 26–35 species sampled; Fig. 4). The best-scoring ML trees from the 29- and 46-taxon matrices (SI 3) placed Alnus as sister to all remaining Betulaceae and Corylus as sister to Ostryopsis, but with low bootstrap support. Investigation of the bootstrap support for alternative placements (Table 2) revealed low support for Alnus and Betula as sister groups from the concatenated data, but high (83/99.8) support for Ostryopsis as sister clade to Carpinus/Ostrya. A Shimodaira-Hasegawa test showed that the placement of Alnus, Betula, and Ostryopsis preferred by the Bayesian analysis (below) was not significantly worse than the topology of the best-scoring ML tree (Table 3).

Divergence times and their 95% CIs for important nodes under strict and relaxed clock models with alternative fossil calibrations are shown in Table 4. Figure 2 shows the chronogram obtained under a relaxed clock calibrated with Endressianthus (71 Myr at the node marked in Fig. 2). The ages of the most recent common ancestors of the extant species of the six genera range from the middle to late Miocene (Fig. 2, Table 4). The split between the two subfamilies is inferred as Paleocene/Eocene (63–43 Myr), and the divergence between Alnus and Betula occurred soon thereafter (60–38 Myr). The divergence between Corylus and the remaining Coryloideae is much younger, dating to the early Oligocene (39–22 Myr). The two species of Ostryopsis diverged from each other c. 5.6 Myr ago. In relaxed clock runs, the ucldev and coefficient of variation
parameter both were 0.34 or 0.35, indicating that the 29-taxon-9321-nucleotide matrix is relatively
clock-like. Substitution rates are illustrated by the blue and red coloring of branches in Figure 3.
There is an increase along the stem lineage of the Coryloideae followed by a slowdown in their
crown group to the level observed in the Betuloideae. Median substitution rates range from 0.7 x
10⁻⁵ at the coryloid root to 0.3 x 10⁻⁵ in some branches of Alnus, Betula and Corylus, similar to
plastid substitution rates of 0.67 or 0.37 x 10⁻⁵ subs/site/Myr inferred by Bousquet et al. (1992).

**DISCUSSION**

**RELATIONSHIPS IN BETULACEAE**

The relationships among the six genera of Betulaceae found under Bayesian (Fig. 2), maximum
likelihood (Fig. 2; SI 3; OSA), and distance optimization (Fig. 1; SI 2) support the mutual
monophyly of the two subfamilies more clearly than did earlier studies (summarized in Table 1).
An analysis of relationships in the order Fagales, however, had found Alnus and Betula as
successive sisters to the remaining genera, instead of as sister groups to each other (Li et al., 2004).
We the current amount of DNA data, this question may not be resolvable, especially as long-branch
attraction from the outgroup Ticodendron may be affecting the placement of Alnus (Table 2; Fig. 1).

The slightly contradictory signal coming from the plastid genes, non-coding plastid regions,
and nuclear spacers and introns can be seen by comparing the placements of Ostryopsis relative to
Corylus and Carpinus/Ostrya in Figures 1, 2, and 3. The latter two genera differ mainly in their
involucral morphology, which in Carpinus consists of a prominent unilateral wing (fig. 23 in
Manchester & Crane, 1987), in Ostrya of an utriculate envelope that completely encloses the nutlet
(fig. 24 in the same study). The nuclear GBSSI sequences support the mutual monophyly of the two
genera (Fig. 4), while nuclear nitrate reductase gene sequences indicate that Ostrya may be nested
within Carpinus (Li, 2008). We found that the latter marker cannot unambiguously be aligned
across genera of Betulaceae (see NEXUS file; Supplementary Information).

**RECONCILING A DEEP FOSSIL RECORD AND RELATIVELY YOUNG**

**MOLECULAR DIVERGENCE TIMES OF EXTANT BETULACEAE**

Inferring divergence times in the Betulaceae has proven more challenging than one would expect
from their good fossil record. Angiosperm-wide molecular dating efforts placed the split between
Betulaceae and Casuarinaceae at 35–37 or 27–29 Myr and that between Alnus and Betula at 19–25
or 18–20 Myr (Wikström, Savolainen & Chase, 2001, Bell, Soltis & Soltis, 2010), dates that are
much younger than those inferred from fossils, which show that Casuarinaceae, Betulaceae, and
Ticodendraceae had differentiated by 86 and 71 Myr ago (Crane et al., 1990, Scriven & Hill, 1995,
Chen et al., 1999, Friis et al., 2003, Manchester, 2011). Both clock studies used nuclear 18S rDNA
and plastid \textit{rbcL} and \textit{atpB} genes, with the first relying on a single calibration point (the divergence of Cucurbitales and Fagales), the second on 36 calibration points. The reason for the too young ages probably lies in the near absence of phylogenetic signal in the 18S data for Fagales, meaning this marker cannot be used to infer divergence times from genetic branch lengths, and the short branch lengths from \textit{rbcL} and \textit{atpB}. The former shows 16, the second 14 substitutions between the outgroup \textit{Ticodendron} and the ten species of \textit{Betula} and \textit{Alnus} in our matrix, but Wikström \textit{et al.} (2001) and Bell \textit{et al.} (2010) included a single species from each of these genera, resulting in their young inferred divergence times.

Compared to the (too) young ages inferred in these angiosperm-wide studies, the ages inferred in the Betulaceae-focused work of Forest \textit{et al.} (2005) were surprisingly old, in several cases exceeding earliest paleobotanical occurrences of Betulaceae. The number of species sampled by Forest \textit{et al.} was similar to ours (26 vs. 27 Betulaceae), but only 462 nucleotides were used compared to the 9321 used here, and fossils were applied as single fixed calibration points, rather than as prior probability distributions (such as the gamma distributions used here; \textit{Material and Methods}). Forest \textit{et al.} also focused on median ages obtained with 10 alternative placements of five fossils, meaning that each fossil’s effect entered the median twice, once when that fossil was assigned to the crown, once when it was assigned to the stem. Of the five fossil records used to fix the crown or stem ages (\textit{Alnus} 65 Myr; \textit{Betula} 49 Myr; \textit{Corylus} to 49 Myr, \textit{Carpinus} 49 Myr, \textit{Ostrya} 33.7 Myr), the \textit{Alnus} constraint is also used here. Our other constraints were \textit{Endressianthus} (Friis \textit{et al.}, 2003) for the Betulaceae stem lineage and \textit{Cranea} (Manchester and Chen, 1996, 1998) for the Coryloideae stem lineage because we wanted to use \textit{Corylus}, \textit{Carpinus}, and \textit{Ostrya} fossils to cross-validate our DNA-based estimates.

Cross validation (against other fossil evidence) shows that the inferred divergence times match the fossil record well (Fig. 3). Thus, fruits and leaves assignable to \textit{Ticodendron} (\textit{Ferrigocarpus}) have been found in early mid-Eocene deposits from Oregon (48.32 Myr) and in the London Clay (ca. 50 Myr old; Manchester, 2011). Since \textit{Ticodendron} consists of a single surviving species, these fossils cannot be assigned in a DNA phylogeny. Nevertheless, they indicate that the split between Ticodendraceae and Betulaceae must be at least 50 Myr old, which does not conflict with our constraint of this split to 71 Myr, based on the \textit{Endressianthus} fossil (in one run). Also, the age of 49 (38–60) Myr that we inferred for the \textit{Alnus/Betula} split using the \textit{Endressianthus} calibration fits reasonably well with the earliest \textit{Alnus} foliage and ovulate cones from the 49–52 Myr-old McAbee flora of British Columbia, a site that also yielded \textit{Betula leopoldae} leaves, catkins and pollen (Crane & Stockey, 1987, Dillhoff \textit{et al.}, 2005). Third, the age of 53 (43–63) Myr we inferred for the Coryloideae stem lineage (when this was not itself constrained) overlaps with the earliest fossils of Coryloideae from late Paleocene/Thanetian (56–59.2 Myr). The earliest fossils
assigned to the genera *Corylus* and *Carpinus* appear in the early Eocene Klondike Mt. Formation dated to 50–49 Myr (Pigg, Manchester & Wehr, 2003); they were used by Forest et al. (2005) to fix crown or stem ages of these genera to 49 Myr. The ages inferred here for the *Corylus* and *Carpinus/Ostrya* stem lineages are 29 (22–39) Myr and 15 (10–21) Myr (Fig. 2, Table 4), which is 20 and 34 Myr younger than the Klondike fossils, suggesting that these fossils represent extinct precursors.

The main earlier effort to date the Betulaceae (Forest et al., 2005) could not use the *Endressianthus* fossil (Friis et al., 2003) as a constraint because of the study’s taxon sampling, which precluded assignment of this fossil (non-parametric rate smoothing does not permit constraining the root itself). The study did, however, use the earliest *Alnus*-type pollen (*Alnipollenites, Paraalnipollenites*) also used here to fix the age of the *Alnus* stem lineage to 65 Myr (while we assigned a gamma prior with an offset at 58 Myr, allowing a proportion of the inferred ages to be 62 Myr or older). With this calibration, Forest et al. inferred an age for the Ticodendraceae/ Betulaceae split of 70.8 Myr (their table 3, using ML branch lengths; 86.5 Myr using ACCTRAN branch lengths, and 81.2 Myr using DELTRAN branch lengths), nicely matching the age of *Endressianthus* (71 Myr). The same calibration gave an age for the Coryloideae crown group of 41.7–57.9 Myr (Forest et al.’s table 3), in agreement with the Coryloideae fossil record, for example, *Cranea* and *Palaeocarpinus* known from late Paleocene (56–59 Myr) and early Eocene (41–56 Myr; Sun & Stockey, 1992, Manchester & Guo, 1996, Manchester et al., 2004). When Forest et al. (2005) instead fixed the age of the most recent common ancestor of the six *Alnus* species in their 462-nucleotide long ITS tree to 65 Myr, they obtained ages of 130.5 Myr for the Ticodendraceae/ Betulaceae node and 76.9 Myr for the Coryloideae crown group, exceeding oldest fossils of these clades. Conversely, when they did not fix the ages of *Corylus* and *Carpinus*, but only used their *Alnus/Betula* stem constraint of 65 Myr, they inferred ages of 15.1 Myr for the *Corylus* crown and 24.3 for the *Carpinus* crown, very similar to our results. Their oldest inferred ages always resulted from fixing the ages of genus-level crown groups to Eocene times.

Placing loose prior probabilities on stem lineage ages, not crown ages, is usually preferable based on two arguments: First, it is implausible that the most recent common ancestor of a handful of living species that barely differ in >9000 nucleotides of plastid, mitochondrial and nuclear DNA lived 65 Myr ago. Clades that survive that long are either extremely species-rich or show long genetic branches. Second, following Doyle and Donoghue (1993), fossils should be assigned conservatively, that is, as constraining the ages of stem lineages, not crown groups. Placing five fossils at ten nodes and then focusing on the median age obtained with all of them, also must be diffusing any correct temporal signal contained in the genetic branch lengths.
Although the inferred median substitution rates (from $0.7 \times 10^{-5}$ at the coryloid root to $0.3 \times 10^{-5}$ in some branches of *Alnus*, *Betula* and *Corylus*) are normal for woody angiosperms (Albert et al., 1994), there clearly is an increase along the stem lineage of the Coryloideae followed by a slowdown in their crown group to the level observed in the Betuloideae (Fig. 2). The branch length differences (and rate shifts) may have to do with average seed dispersal and gene flow distances that could influence species formation. Pollen and seeds of *Betula* are carried over significant distances (e.g., Ford, Sharik & Feret, 1983, Hjelmroos, 1991, Matlack, 1991), partly facilitated by climates with continuous ice/snow covers during winter (D- and E-type climates; SI 4) allowing for secondary dispersal (Matlack, 1989). Birch pollen can be detected across Scandinavia before the flowering season of local birch populations (Hjelmroos, 1991, Skjøth et al., 2007), and backward trajectory analyses suggest pollen dispersal over long distances in northeastern (Šauliene & Veriankaite, 2006) and northwestern Europe (Skjøth et al., 2008). *Betula* is also the most cold-tolerant genus among the northern hemisphere Fagales, with some species extending into high-alpine and arctic environments (Dfc, Dfd, ET climates according to Köppen-Geiger; Kottek et al., 2006; SI 4). *Alnus* resembles *Betula* in its pollen and seed dispersal (Ridley, 1930), and its species also cover a wide range of climates (SI 4), although they all depend on ready access to ground water (Tallantire, 1992, Prieditis, 1997). These ecological traits may explain why *Alnus* and *Betula* were able to diversify into new habitats (Eocene; Mai, 1995).

By contrast, *Corylus*, just like *Fagus* and *Quercus*, relies on jays for dispersal (*Garrulus*, Corvidae; Haffer & Bauer, 1993) as well as other stationary birds and small mammals (Ridley, 1930). By contrast, the winged seeds of *Carpinus*, *Ostrya*, and *Ostryopsis* are wind-dispersed (Ridley, 1930), although they may not cover the same distances as *Betula* and *Alnus* seeds. Our sparse species sampling, however, precludes a more in-depth analysis. The inferred substitution rate shift in the Coryloideae between 25 and 15 million years ago (Fig. 2; Table 4) precedes the mid-Miocene climatic optimum (Zachos et al., 2001) and may relate to new niches that became available after the mid-Miocene optimum.

**Conclusions**

Using curated GenBank data is becoming ever more important as researchers construct huge matrices from public data repositories (Bininda-Emonds, 2004; McMahon and Sanderson, 2006; Chatterjee et al., 2009). Starting from more than 900 available Betulaceae sequences in GenBank, we built 46- and 29-species matrices of 9321 aligned nucleotides that balanced the trade-off between taxon and DNA sampling. However, instead of choosing a single placeholder accession for each species as is usually done, we computed species consensus sequences and then used the
criterion of “tree-like signal” to select the most suitable of available Betulaceae species. The resulting networks and phylogenies suggest that *Ostryopsis* is indeed the sister to *Carpinus/Ostrya* and that the Betuloideae and Coryloideae are mutually monophyletic, which was not clear before. We used relatively few of the many Betulaceae fossils as constraints, preferring to instead use them to cross-validate inferred node ages. Inferred substitution rate shifts in a few places in the phylogeny may be linked to different rates of allopatric species formation, but this needs to be tested with a denser phylogenetic, geographic, and ecological species sample.

**ACKNOWLEDGMENTS**

We thank Jianquan Liu for sharing his unpublished sequences of *Ostryopsis*. GWG acknowledges funding by the Vetenskåpradet (Swedish research council), project no. 2008-3726.
REFERENCES


Skjøth, CA, Sommer, J, Stach, A, Smith, M, Brandt, J. 2007. The long-range transport of birch (Betula) pollen from Poland and Germany causes significant pre-season concentrations in Denmark. Clinical and Experimental Allergy 37: 1204–1212.


Figures and Tables

**Figure 1.** Neighbor Net splits graph from the 46-taxon matrix of 9321 aligned nucleotides of nuclear and plastid sequences. Values at edges refer to ML bootstrap support from 350/1000 replicates, with the first value obtained from the 46-taxon matrix, the second from the 29-taxon matrix. “A” indicates the outgroup-inferred root from ML tree searches and “B” that from the Bayesian inference (compare Fig. 3). Abbreviations: NA, North American spp. of *Ostrya*; EA, Eurasian species of *Ostrya*.

**Figure 2.** A chronogram for the Betulaceae obtained under a relaxed clock model applied to an alignment of 29 species and 9321 aligned nucleotides (*Materials and Methods*). Labels above or below branches indicate support from, in the following order, Bayesian analysis of the 29-taxon matrix, and ML bootstrapping of the 29- and 46-taxon matrix. Labels at nodes indicate 95% posterior probability intervals. The geological time scale is in million years and follows Cohen et al. (2012); ranges of modern genera and extinct fossil lineages shown by gray bars. The geographic origin of each accession is shown after the species name: (1), Australasia; (2), Central America (incl. Mexico); (3), eastern N. America; (4) (cool) temperate northeastern Asia (incl. Siberia, Russian Far East, Mongolia, N.E. China, N. Korea, and N. Japan); (5), Himalayas foothills and adjacent mountain areas in S.E. Asia and Southwest China; (6) (temperate) western Eurasia; (7), Andes; (8), S. W. China (Yunnan, Sichuan, Guizhou); (9) Mountainous areas in N.C. China; (10) Rio Grande valley, N. America; (11) Mountainous areas in N. Vietnam and/or S. China (incl. islands of Hainan and Taiwan); (12) (warm) temperate East Asia (C. and E. China, Korea, Japan); (13) Eastern Mediterranean hinterland.

**Figure 3.** Unrooted maximum likelihood tree based on all available nuclear GBSSI sequences (matrix included in OSA; referenced as K.-O. Yoo and J. Wen, unpublished [submitted 2008] in GenBank). Species of *Carpinus* and *Ostrya* form sister clades, with moderate to high support. Edge thickness represents low (<50, thin), moderate (50–90, medium) and high BS support (>90; thick lines). Exact values indicated for interior edges. The taxon sampling is comparable to that of the concatenated 46-taxon matrix. Species not covered by the 46-taxon matrix are marked in gray. Two ingroup species (*Carpinus tschonoskii* and *Ostryopsis davidiana*) and the two outgroups in the 46-taxon matrix lack GBSSI sequences. “A” indicates the outgroup-inferred root from ML tree searches and favored by the Fagales dataset of Li et al. (2004); “B” indicates the root here inferred under Bayesian inference and favored by most other studies with low to moderate support (compare Fig. 2; Tables 1, 2).
### Table 1. Support for genus relationships within the Betulaceae based on different taxon and gene samples. Numbers in square brackets refer to the number of species (according to the original papers; see SI 1 for identification or sequencing errors). ‘NA’, clade not represented in tree; ‘~’, according bipartition not testable. Clades that received moderate to high support independent of gene- and taxon-sampling highlighted in bold: ‘BS$_{NJ/P}$’, Bootstrap support under NJ or MP; PP, Bayesian posterior probabilities.

<table>
<thead>
<tr>
<th>Clade</th>
<th>Kato et al. (1998) matK BS$_p$</th>
<th>Li et al. (2004) Multigene† BS$_p$/PP</th>
<th>Forest et al. (2005); 5S IGS+ITS BS$_p$/PP</th>
<th>Li et al. (2005) ITS BS$_p$</th>
<th>Yoo &amp; Wen (2002, 2007); ITS/Morph, multigene§ BS$<em>p$/BS$</em>{p}$, BS$_p$/PP</th>
<th>Li (2008); NIA BS$_{NJ}$/BS$_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betulaceae</td>
<td>100</td>
<td>100/1.0</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>Betuloideae*</td>
<td>64</td>
<td>79/1.0</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>Betula</td>
<td>~ [1]</td>
<td>~ [1]</td>
<td>100/1.0 [7]</td>
<td>100 [34]</td>
<td>~, <del>/</del> [1]</td>
<td>~ [100/100 [3]</td>
</tr>
<tr>
<td>Coryloideae*</td>
<td>100</td>
<td>100/1.0</td>
<td>96/1.0</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>C.-O.-Os.</td>
<td>NA</td>
<td>NA</td>
<td>72/0.94</td>
<td>48</td>
<td>&lt;50/&lt;50, &lt;50/0.5</td>
<td>86/NA**</td>
</tr>
<tr>
<td>Ostryopsis (Os.)</td>
<td>~ [1]</td>
<td>~ [1]</td>
<td>~ [1]</td>
<td>~ [1]</td>
<td>~, <del>/</del> [1]</td>
<td>~ [1]</td>
</tr>
<tr>
<td>Co.-Ox.</td>
<td>99</td>
<td>88/1.0</td>
<td>NA</td>
<td>NA</td>
<td>NA, NA</td>
<td>NA</td>
</tr>
<tr>
<td>C.-O.</td>
<td>100</td>
<td>100/1.0</td>
<td>95/1.0</td>
<td>82</td>
<td>93/50, 79/0.97</td>
<td>98/91</td>
</tr>
<tr>
<td>Betuloideae</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>99</td>
<td>&lt;50/&lt;50</td>
<td>99</td>
</tr>
<tr>
<td>Coryloideae split*</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
</tbody>
</table>

*The mutual monophyly of Betuloideae/Coryloideae can only be tested in rooted trees.

†Three plastid genes (atpB, matK, rbcL); one plastid intron (trnL), one nuclear ribosomal RNA gene (18S rDNA), and one mitochondrial gene (matR). The signal stems mostly from the matK partition, with the trnL data forcing Alnus as sister to the remaining Betulaceae, while matK indicating an Alnus/Betula sister relationship (Kato et al., 1998; G. Grimm, pers. obs.).

‡Sampling included multiple accessions from the same species or individuals.

§One plastid gene and two intergenic spacer regions (matK, trnL-trnF, trnH-psbA)

**In the parsimony tree shown in Li (2008), Corylus and Ostryopsis are swapped, the resulting Corylus/Carpinus/Ostrya clade receiving a BS$_p$ of 56.
Table 2. ML bootstrap support for Betulaceae generic relationships in the 46-taxon, and 29-taxon data sets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutual monophyly of Betuloideae and Coryloideae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML Betuloideae clade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betula sister to other Betulaceae</td>
<td>&lt;5</td>
<td>&lt;1</td>
<td>35</td>
<td>~</td>
<td>18</td>
<td>10</td>
<td>~</td>
<td>40</td>
</tr>
<tr>
<td>BI Betuloideae clade</td>
<td>48</td>
<td>61</td>
<td>&lt;10</td>
<td>~</td>
<td>22</td>
<td>&lt;5</td>
<td>~</td>
<td>25</td>
</tr>
<tr>
<td>Position of Ostryopsis within Coryloideae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML Ostryopsis sister to Corylus</td>
<td>&lt;1</td>
<td>13</td>
<td>&lt;10</td>
<td>&lt;1</td>
<td>31</td>
<td>13</td>
<td>17</td>
<td>&lt;5</td>
</tr>
<tr>
<td>BI Ostryopsis sister to Carpinus/Ostrya</td>
<td>99.8</td>
<td>83</td>
<td>53</td>
<td>99</td>
<td>45</td>
<td>19</td>
<td>58</td>
<td>95</td>
</tr>
<tr>
<td>Ostryopsis sister to other Coryloideae</td>
<td>&lt;1</td>
<td>&lt;5</td>
<td>27</td>
<td>&lt;1</td>
<td>23</td>
<td>45</td>
<td>15</td>
<td>&lt;5</td>
</tr>
</tbody>
</table>

*aFlag (ML, BI) indicates ML-preferred (best-scoring 29-/46-taxon trees) and Bayesian-preferred (Fig. 3) phylogenetic splits.

Table 3. Results of the Shimodaira-Hasegawa test based on the 46-taxon data set. The Bayesian topology is shown in Fig. 3.

<table>
<thead>
<tr>
<th>Topology</th>
<th>ln L (GTR+□)</th>
<th>D(L)</th>
<th>Standard deviation</th>
<th>Significantly worse (p&lt;0.01) than best-scoring ML tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best-scoring ML tree</td>
<td>-23297.441657</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>Alternative A: Betuloideae clade</td>
<td>-23299.413332</td>
<td>-1.971675</td>
<td>2.680129</td>
<td>No</td>
</tr>
<tr>
<td>Alternative B: Ostryopsis sister to Carpinus/Ostrya</td>
<td>-23301.058958</td>
<td>-3.617301</td>
<td>5.783962</td>
<td>No</td>
</tr>
<tr>
<td>Bayesian-preferred topology (= A &amp; B)</td>
<td>-23299.279698</td>
<td>-1.838041</td>
<td>5.250697</td>
<td>No</td>
</tr>
</tbody>
</table>
Table 4. Estimated node ages in million years (Myr) for selected divergence events under a strict clock model and a relaxed clock model using different calibration fossils as indicated above columns. Ages are in million years, and the values in brackets are the 95% posterior probability intervals.

<table>
<thead>
<tr>
<th>Nodes of interest</th>
<th>Molecular clock model</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Betulaceae stem 71</td>
<td>Betulaceae stem 71</td>
<td>Corylus stem 56</td>
<td>Corylus stem 56</td>
<td>Alnus stem 58</td>
</tr>
<tr>
<td>Strict</td>
<td>Constrained</td>
<td>Constrained</td>
<td>154 (178-133)</td>
<td>133 (175-101)</td>
<td>92 (105-80)</td>
</tr>
<tr>
<td>Relaxed (Fig. 3)</td>
<td></td>
<td></td>
<td>53 (63-43)</td>
<td>113 (126-99)</td>
<td>98 (124-77)</td>
</tr>
<tr>
<td>Betulaceae Stem</td>
<td></td>
<td></td>
<td>52 (58-46)</td>
<td>113 (126-99)</td>
<td>98 (124-77)</td>
</tr>
<tr>
<td>Betulaceae crown group</td>
<td>46 (52-39)</td>
<td>99 (114-84)</td>
<td>91 (118-68)</td>
<td>Constrained</td>
<td>Constrained</td>
</tr>
<tr>
<td>Alnus/Betula Crown</td>
<td>26 (30-22)</td>
<td>29 (39-22)</td>
<td>Constrained</td>
<td>Constrained</td>
<td>33 (39-28)</td>
</tr>
<tr>
<td>Coryloideae Crown</td>
<td>23 (27-19)</td>
<td>25 (34-18)</td>
<td>50 (55-45)</td>
<td>49 (56-41)</td>
<td>30 (35-25)</td>
</tr>
</tbody>
</table>

**ONLINE SUPPORTING MATERIAL**

**SI 1.** Notes on data curation, including notes on apparently erroneous sequences in GenBank

**SI 2.** Distance statistics

**SI 3.** The best-scoring ML trees inferred from 29- and 46-taxon matrices.

**SI 4.** Geographic and climatic distribution of Betulaceae with a graphical overview (Fig. S1) and tabulated data (File S1).

**Figure S1.** Map of the geographical and climatic distribution of Betulaceae. The pie charts show the number of species and proportion of genera divided by large scale (bio)geographic unit and climate zones (details provided in File S1). Base map from Kottke et al. (2006; freely available via [http://koeppen-geiger.vu-wien.ac.at/](http://koeppen-geiger.vu-wien.ac.at/)). Red tones: equatorial (tropical) climates (A-climates); yellow tones: Dry climates (deserts, BW, and steppe, BS); light green tones: warm temperate, summer-dry climates (Cs), lush green tones: warm temperate, fully humid climates (Cf), brown tones, warm temperate, winter-dry climates (Cw); pink/purple tones: snow climates (D); blue tones: polar climates (tundra, ET, and frost climate, EF).
**File S1.** An XLSX file with tabulated distribution and climate data for Betulaceae species, compiled from various data sources as indicated. See sheet “Readme!” for further content and instructions.
Figure 2

202x143mm (300 x 300 DPI)
Figure 3

![Tree Diagram]

207x157mm (300 x 300 DPI)