Dissecting the Fagales multigene tree: When topological incongruence adds up to fine support

Abstract − The investigation of systematic relationships has greatly benefited from the application of multigene analyses, however, phylogenetic trees should be treated with caution considering the resultant structure of subtrees, in particular when such a reconstruction is interpreted as the sole basis to define phylogenetic units and infer inclusive common origin (monophyly s.str.; holophyly). In this study, I will demonstrate for the order Fagales that even well supported nodes in a multigene phylogram may be misleading in an evolutionary context. Firstly, the amount of incongruence among the single gene topologies that have been combined to form the basis of the multigene data set is investigated and documented. In a second step, the data from the six combined partitions (nuclear ribosomal 18S rDNA, mitochondrial *matR* gene, chloroplast *atpB, matK,* and *rbcL* genes, and *trnL* intron/*trnL-trnF* intergenic spacer) are used to put forward and to test alternative hypotheses. Third, signal from ribosomal spacer data sets is accessed using a combination of classic phylogenetic tree-building (bootstrapping, Bayesian analysis) and phylogenetic networks. The latter data have been shown to be prospective for investigating intergeneric and intrageneric relationships. Finally, the results of the multigene analyses and spacer analyses are brought into agreement.

Introduction

The molecular systematics of the Fagales are thought to be fairly resolved based on several studies,[1] both at the inter- and intrafamily level. Li et al.[2] reconstructed a comprehensive multigene phylogeny based on concatenated sequence data of six gene regions comprising data from all three plant genomes, which exhibited numerous well supported clades. This study resolved both inter- and intrafamilial relationships to an appealing degree and the authors concluded that combining the data from several genes allowed a better resolution than reconstructions based on the single genes. However, several results at the family level of the phylogenetic synopsis shown were in contrast to other data sets, e.g. much broader sampled single-gene data (mostly from the nuclear-encoded ribosomal spacers). In particular, the phylogenetic position of genera within each family subtree was differing. According to Li et al.[2], *Alnus* is sister to a clade comprising all other Betulaceae; within this clade, *Betula* is sister to a clade comprising two subclades, *Carpinus* + *Ostrya*, and *Corylus* + *Ostryopsis*. Support via parsimony nonparametric bootstrapping (BS_P) and posterior probabilities (PP) for each branch in the subtree was convincingly high ($BS_P \geq 99$; PP = 1.0). The long-branching

Ticodendron (monotypic Ticodendraceae) was highly supported as the sister taxon of the Betulaceae. Using nuclear encoded ribosomal DNA (rDNA) spacer data of Betulaceae (5S intergenic spacer, 5S-IGS; internal transcribed spacers ITS1 and ITS2 of the 35S rDNA cistron) and *Ticodendron* as outgroup, Forest et al.[3] found, however, that *Alnus* and *Betula* are sistertaxa (BS_P=79, PP=1.0); this *Alnus-Betula* clade was sister to a clade comprising the other Betulaceae, which also received high support (BS_P=96, PP = 1.0). Phylogenetic structure within the latter clade was found to be slightly differing from the corresponding subtree in Li et al.[2]: *Corylus* formed a sister clade to the remaining taxa *Carpinus, Ostrya,* and *Ostryopsis* (represented by a single accession; BS_P=72, PP=0.94). As in Li et al.[2], *Carpinus* was sister to *Ostrya*. With exception of *Carpinus* (BS_P=65; PP=0.97), branches defining a genus' root received high support (BS_P = 93/100; PP = 1.0).

In case of the Fagaceae, the 6-gene phylogeny was interpreted towards three basic lineages,[2] namely *Fagus, Trigonobalanus* (its relatives *Colombobalanus* and *Formadendron* were not included), and a clade comprising all the other taxa (*Castanea, Castanopsis, Lithocarpus, Quercus*; *Chrysolepis* not included), which has also been inferred based on ITS data, although with low support[4] (see ref.[5] for a critical re-evaluation). A sistertaxon relationship between *Castanea* and *Castanopsis* was indicated with moderate support (BS_P=67, PP=0.97); a close (sister) relationship between these two taxa can also be deduced based on broad-sampled ITS data, however, only by using distance-based phylogenetic networks but not phylogenetic trees.[5] Further intergeneric relationships are basically unresolved up to now despite the availability of exemplary multigene data, which is due to the miscellaneous signal from intrageneric lineages within *Quercus,* dissolved as a clade by the multigene data (discussed in ref.[5]).

The third Fagales family comprising several genera are the Juglandaceae. *Engelhardia* was placed as sister to other Juglandaceae, the latter formed a clade with an unresolved basal trichotomy (*Annamocarya, Platycarya*, other taxa).[2] A sistertaxon relationship was inferred for *Carya* and *Juglans* (BS_P=999, PP=999), whereas the exact position of *Cyclocarya* and *Pterocarya* in relation to the putative sister taxon pair *Carya* + *Juglans* could not be finally resolved.[2] The distinction between the two major clades (ranked as subfamilies[1]) within the Juglandaceae, the Engelhardoideae (*Engelhardia, Alfaroa, Oreomunnea;* the latter two not included in ref.[2]) and Juglandoideae (remainder), was confirmed using a combination of three noncoding sequence regions, ITS, *atp*B-*rbc*L spacer, and the *trn*L-*trn*F intergenic spacer (and a morphological partition).[6] *Juglans* was, however, moved to a clade comprising *Cyclocarya* and *Pterocarya*, but not *Carya*.[6]

Regarding interfamily relationships, the phylogenetic backbone, the two included outgroup taxa (*Celtis, Hamamelis*) placed the all-Fagales root between *Nothofagus* and the remaining Fagales, and the Fagaceae were recognized as the second diverging lineage as in earlier single-gene trees focussing on angiosperm relationships and including several members of the Fagales.[7,8] As in earlier studies based on a single or several genesREF, *Rhoiptelea* (monogeneric Rhoipteleaceae) was placed as sister taxon of the Juglandaceae (BS_P=100, PP=1.0); a new finding was that the *Rhoiptelea*-Juglandaceae clade was resolved as the sister clade to the Myricaceae (included *Comptonia* and *Myrica*; BS_P=63, PP=0.95).[2] *Ticodendron* (monogeneric Ticodendraceae) was supported as sister taxon of the Betulaceae, and *Casuarina* (monogeneric Casuarinaceae) as sister to Betulaceae-*Ticodendron* (BS_P=100, PP=1.00).[2]

Phylogenetic tree reconstructions may be affected by systematic bias (reviewed in ref.[9]). For example, increasing the number of used nucleotides by adding data from various gene regions has been shown to artificially increase support (BS, PP) of branches that may be erroneous.[10] Single-gene topologies are often (partially) incongruent to each other and to a combined tree, but many multigene studies regard this incongruence as insignificant based on the outcome of the incongruence length difference (ILD) test.[11] Simulation studies show, however, that the ILD test, also used by Li et al.[2], can be insufficient to identify incongruence of concatenated genes.[12-14] Another problem arises from a known problem: the effect of outgroups to define the root of an ingroup. Adding an outgroup can substantially distort the ingroup topology in simulations[15], but also real-world studies[16] challenge the general potential of outgroups to infer a reliable ingroup root. Is there a similar effect for internal roots of any possible subtree, and could this explain why different data sets recognize different grades and clades? For example, in the case of the potential subfamily Betuloideae (*Alnus, Betula*), which formed a grade in the 6-gene phylogeny,[2] but a clade based on broadly sampled sequence data from two nuclear spacer regions,[3] the two alternatives differ only by the placement of the Betulaceae root as inflicted by *Ticodendron* as the sister- or outtaxon (Fig. 1). Last but not least, disproportionally distinct (or erroneous, see ref.[17], for dubious *rbc*L sequences stored in sequence databases) sequences of a single partition may distort the topology locally, but otherwise remain undetected because of the stabilizing effect of the remaining partitions in the concatenated data matrix, which can overrule an incompatible (potentially erroneous) signal.

At the example of the Fagales, I will demonstrate that phylogenetic tree-building and analyses of concatenated data using a selection of several genes may have certain pitfalls. Pitfalls easily detected and avoided, if phylogenetic reconstruction is not reduced to the single

inference of a combined tree, as commonly done in systematic botanical studies. Using long-known pair-site tests to test for topological incongruence such as the SH-test[18] and additional analyses of subsets of the concatenated data, the incongruence between the concatenated partitions and influence of each partition on the multigene tree can be easily revealed and visualized using consensus networks.[19] 'Bipartition networks',[20] a special type of consensus networks, complement the traditional phylogenetic tree-building by visualizing support for alternative and competing phylogenetic splits, which are likely to occur if different gene regions with (partly) incompatible signal are combined. Moreover, I will outline how distance-based analyses (including planar phylogenetic networks[21,22]), often dispraised as "phenetic" or "non-phylogenetic" (for the historical reasons see chapter 10 in ref.[23]; numerous anonymous reviewers, pers. comm. 2002–2011), can help to interpret (*i)* the placement of taxa in a phylogenetic trees, (*ii*) decrease of branch support, and (*iii*) present a means to define stable phylogenetic systematic units.

Methods

Primary incongruence among the partitions combined for the 27-taxa data[2] is investigated using trees based on each single-gene partition, the multigene (6-gene) tree based on all original partitions, and six 5-gene trees, each with one partition excluded from analysis. Phylogenetic trees were inferred based on the primary character matrices (sequence alignments) under parsimony (MP) and maximum likelihood (ML) as optimality criteria and based on distance matrices using uncorrected p (d_p) and ML-based (d_{ML}) pairwise (Hamming) genetic distances with PAUP* 4.0 beta 10[24] and RAxML 7.2.6[25].

Under MP, branch-and-bound searches resulted in the optimal topologies. Branch support was established by nonparametric bootstrapping[26] using 100,000 replicates, each with a single tree optimized by TBR swapping and the MulTrees option deactivated.[27]

Under ML, RAxML[17,25,28] used ten (standard) or fifty inferences to find the best-known topology,[17,25,28] each one starting from a parsimony tree and using the GTRCAT setting for model optimisation during the run,[29] the final tree being optimized under GTR + Γ model (). Branch support was established by 'fast bootstrapping'[28] under ML (up to 1,000 replicates) as implemented in RAxML and Bayesian posterior probabilities.[30] Number of necessary bootstrap replicates was selected based on the MRE bootsstop criterion[31] implemented in RAxML. Bayesian analyses (BI) were performed with MrBayes 3.1[32] and used 2,000,000 generations, 10 parallel runs with one cold chain and no heated chain (as recommended for

matrices with few taxa); pre-convergence trees were discarded (see Electronic Supplement ES999 for details).

Distance-based phylograms were inferred with the BioNJ algorithm.[33] The BioNJ algorithm represents a modification of the NJ algorithm[34], and can be used to fast reconstruct[23] phylogenetic trees based on distance matrices that fulfil the minimum evolution criterion.[23,34,35] The neighbour-net (NN) algorithm[21,22], implemented in SplitsTree[9] was used to compute distance-based phylogenetic networks.

Paired-sites tests (KH test[36]; SH test[18] ADD AU test/CONSEL stuff) were used to check for significant topological incongruence among single-gene ML phylograms using PAUP* and RAxML. For the KH test (PAUP), the RELL option was used with 10,000 replicates on each single-gene matrix as suggested in ref.[23]. In addition to the statistical tests, consensus networks[19] allowed visualizing topological conflict (computed with SplitsTree 4[9]).

MAKE "PROFILE PROFILE ALIGNMENT" In addition to the published phylogenetic trees, the ITS data available from the gene banks are analysed at the genus and family level. Where applicable (if alignable), phylogenetic reconstructions were performed. Multiple alignments have been computed for such genera with broad ITS data basis to assess the intrageneric and intergeneric ITS divergence, highly similar sequences (lacking major length polymorphism; *Lithocarpus* was treated separately) have been combined to strict consensus sequences to facilitate intra- and interfamilial analysis and compute an alignment at ordial level for comparative oligonucleotide motif analysis. The data (772 ITS accessions representing 293 species) of each genus could be condensed to 130 consensus sequences representing the following members of the Fagales and sufficiently reflecting the ITS divergence of this group (Table 3 Results of the 5-gene ML analyses, excluding one of the six original partitions..

Table): the Betulaceae *Alnus* (20 ITS accessions screened and analysed representing 20 species), *Betula* (20/17), *Carpinus* (44/18), *Corylus* (49/18), *Ostrya* (10/6), *Ostryopsis* (3/2); the Fagaceae *Castanea* (4/3), *Castanopsis* (22/16), *Chrysolepis* (2/2), *Colombobalanus* (1/1), *Fagus* (237/8), *Formadendron* (2/1), *Lithocarpus* (76/41), *Quercus* (159/64), *Trigonobalanus* (1/1); the Juglandaceae *Annamocarya* (1/1), *Alfaroa* (4/4), *Carya* (12/8), *Cyclocarya* (2/1), *Engelhardia* (2/2), *Juglans* (43/15), *Oreomunea* (1/1), *Platycarya* (6/3), *Pterocarya* (2/1); the Myricaceae *Comptonia* (2/1), *Morella* (15/10), *Myrica* (5/2); *Nothofagus* (Nothofagaceae; 25/24, including new data); *Rhoiptelea chiliantha* (Rhoipteleacea; 1 accession); *Ticodendron incognita* (Ticodendraceae, 1 accession). Several pseudogenous and older data of lower

quality was excluded from analysis. Only a single and fragmentary ITS sequence [37] is stored of the Casuarinaceae (*Casuarina equisetifolia* subsp. *equisetifolia*). Hence, we do not refer to *Casuarina* and Casuarianceae. ITS sequences of *Hamamelis* spp. (19/5) have been included as comparative data. In contrast to the second outgroup used in Li et al.[2], *Celtis*, the ITS of *Hamamelis* is possibly exhibiting one of the least derived ITS sequences among eudicots. Conserved ITS motives of *Hamamelis* are highly similar not only to the assumed 'basalmost' Fagales or the consensual state, but also putative less derived taxa (considering the ITS sequences) of other eudicot lineages such as the Sapindales[20] and Proteales (own observations).

Results

Phylogenetic ambiguity expressed in the concatenated data

The best-known ML tree (Fig. 1) based on the concatenated data is largely in agreement to the original phylogenetic synopsis,[2] except for the Juglandaceae subtree: in contrast to ref.[2], *Carya* is recognized and well supported as sister of *Annomocarya*; this clade is placed as sisterclade to the remaining "core Juglandaceae"[2] (= Juglandoideae Eaton[1,6]), which form a moderately to well supported clade (Fig. 2). Within the latter, *Platycarya* represents the first diverged taxon (moderate support), the relationship between the remaining three genera, *Juglans, Cyclocarya,* and *Pterocarya*, is essentially unresolved (Fig. 2; all possible combinations receive equally low support, Fig. 3). Branch support is relatively high along the backbone (independent of the method), with the exception of one branch: a sister relationship between Myricaceae and *Rhoiptelea*+Juglandaceae receives only moderate support (BS_{ML}=62/74, BS_P=???, PP=???; same in ref.[2]), which is due to a less dominant and incompatible signal in the data favouring a second alternative, a sister relationship between the Myricaceae and the *Casuarina*-*Ticodendron*-Betulaceae clade (BS_{ML}=38/26, BS_P=???, PP=???; Fig. 3). Towards the tips, the concatenated data is increasingly indecisive, which is expressed by the box-like structures in Figure 3 (bipartition networks based on the parsimony bootstrapping and the Bayesian analysis are included in the ES). Only the Betulaceae subtree is fully resolved, according branches receive high support ($BS_{ML} \geq 95$, $BS_P \geq$???, PP $\geq$???). In particular, relationships among the members of the subfamily Quercoideae Ørsted in a broad sense[1], i.e. a clade including all Fagaceae except *Fagus*, are not unambiguously resolved; the data prefers either *Trigonobalanus* or *Lithocarpus* as first diverging lineage (Fig. 3).

Regarding the position of *Quercus* the combined data is unable to decide between mainly two alternatives: placing *Quercus* within a clade including *Castanea*+*Castanopsis* (and *Lithocarpus*) or as sister taxon to *Trigonobalanus*. The lowered support for the sister relationships between *Castanea* and *Castanopsis* ($BS_{ML} \geq 95$, BS_P=???, PP=???) is also due to a minor signal in the data supporting a sister relationship between *Castanopsis* and *Lithocarpus* (Figs 2, 3).

A straightforward means to visualize the level of incompatibility in a data set are planar phylogenetic networks based on a pairwise (Hamming) distance matrix. In the case of the concatenated data, a distance matrix based on simple, uncorrected pairwise ("p") distances already captures the same phylogenetic signal expressed in the ML (and MP) tree inferences and branch support analyses to a sufficient degree (Fig. 4A; for results using ML-corrected distances see ES). In the according neighbour-net (NN) splits graph (Fig. 4B), a planar phylogenetic network, the tree-like portions refer to the clades (branches) in the phylogenetic trees, which received (very) high support from bootstrapping and Bayesian analyses. Undoubted monophyletic groups, such as the APG-recognized familiesREF, can be straightforwardly identified in the graph (Fig. 4B). The reason for this is that intrafamily distances are generally lower than interfamily distances (Fig. 5; ES). In addition, ambiguous phylogenetic relationships are represented by (more or less) prominent box-like structures. Based on the NN splits graph, the nature of the miscellaneous signal backing several alternative (ambiguous) relationships in tree-based analyses (Figs 2, 3) can be understood. In the case of the internal, partly ambiguous relationships within the Quercoideae and Juglandoideae clades (Figs 2, 3), patterns of overall similarity are not decisive; a stringent phylogenetic signal from the concatenated data matrix is missing. However, two taxa, *Trigonobalanus* (Quercoideae) and *Platycarya* (Juglandoideae, Fig. 1) are distinctly different from the remainder of the according clade (Fig. 4B), which would render these two taxa as good candidates for the first diverged genera within the according subtrees. The ambiguous signal regarding the position of the Myricaceae (cf. Fig. 2) is linked to the fact that the Myricaceae are only slightly more similar to *Rhoiptelea*-Juglandaceae than to *Casuarina*-*Ticodendron*-Betulaceae (Fig. 4B). Combined with the evidence from the tree-based analyses, it can be concluded that the Myricaceae-*Rhoiptelea*/Juglandaceae sister relationship preferred by the concatenated data is indeed only one of three topological possibilities (the other two being Myricaceae sister to other core higher hamamelids and Myricaceae sister to *Casuarina*/ *Ticodendron*/ Betulaceae), and requires further investigation. The earliest divergences (in

whatever sequence) within the "core higher hamamelids"[2] appear to have occurred relatively fast after each other.

Further phylogenetic information provided by distances

More phylogenetic relevant information can be directly drawn from the NN splits graph (Fig. 4B) and the underlying distance matrix (Fig. 5; ES). The signal from the outgroups is not entirely compatible, and, with regard to the long terminal and connecting edges produced by the two outgroup taxa and *Nothofagus*, ingroup-outgroup long-branch attraction (LBA) may account for misplacing the ingroup (all-Fagales) root (alternative roots are indicated in Fig. 4B, see Discussion below). *Fagus* is markedly different from the other Fagaceae and less distant to any ingroup taxon, in particular to *Nothofagus*, and the two outtaxa than the remainder of the Fagaceae. Based on the overall intra- and interfamily genetic divergence and phylogenetic position (Figs 2–5; ES), it would reasonable to raise the Quercoideae to the family level, since both *Rhoiptelea* and *Ticodendron* are recognized as distinct families and not included in the Juglandaceae or Betulaceae. In addition to *Fagus*, also *Rhoiptelea* and the Myricaceae (to some degree) are apparently less derived from their putative common ancestor (and the common ancestor of all Fagales) than the remaining taxa; with *Rhoiptelea* placed in an ancestor-like fashion to the Juglandaceae. In contrast, both *Casuarina* and *Ticodendron*, the sister lineages of the Betulaceae, are markedly derived as exhibited by long terminal edges. The position of *Ticodendron* and *Casuarina,* and *Alnus* and *Betula*, at opposite sides of the subgraph (Fig. 4B) may be indicative of incompatible signals in the underlying data regarding the exact position of these taxa. Also, it can be seen that none of the sister lineages (*Fagus, Rhoiptelea, Casuarina, Ticodendron*) of 'crown' groups (quercoid clade, Juglandaceae, Betulaceae) is significantly more similar to a particular taxon of the latter; which could be expected if they represent sister lineages.

Incongruence induced by single genes

The cause and extent of incongruence in the concatenated data can be further analysed using the single-gene matrices. The limitation to 27 taxa allows using the Branch-and-Bound (B&B) algorithm under parsimony to find the optimal, most parsimonious tree-like solutions. According to refs[24,38], B&B ensure finding the really optimal topology or topologies (most parsimonious trees, MPT) based on a given alignment. Thus, the MPT via B&B are reflecting equally optimal solutions under parsimony, i.e. equally 'best' and alternative phylogenies. In contrast to the commonly used strict consensus tree to summarize MPT, a consensus network of all MPT computed via B&B allows visualizing all equally parsimonious topological

alternatives based on the single-gene matrices at the same time (Fig. 6). As seen in the consensus network of all MPT (Fig. 6) some relationships indicated in the 6-gene tree are indeed unambiguous among all partitions under MP: all families are consistently recognized as clades (with varying support; Table 1), *Fagus* is clearly separated from the Quercoideae clade, *Rhoiptelea* is sister to the Juglandaceae, and a *Casuarina*-*Ticodendron*-Betulaceae clade is found in all MPT. Aside from these, all other relationships vary among the MPT, i.e. each single gene partition largely seems to favour a different topology or several topologies. Topologies that are partly incongruent to each other but equally optimal under MP (Fig. 6). This is also expressed by the best-known ML trees inferred based on the single-gene matrices (Fig. 7), the Pearson correlation coefficients between ML bootstrap replicate collections based on subsequently filtered data, and the results of a KH and SH tests using the preferred ML topologies based on each single-gene matrix (Table 2; see following paragraphs).

Nuclear encoded 18S rDNA: overall little signal, a moved all-Fagales root, and zero-branch attraction 'supporting' a Juglandaceae-Myricaceae sister relationship

The 18S rDNA data adds, if any, only faint signal to increase overall resolution and support but induce primarily splits incompatible with the other five partitions: the 18S rDNA-preferred topology and ML bootstrap results are significantly incongruent to those preferred by other data sets and the concatenated data (Table 2). Specifically, the Fagales root (inferred by position of *Hamamelis* and *Celtis*) moves towards the 'core higher hamamelids', hence, *Nothofagus* becomes sister to the Fagaceae (Fig. 7A; BS_{ML} = 77; PP = 0.98). The reason for this is a general 18S rDNA similarity of the 'core higher hamamelids' to each other ($d_p \leq 0.01$; with the exception of *Carya* and *Casuarina* to a lesser degree) contrasting an increased genetic divergence among members of the Fagaceae (d_p = 0.02–0.03; 0.01–0.02 in the case of *Fagus*), which, in the case of *Castanopsis* vs. *Trigonobalanus*, even exceeds the difference to *Nothofagus* and the outtaxaon *Celtis*. This situation is strikingly different from the other partitions, based on which, *Nothofagus* is always the genetically most distinct taxon within the ingroup (Fagales), only to be topped by the two outtaxa (ES). Betulaceae and *Ticodendron* are highly similar to each other; the extremely low divergence in this group naturally provides little phylogenetic signal (Fig. 7A; ES). This is even more the case for the Juglandaceae (except *Carya*), *Rhoiptelea,* and the Myricaceae. The according subtree (Fig. 7A) is essentially collapsed; the numerous alternative bipartitions receive diminishing support. This is not surprising, since the 18S rDNA of all these taxa is highly similar to identical (ES). The

placement of the significantly distinct, possibly aberrant, *Carya* within this Juglandaceae-*Rhoiptelea*-Myricaceae clade in the ML tree correlates to the relatively higher similarity between *Carya* and this group compared to the other included taxa. The high support for the Myricaceae subclade within this clade stems from the fact that both sequences are identical and slightly more different to the (nearly) identical Juglandaceae (excluding *Carya*).

Chloroplast *atp*B gene: Differential divergence patterns at various levels

The *atp*B data contributes to a much higher degree to the combined tree shown in Figure 2 than 18S rDNA data. Intrafamily divergence is typically lower than interfamily divergence, and even within a family, the patterns of similarity vary, which finds its representation in the ML tree and branch supports (Fig. 7B). In contrast to the combined tree, *Betula* and not *Alnus* is resolved as the first diverging lineage within the Betulaceae (BS_{ML} = 92/90; PP = 0.997/0.92; Fig. 7B); in a distance framework, however, *Alnus* would be a probable sister taxon to *Betula*. The according alternative topology is not significantly worse than the best-known shown in Figure 7B. The Myricaceae are sister to the other 'core higher hamamelids' in all MPT and the ML tree, but the other two alternatives (see above; Fig. 5) receive similar support (ES). The high number of MPT is due to the unresolved internal relationships in the Quercoideae subtree, the *atp*B region of *Castanea, Castanopsis, Lithocarpus* and *Quercus* is virtually identical, and the position of *Nothofagus* in comparison to the outgroup taxa under parsimony (it is equally parsimonious to place *Nothofagus* as sister to one of the outtaxa than as sister taxon to the remainder of the ingroup).

Chloroplast *mat*K gene: well differentiated at all levels, forming the 'backbone' of the combined tree

Similar to the *atp*B data, and even more pronounced, the *mat*K data exhibit a nice differentiation, both below and above the family level (exemplarily illustrated by the pairwise genetic distances; ES). Clades and sister relationships indicated by the 6-gene tree can be traced in detail using patterns of overall genetic similarity in the *mat*K. Thus, the *mat*K-preferred tree (Fig 7C) is in best agreement with the synopsis (Table 2; cf. Fig. 2), the number of MPT is only due to permutations among all Juglandoideae and Quercoideae, which are all equally optimal under MP; ambiguities resolved to some degree under ML. The only difference is that, like based on *atp*B, *Betula* and not *Alnus* is sister to the remaining Betulaceae under MP, under ML each possible sistertaxon relationship is equally probable (ES). Notably both taxa are, like in the case of *atp*B most similar to each other (d_p = 0.02 vs. ≥

0.03 to other taxa) and are ± equally similar to other taxa. The increased and finely differentiated divergence in the *mat*K data predefines many relationships, which are also recognized based on the concatenated data. Even *mat*K unresolved relationships or low supported branches reduce the topological possibilities: a bipartition supported by any of the other data sets contrasting the *mat*K data is topologically unfavourable based on the concatenated data (independent of the optimality criterion). Only the 6th partition, the *trn*L intron and *trn*L-*trn*F spacer (*trn*L in the following) contributes an equally strong, however largely incongruent, signal (see below).

The mitochondrial *mat*R gene: like the 18S rDNA, but with other odds

If *mat*R is used, the Fagales backbone essentially collapses (Fig. 7D); this partition appears to be of little use to resolve deeper relationships in the Fagales. Betulaceae (with highly similar to identical *mat*R sequences; ES) + *Ticodendron* are still recognized as a clade, but support is low (BS_{ML} = 32; PP = 0.52; $d_p \leq 0.01$ vs. ≥ 0.01 compared to other taxa/groups of taxa). Only Fagaceae (four genera highly similar to identical), Juglandaceae (five genera highly similar to identical), and, naturally, Myricaceae (both taxa highly similar) are still recognized and supported as clades (Fig. 7D). At odds to all other partitions and the 6-gene tree, *Casuarina* is resolved as sister to the Myricaceae with high support (BS_{ML} = 89; PP = 0.97). Regarding the distance patterns the latter may be due to local LBA: all other taxa of the 'core higher hamamelids' are similar to each other, and only *Casuarina* and the Myricaceae are distinct. The phylogenetic position of *Rhoiptelea,* which is highly similar to *Alnus* (Betulaceae!), and relatively similar ($d_p \sim 0.01$) to most Fagales (except *Quercus* and *Nothofagus*), is accordingly unresolved (Fig. 7D; ES). *Quercus* (Fagaceae), on the other hand, is most distinct taxon among the ingroup. Its situation is analogous to the situation of *Carya* regarding the 18S rDNA data; the data need to be verified. In the sum, the mitochondrial *mat*R data, like the nuclear 18S rDNA data, resolves (supports) only relatively few relationships, which are, nevertheless, incongruent to those indicated by the other (plastid) data sets (topological incongruence to all other preferred trees is significant, Table 2) and lacks signal to further support or resolve relationships not pre-defined by *mat*K. However, the topologies of subtrees below family level indicate relationships, which are in some agreement to nuclear-encoded rDNA spacer data (below) and in contrast to the (*mat*K-dominated) 6-gene topology.

Chloroplast *rbc*L gene: a most flexible signal

The divergence patterns in the *rbc*L largely match the situation in the *mat*K data (see above), although at a lower amplitude (ES). Thus, support along the backbone is relatively high because of sufficient signal strength, with support values collapsing towards the leaves of the tree, as often the case in the other single-gene analyses and the 6-gene tree to a lesser degree. But, standing-alone, the signal from the *rbc*L appears insufficient to obtain a reliable Fagales phylogeny: the topologies of the ML tree and the MPTs are quite different to the topologies preferred otherwise. *Ticodendron* and the Myricacea form a sister clade to the other Fagales except *Nothofagus*; *Engelhardia* and *Rhoiptelea* are nested within the Juglandoideae, the Betulaceae are recognized as a clade, but with low support (BS_{ML} = 25, PP = 0.46). Some of the indicated intrafamily relationships agree, like in the case of *mat*R, with reconstructions based on ITS data that differ from the 6-gene tree. The topological indifference of the *rbc*L is highlighted by the KH and SH tests: the *rbc*L-favoured topology is not significantly better than the *atp*B- and *mat*K-inferred topologies, i.e. the *rbc*L data can be brought easily in agreement with the topologies favoured by *atp*B and *mat*K data, even if its ML- and MP-optimized trees would prefer partly incongruent relationships.

Chloroplast *trn*L intron and *trn*L-*trn*F spacer (*trn*L): strong signal, but mostly incompatible with the rest

The three MPT based on *trn*L data differ only in the position of *Juglans, Cyclocarya* and *Pterocarya* to each other, hence, these data appear superficially the most decisive of the six combined partitions. The inter- and intrafamily divergence has the same amplitude than found for *mat*K (ES). The distance patterns of each taxon are differential, with the exception of *Juglans, Cyclocarya,* and *Pterocarya.* For these three Juglandaceae taxa, each sisterclade alternative is equally probable. In striking contrast to *mat*K, the *trn*L-preferred tree (Fig. 6F) is largely incongruent to the other single-gene trees and the 6-gene tree (Figs 2, 6; Table 2; SH and KH test p-values ≤ 0.0002). For instance, in contrast to all other five single-gene trees and the multigene tree, *Betula* is placed as sister to *Carpinus* in an overall weakly resolved Betulaceae subtree, and not as first diverging Betulaceae lineage (); the weak resolution is due to the high similarity of most Betulaceae (except *Alnus*) to each other (ES). *Alnus,* on the other hand, is markedly distinct from the remaining Betulaceae, nearly as distinct as the putative sister taxon of the Betulaceae, *Ticodendron.* Thus, *Alnus* is (and can) only be placed

as sistertaxon to all other Betulaceae, which is highly supported (BS_{ML}=100) based also on the concatenated data.

Stability of multigene topologies

Given that some highly supported relationships in the 6-gene tree could only be recovered based on a single of the six concatenated partitions, the question arises how reliable are moderate or highly supported branches in multigene analyses. In order to assess the impact of a single partition on the topology, a single partition was excluded per run from analysis. According to the results of the 5-gene ML tree inferences and bootstrap analyses (Fig. 7), several relationships indicated by the 6-gene tree are possible data sampling artefacts, i.e. imprints from a single partition that overrules the signal of the other five partitions (Table 3). The moderately supported sister relationship between Myricaceae and the *Rhoiptelea*-Juglandaceae clade (6-gene data; Figs 1, 7A; ref. [2]; but see Figs 2, 3) is a molecular imprint of the 18S rDNA partition; if this partition is excluded from analysis the alternative phylogenetic split promoting a Myricaceae-Betulaceae clade (incl. *Casuarina, Ticodendron*), less supported based on the combined data (Figs 2, 4), receives high support (Table 3). Exclusion of the *atpB* (chloroplast) and *matR* (mitochondrion) genes only effects low- to moderately supported (BS_{ML} < 80; Figs 1, 2) relationships within the Quercoideae (Fig. 5B, D, E). The highly supported sister relationship between *Corylus* and *Ostryopsis* (Figs 1, 7C; cf. ref.[2]) is *matK*-induced. If this partition is eliminated, *Corylus* is placed as sister to *Carpinus* + *Ostrya* with moderate support (Table 3; also reported based on nuclear spacer data[3]). The moderately supported *Juglans-Cyclocarya-Pterocarya* clade is mostly backed up by signal from the *rbcL* gene (Fig. 7F; Table 3). However, since the support of the 5-gene data-favoured alternative of a *Platycarya-Cyclocarya-Pterocarya* clade is low (Fig. 7F; Table 3), neither alternative should be easily rejected as possibility. If the *trnL* partition is excluded, the resultant 5-gene ML tree shows a Betuloideae Arnott clade[1,3], i.e. *Alnus* sister to *Betula*, which receives moderate support (Fig. 5F). The alternative of a Betuloideae grade (Figs 1, 7A–E; cf. ref.[2]) receives only diminishing support (Table 3).

Rare changes: Highly conserved ITS motives

Parts of the ITS region are structurally and sequentially highly conserved among all plants.[39-43] In particular, the ITS region includes the 5.8S rDNA, which is c. 160 nucleotides long and structurally highly constrained.[39,44] Table 4 and Figure 8 show the conservativeness and variability of the 5.8S rDNA in the case of the Fagales, *Hamamelis* is included for comparison. Using the phylogenetic reconstructions as a guideline above, one can plot the

evolution of the 5.8S rDNA, which may then provide some additional evidence to decide between alternative scenarios. Based on the 5.8S rDNA sequence three lineages within the Betulaceae can be characterized ('barcoded'): *Alnus* + *Betula* (= Betuloideae), *Ostryopsis, Carpinus* + *Corylus* + *Ostrya*. This agrees with evidence from *matR* and phylogenetic trees based on 5S-IGS and ITS nuclear spacer data.[3] An accordings alternative topology is only rejected by ... *Ticodendron* is significantly different from all Betulacea. The Juglandoideae and *Rhoiptelea* share exactly the same 5.8S rDNA, whereas 5.8S rDNA sequences of the three genera of the Engelhardioideae are distinct from this consensus and each other (ES). The 5.8S rDNA of the Myricaceae is highly characteristic and identical among genera and species (except for a single site variability in *Morella*). Within the Fagaceae, the 5.8S rDNA highlights the deep phylogenetic split between *Fagus* and the Quercoideae (cf. Figs 2, 4, 6, 7; ES), and their differential relationship to *Nothofagus* (*Fagus*-5.8S is more similar to that of *Nothofagus* than the ones of the Quercoideae; cf. Fig. 4, 6A; ES). Most Quercoideae taxa share the same 5.8S rDNA variant (except *Chrysolepis, Colombobalanus,* and *Formanodendron*), which, in the species-rich genera such as *Castanopsis, Quercus* and *Lithocarpus*, is complemented by 5.8S rDNA variants that can be directly derived from the common variant. The 5.8S rDNA of *Hamamelis* (conserved within the genus) differs by four nucleotides from the strict consensus of all Fagales. Overall, the divergence patterns found in the 5.8S rDNA mirror the situation in the 18S rDNA data (see above; Fig. 6A; ES)

Discussion

Incongruence: Phylogenetic implications

Each of the data sets that were combined for the 6-gene 27-taxa set favour topologies, which are, for a significant part, incongruent to each other (Figs 5, 6; Table 2). Both the single-gene and the 5-gene analyses demonstrate that many signals from the concatenated partitions are incompatible and inflict numerous incongruences (affected taxa highlighted in Figs 6, 7). This is in contrast to the result of the ILD test performed originally,[2] which did not indicate a significant incongruence. The position of *Casuarina,* the Myricaceae, the all-Fagales root (considering the position of Fagaceae to *Nothofagus* and the outgroup taxa), and some aspects of intrafamily relationships (Figs 6, 7; Table 3) can be questioned despite moderate to high support based on the concatenated data (Figs 2, 3; cf. ref.[2]). Only by excluding one of the six partitions significant changes in the tree topology are inflicted (Fig. 7; Table 3; ES). The combined "multigene"[2] topology relies mostly on the relatively strong signal provided by the *mat*K partition, plus a weaker signal from the *atp*B and *rbc*L (Table 2)

partitions where additive (but see Fig. 7B, E). Only the *mat*K data alone would allow depicting a phylogeny not much less resolved than based on the concatenated 6-gene data (cf. Figs 2, 5C); the according conclusion of Li et al.[2] does not hold. Moreover, all signals from the other five partitions (Figs 5, 6), in particular the two non-plastid partitions (18S rDNA; *mat*R), that are incompatible to *mat*K are lost, i.e. not represented in the 6-gene tree, with two exceptions:

(i) The *trn*L data constraints the final placement of *Alnus* and rejects alternative placements, which would, at least, be equally likely based on the other five partitions (Figs 5–7; Tables 2, 3) and favoured by ITS and 5S-IGS data.[3] The *trn*L sequence of *Alnus* is relatively distinct to all other Betulaceae (ES); and the placement of *Alnus* as sister to the remaining Betulaceae, i.e. in the subtree comprising all other (more distant) Fagales and the outgroup, may be *trn*L-induced LBA. In contrast to the other alternative relationships favoured by *trn*L (Fig. 6F), where the incompatible *trn*L signal is outperformed by the signal from *mat*K and the other four partitions, the *trnL* signal prevails in this case, because *matK* only provides only a weak to moderate signal for alternative placements. Moreover, the other four partitions, which, in general, contain signal of much lower amplitude than in *matK* and *trnL*, prefer also different alternatives or are indecisive (Figs 6A–E; but see Fig. 7F).

(ii) The placement of Myricaceae is constrained by fact that their 18S rDNA data is highly similar to that of *Rhoiptelea* and the Juglandaceae (ES). This 'zero-branch attraction' outperforms a moderately strong signal from the *matK*, which would favour to place the Myricaceae as sister to the *Casuarina*-*Ticodendron*-Betulaceae clade (Fig. 6C; ES), a signal that prevails in all 5-gene analyses that include *matK* data (Fig. 7; Table 3). The Myricaceae … …

Grades, clades, and our conception of common origin

A minority of systematicists[45-48] put forward a number of arguments, why cladistic systematics should not be equalled with phylogenetic ("evolutionary"[45,47]) systematics. Their basic opinion is that grades in a (molecular) tree that would be interpreted as paraphyla in a Hennigian sense, hence, "invalid" taxa, may be as valid as systematic units as clades, interpreted as monophyla in a Hennigian sense, hence, "valid" taxa. Aside from all good or bad reasons to follow the majority or minority opinion in this matter, it cannot be ignored that the common practise in botanical systematics and phylogenetics is to substantially relax cladistic ideals. Many non-monophyletic taxa are still accepted (in the case of Fagales: subfamilies Betuloideae[1], Engelhardioideae[1,6], and genera *Castanopsis*REF, *Quercus*REF),

barcode studies establish barcodes for taxa without establishing the molecular monophyly first (i.e. providing data to infer a highly supported clade), dating studies rely on numerous fossils as age constraints for molecular clades that were never included in any kind of phylogenetic reconstruction, and so on. The available data and literature on the Fagales provide an excellent example why a strict cladistic system that only recognize (and accepts) a common origin based on the observation of highly supported clades in a phylogenetic tree is simply impractical. Instead a traditional system based on the assumption of common origin and general similarity can be straightforwardly applied backed by molecular data.

Except for the *trnL*-based and the 6-gene tree (Figs 2, 3, 5F; see ref.[2]), which undoubtedly indicate a Betuloideae grade (paraphyletic in a Hennigian[49] or "cladistic" but monophyletic in a Haeckelian or "evolutionary"[45,48,50] sense), all other genetic evidence (Figs 5–7; Table 3; ES; ref.[3]) points towards a common origin of *Alnus* and *Betula* (a monophyletic subfamily Betuloideae, in a Hennigian[49] and Haeckelian[50] sense), and an according 'betuloid clade' can be found and supported in according data-filtered phylogenetic trees (Fig. 7; Table 3; ES; see also ref.[3]) Independent of which evolutionary scenario shown in Figure 1 applies, a "wrong" relationship, hence, a wrong systematic interpretation, would receive substantial support depending on which data have been used. Although it is impossible to distinguish between a paraphyletic or monophyletic (in a Hennigian sense) subfamily Betuloideae, the data is, independent of filtering, rejecting the possibility that the Betuloideae are polyphyletic, i.e. not sharing a (direct) common origin. The data is decisive regarding the question of common origin in general (the basis of pre-Hennigian phylogenetic systematics), but indecisive regarding the question of inclusive and exclusive common origin, which is necessary to the application of cladistic-phylogenetic systematics.

The 18S rDNA not only provides the signal and, hence, the moderate support for a Myricaceae-Juglandaceae clade in the 6-gene tree (incl. *Rhoiptelea*; Figs 2, 3, 5A; ref.[2]; but see Figs 3, 7; Table 3); the same data would also move the Fagales root one node up, and recognize a direct (and inclusive) common origin, a sister relationship, between the *Nothofagus* (the 'Southern' or 'Wrong Beech') and the Fagaceae (including the beech trees). Based on the 18S rDNA data, *Nothofagus* could be re-integrated in the Fagaceae following cladistic principles; or the Fagales could be separated into two "monophyletic" orders: the Fagales s.str and the Juglandales REF. Instead of a Juglandaceae-Myricaceae clade, the remaining data supports a Myricaceae-Betulaceae (incl. *Casuarina* and *Ticodendron*) clade. Thus, one could conclude that both the Juglandaceae-Myricaceae clade and the re-rooted Fagales are branching artefacts inflicted by the 18S rDNA data. This is corroborated by all

analyses so far, which commonly have placed the all-Fagales root between *Nothofagus* and the remainder.[2,7,8,17,51-53] However, the situation here is not that straightforward. The 18S rDNA is the only data set, in which *Nothofagus* is not generally more distinct to all other Fagales than the remaining taxa; it may be possible that the all-Fagales root is only placed between *Nothofagus* and all other Fagales because of LBA between *Nothofagus* and the outtaxa (cf. Fig. 4; note that all based on all available data, the Fagales are much more similar to each other than to any other sequenced angiosperm). Furthermore, the 18S rDNA underlies strong structural constraints at the DNA level causing a (relatively) high sequence conservation, which is the reason that 18S rDNA data is commonly included in analyses not only focussing on deep (or very deep) divergences in angiosperms since the dawn of molecular phylogenetics [51,53-57], but also for other groups of organisms such as protozoans[58,59] and animalsREF. It could theoretically be that the 18S rDNA conserved signals of deep divergences lost in the other partitions; or not captured in the plastid genome at all. In this case, the 6-gene and 5-gene analyses would unavoidably provide tree topologies erroneously rejecting an inclusive common origin (i.e. monophyly s.str.) of *Nothofagus* and the Fagaceae because of a misplaced all-Fagales root. Taken an unrooted all-Fagales tree or distance-based network as basis, a common origin of *Nothofagus* and Fagaceae would be recognized as an alternative to the generally preferred view of *Nothofagus* as sister to all other Fagales. Morphologically, *Nothofagus* provides a better outgroup to reconstruct character evolution in *Fagus* than other Fagaceae[60] (or other Fagales), which could be taken as another evidence for a closer relationship between the two taxa than currently assumed. The problem of changing backbone relationships in the lack of suitable outgroups (because all potential outgroups are already very distant, both in a genetic and phylogenetic sense, to all ingroup taxa) has recently been demonstrated for the group of "basal eudicots" … Matthews vgl. mit Qiu[51]…

Incongruence and incompatibility: Pitfalls and prospects

Given the technical advances in phylogenetic software, it is hard to understand why most multi- or oligogene studies in the field of systematic botany still combine data without a proper assessment of potential incompatibility, which, during reconstruction, may lead to significant incongruence and branching artefacts. In the last issues of *TAXON* and *Molecular Phylogenetics & Evolution,* to take two periodicals with numerous systematic botanical studies relying on few- to many-gene analyses as the major (or only) basis to draw phylogenetic conclusions, 999 out of 999 applied an ILD test, which has been proven to be insufficient[12,13] and 999 didn't applied a test a all. Only 999 of the studies assessed if the

single-gene topologies were significantly incongruent or not; 999 tested for alternative topologies. The example of the putative Myricaceae-Juglandaceae+*Rhoiptelea* sister relationship and the 'betuloid grade' demonstrate the potentially distorting effect of a single partition on a combined analyses. Such an affect can be easily identified by re-analysing subsets of the concatenated matrix, and comparing the topology of the resulting preferred ML trees and the supports of incompatible bipartitions competing to form branches in phylogenetic trees. The most recent versions of RAxML,[28] for instance, include all functions necessary to check for topological incongruence (via SH-test) and to quickly estimate reasonable phylogenetic trees and branch supports (via fast tree-climbing and bootstrapping). In combination with modules implemented in Splits Tree,[9] alternative topologies and competing bipartitions can be easily visualized (Figs 2–6; ES).

In addition, distance-based phylogenetic trees and networks are easy and fast computed using the (Bio)NJ and NN algorithms. If the (Bio)NJ tree largely agrees with the ML tree, and well supported branches thereof, the underlying distance matrix has obviously captured the same phylogenetic signal as the ML tree-inference (and ML bootstrapping), hence, the is no reason to reject the network based on the very same distance matrix to draw further evolutionary conclusions. Based on the NN splits graph, one can intuitively define clusters characterized by high intracluster similarity (Fig. 4; cf. ES), which are distinctly different from the remaining taxa (e.g. the quercoid clade, the Juglandaceae, the Betulaceae). This, by all odds, and in contrast to mainstream systematic beliefs, is a direct evidence for a common origin (cf. Figs 2–4).[23] Furthermore, since the NN splits graph does not distort the distance between two terminals,[22] the graph allows identifying genetically 'primitive' and 'derived' or very distinct taxa. If the graph produces prominent box-like structures that correlate to equally or variably supported topological alternatives additional information is gained regarding the evolutionary unfolding of modern lineages. Pronounced phylogenetic incompatible signals may arise from incomplete lineage sorting, reticulation, heteroachy, or fast ancient radiations.[61,62] The common approach to just discard (collapse) all branches with limited support (e.g. BS < 80, and PP < 0.95), hence, cannot resolve or decide which (if any) scenario applies in a particular case. By combining traditional phylogenetic tree-building and support analyses with distance-based networks, such a distinction is possible.

Another interesting feature of a NN splits graph in an evolutionary context seems to be that, if no outgroup is used and given that the ingroup shares a common origin, the common ancestor, hence, the root, would be located in the centre, or close to the centre of the graph. This was found for non-molecular datasets including (real or putative) ancestors and their

(real or potenatial) descendants,[63,64] a general proof has so far not been tried for molecular data sets. Figure 8 shows some hypothetical examples of phylogenetic trees, the distances between nodes, and subsequently terminals, and the resulting phylogenetic network based on the according distance matrices. Although this is not a proof, the results are encouraging: only in the case of strong heteroachy, the root (as indicated by the position of the common ancestor) is not close the centre of the graph. Given this results, the 18S rDNA-indicated root, rejected by the combined data, seems to be not unplausible.

Not used so far

In Figure**Error! Reference source not found.** the uncorrected p-distance is plotted against the z-axis of the diagram for each pair of taxa and expanded to a surface area. The taxa have been grouped according to their systematic position along the x- and y-axis. Such groups that form well supported clades appear as depressions in the surface; outgroups and distantly related taxa pairs are forming ridges and heights. Even deeper phylogenetic relationships are readily visible (e.g. the core higher hamamelids or the proposed sisterclade relationship between Myricaceae and the *Rhoiptelea*-Juglandaceae clade). Lowest distances to non-sister taxa and outgroups are found in Myricaceae and *Fagus*, highest distances characterize *Rhoiptelea, Casuarina,* and *Ticodendron* (Figs**Error! Reference source not found.**, **Error! Reference source not found.**); the correlation to features of the NN splits graph, the reconstructed phylogenetic trees and the bipartition networks are imminent.

Roots and cladistic interpretation of subtrees

The topological congruence between the NJ trees and the phylogenetic synopsis provided in [2] demonstrate that the phylogenetic process one tries to model using ML and MP can be reduced to a simple natural phenomenon: significantly higher similarity of gene sequences among closest relatives (“nearest neighbours”) and their dissimilarity to distantly related taxa. It also demonstrates that in this case ML, and even less MP, can catch any further phylogenetically signal not already comprised in the pairwise distance distribution. For example, *Alnus* and *Betula* are more distant to each other and basically equally distant to the core Betulaceae node (*Carpinus, Carpinopsis, Ostrya, Ostryopsis*), accordingly incongruent topologies are found, and *Alnus* and *Betula* are not recognized as sister taxa, instead they are placed as a grade in relation to the most terminal and undisputed (comprising the most similar

taxa) clade. Due to the *trn*L-F data, *Betula* is less distant than *Alnus* to the core Betulaceae, and this forces *Alnus* as sister to all other Betulaceae. ITS (and 5S IGS) data of *Alnus* and *Betula* comprises far enough evidence to show that *Alnus* is the closest living relative of *Betula*, which is in agreement to alternatives proposed based on most included gene regions. Why they are not recognized as sister taxa by the combined multi-gene data set? Because of the Betulaceae-root that is placed due to the inclusion of *Ticodendron* (and subsequently all other Fagales). *Ticodendron* is obviously the closest living relative of the Betulaceae, it is less

휀 Ò 쨀 Ê 숀 ½ 렀 ,

found to *Alnus, Betula,* and the basic node of the core Betulaceae. Since *Alnus* and *Betula* are recognized as Betulaceae, *Ticondendron* can only be placed as sister to all Betulaceae. The relatively high divergence between *Ticodendron* and all Betulaceae, the comparably increased divergence of *Alnus* and *Betula* to the other Betulaceae (as also reflected in the ITS), and the generally low divergence among the other Betulaceae result in a 'short-branch attraction' (connecting *Betula* closer to the core Betulaceae), or LBA between some 'ingroup' taxa (i.e. *Alnus, Betula*) and the Betulaceae's 'outgroup' (i.e. *Ticodendron*). The most parsimonious (and likeliest) solution is to place *Ticodendron* as 'sister taxon' to the most distinct Betulaceae, namely *Alnus* because of the *trn*L-F data. Placing *Ticondendron* as sister to *Betula* would be less parsimonious and probable because *Betula* is closer related to *Alnus* or the other *Betulaceae*, a placement among the other Betulaceae is obviously least parsimonious and unlikely. As consequence, the inferred Betulaceae 'root' is a reconstruction artefact, due to *trn*L-F-induced LBA between *Alnus* and any non-Betulaceae, and accordingly, the phylogenetic interpretation of the Betulaceae subtree as given erroneous. A wrongly inferred subtree root as in the case of *Ticodendron* and the Betulaceae, could also apply to all other internal roots, and in particular, the all-Fagales-root inferred by inclusion of *Celtis* and *Hamamelis*. Alternatively placed all-Fagales-roots do not affect the support of any other bipartition during phylogenetic analyses, but the phylogenetic interpretation thereof. For example, if one assumes that the most conserved gene region included (18S rDNA) infers the best possible root, *Nothofagus* must be interpreted as sister lineage to the Fagaceae, and not, as sister lineage to all other Fagales. The current *Nothofagus*-Fagaceae grade in the multigene tree, that would render an according group 'paraphyletic', may well be a slightly mis-reconstructed *Nothofagus*-Fagaceae clade, which would be interpreted as monophyletic.

Conclusion

Taxa that share significantly similar sequences in six gene regions and that are substantially different to other sampled taxa in these gene regions, are without a doubt closest modern relatives, hence, they are likely to share a common origin, but whether they are monophyletic or sister taxa in a strict, cladistic sense is hard to decide. As shown, if one soley relies on a phylogenetic tree based on multigene data (of unknown compatibility), we might simply visualize sequence similarity or dissimilarity (e.g. distinct *Alnus-trnL*; highly similar 18S rDNA of Myricaceae and Juglandaceae) but not necessarily model the exact evolutionary pathways, i.e. the actual phylogeny. Grades *and* clades in a molecular-based phylogenetic tree may be representations of paraphyla *and* monophyla in a strict, Hennig'ian sense, which makes it difficult to rely on the latter as the only "valid" taxonomic group. In this context, it is minor importance whether trees are based on distances, on likeliness and a substitution model, or on character changes as under parsimony, if the data set is divergent and significant enough (independent of the number of different gene regions or base pairs included). It has to be further tested, if rather simple and fast algorithms such as NJ and NN consistently recognize groups of putative common origin, i.e. monophyly in a general (non-cladistic) sense, based on multi-gene data as it is the case in the Fagales. Without a doubt, the opportunity of any network is that second-best or third-best alternatives to group taxa or topological alternatives of trees are not lost, but included in the reconstruction of splits graphs (NN, bipartition) or consensus networks. Furthermore, topological tests and data significance analyses are crucial to avoid that a single gene forces relationships in conflict with the remaining data.

Moreover, one may want to concretize the phylogenetic information 'behind the graph,' in particular such branches that differ among methods and data sets used, and are not recovered the same way by network approaches. As shown, it may be more prospective to infer the root of a tree (or any subtree) by ingroup comparison rather than to completely rely on outgroups or putative sister taxa. In particular, one should take into account not only the support of each branch but also the branch lengths' diversity in phylograms and distance-based split networks as an additional source of phylogenetic information.

Figures and Tables

Table 1 Support at the family level from the concatenated and single-gene data.

Table 2 Results of the KH and SH test. Pearson correlation coefficients of bootstrap frequencies are given for comparison.

Table 3 Results of the 5-gene ML analyses, excluding one of the six original partitions..

Table 4 Variable sites in the 5.8S rDNA of Fagales and *Hamamelis*

Figure 1 Depending on where the root is placed, hence, to which branch the *Ticodendron* as the outtaxon is connected, the subfamily Betuloideae (*Alnus* + *Betula*) forms a grade or a clade (interpreted as evidence for paraphyly and monophyly according Hennig's concept), although the phylogenetic relationships within the ingroup remain unchanged.

Figure 2 Best-known ML tree based on the concatenated data, substitution rates were independently optimized for each of the six gene partitions. Numbers along branches refer to support based on nonparametric bootstrapping under ML (not partitioned, using six partitions) and MP and Bayesian inference (PP).

Figure 3 Bipartition network based on the ML bootstrap sample (each partition independently optimized). Edge lengths are proportional to the number of bootstrap replicate trees showing the according bipartition, i.e. reflects the bootstrap support of the according edge bundle (branch in phylogenetic trees).

Figure 4 Distance-based phylogenetic analyses based on simple (uncorrected p) Hamming distances. **A**, Unrooted NJ phylogram. The tree has the same topology as best-known ML tree in Fig. 1, except that *Platycarya* instead of *Carya*+*Annamocarya* is placed as sister to the remaining core Juglandaceae (Juglandoidea). **B,** Neighbour-net splits graphs based on uncorrected pairwise distances. Phylogenetically ambiguous taxa are connected to the graph involving box-like structures; tree-like portions straightforwardly identify unambiguous clades that received high support (cf. Figs 1, 2). Some taxa (*Fagus, Rhoiptelea,* and the Myricaceae) are closer to the centre of the graph, hence, closer to the putative roots (genetically less derived; discussed in the text).

Figure 5 Three-dimensional area graph of the distance distribution. 'Mountains' and 'ridges' indicate relatively high pairwise distances, 'gorges' and 'valleys' low pairwise distances. Note that all families recognized by APG comprising more than one genus, except for the Fagaceae (Fagus is relatively distinct from the remaining genera), are characterized by low distances to members coupled with markedly higher distances to non-members. The same

holds for the members of two identified suprafamily clades (Fig. 2; cf. ref.[2]), the Juglandaceae-*Rhoiptelea*-Myricaceae and the 'core higher hamamelid' clade.

Figure 6 Consensus network of all MPT based on all six single-gene matrices. Box-like portions indicate topological incongruence between and among the MPT computed, thickened lines correlate to well-supported branches in Fig. 1. Only splits that occurred in 10% or more of MPT are shown, edge lengths are mean branch lengths of the MPT. Equally long incongruent edges define the position of Myricaceae among the core higher hamamelids: the Myricaceae can be placed as sister clade to all other core higher hamamelids (red), to the *Rhoiptelea*-Juglandaceae clade (green), or to the *Casurina*-*Ticodendron*-Betulaceae clade (blue). Note that only the latter two receive measurable support from nonparametric bootstrapping and Bayesian analyses.

Figure 7 ML trees based on single-gene matrices; numbers at branches indicate bootstrap support and posterior probabilities, respectively (BS_{ML}/ BS_P/ PP). **A,** 18S nrDNA data: the outgroup-inferred ingroup root moves, with the result that *Nothofagus* is recognized as sister to the Fagaceae and not all Fagales. **B,** *atp*B data. **C,** *matK* data **D,** *matR* data (mtDNA). **E,** *rbcL* data. **F,** *trnL* data.

Figure 8. ML trees based on five-gene matrices; numbers at branches indicate bootstrap support and posterior probabilities, respectively (BS_{ML}/ BS_P/ PP). **A,** 18S nrDNA data excluded: Myricaceae are supported as sister to Betulaceae and relatives. **B,** *atp*B data excluded. **C,** *matK* data excluded. **D,** *matR* data excluded. **E,** *rbcL* data excluded. **F,** *trnL* data excluded.

Figure 9. Secondary structure model of the 5.8S rDNA/5.8S rRNA[44] summarizing the data of 999 ITS sequences available for members of the Fagales and *Hamamelis*. Sites exhibiting variation (and type of variation) are highlighted.

Figure xx Bipartition networks based on the 18S rDNA and *atp*B data sets. Edge lengths are defined by the 'weight' of each split; here: the PP computed based on the non-burned Bayesian inferred saved trees (BIST). Edge bundles of alternative phylogenetic relationships considering the position of the Myricacea are coloured. A, BP network based on 18S rDNA data, only splits are shown that occurred in ≥15% of 49,824 BIST. B, BP network based on *atp*B data and 47,230 BIST.

1 Stevens, P. F. Vol. 2008 (2001 onwards).
2 Li, R.-Q. *et al.* Phylogenetic relationships in Fagales based on DNA sequences from three genomes. *International Journal of Plant Science* **165**, 311-324 (2004).
3 Forest, F. *et al.* Teasing apart molecular- versus fossil-based error estimates when dating phylogenetic trees: a case study in the birch family (Betulaceae). *Syst. Bot.* **30**, 118-133 (2005).
4 Manos, P. S., Zhou, Z. K. & Cannon, C. H. Systematics of Fagaceae: Phylogenetic tests of reproductive trait evolution. *International Journal of Plant Science* **162**, 1361-1379 (2001).
5 Denk, T. & Grimm, G. W. The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. *Taxon* **59**, 351-366 (2010).
6 Manos, P. S. *et al.* Phylogeny of extant and fossil Juglandaceae inferred from the integration of molecular and morphological data sets. *Syst. Biol.* **56**, 412-430 (2007).
7 Hilu, K. W. *et al.* Angiosperm phylogeny based on *matK* sequence information. *Am. J. Bot.* **90**, 1758-1776 (2003).
8 Savolainen, V. *et al.* Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* **49**, 306-362 (2000).
9 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254-267 (2006).
10 Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of live. *Nat. Rev. Genet.* **6**, 361-375 (2005).
11 Farris, J. C., Källersjö, M., Kluge, A. G. & Bult, C. Testing significance of incongruence. *Cladistics* **10**, 315-319 (1994).
12 Dowton, M. & Austin, A. D. Increased congruence does not necessarily indicate increased phylogenetic accuracy - The behavior of the incongruence length difference test in mixed-model analyses. *Syst. Biol.* **51**, 19-31 (2002).
13 Hipp, A. L., Hall, J. C. & Sytsma, K. J. Congruence versus phylogenetic accuracy: Revisiting the Incongruence Length Difference test. *Syst. Biol.* **53**, 81-89 (2004).
14 Czarna, A., Sanjuán, R., González-Candelas, F. & Wróbel, B. Topology testing of phylogenies using least squares methods. *BMC Evol. Biol.* **6**, 105 (2006).
15 Shavit, L., Penny, D., Hendy, M. D. & Holland, B. R. The problem of rooting rapid radiations. *Mol. Biol. Evol.* **24**, 2400-2411, doi:10.1093/molbev/msm178 (2007).
16 Graham, S. W., Olmstead, R. G. & Barrett, S. C. H. Rooting phylogenetic trees with distant outgroups: A case study from the Commelinoid monocots. *Mol. Biol. Evol.* **19**, 1769-1781 (2002).
17 Stamatakis, A., Göker, M. & Grimm, G. W. Maximum likelihood analysis of 3,490 rbcL sequences: Scalability of comprehensive inference versus group-specific taxon sampling. *Evol. Bioinf.* **6**, 73-90 (2010).
18 Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114-1116 (1999).
19 Holland, B. & Moulton, V. in *Algorithms in Bioinformatics: Third International Workshop, WABI, Budapest, Hungary. Proceedings* Vol. 2812 *Lecture Notes in Bioinformatics (LNBI)* (eds G. Benson & R. Page) 165-176 (Springer Verlag, 2003).
20 Grimm, G. W., Renner, S. S., Stamatakis, A. & Hemleben, V. A nuclear ribosomal DNA phylogeny of *Acer* inferred with maximum likelihood, splits graphs, and motif analyses of 606 sequences. *Evol. Bioinf.* **2**, 279-294 (2006).
21 Bryant, D. & Moulton, V. in *Algorithms in Bioinformatics, Second International Workshop, WABI* Vol. 2452 *Lecture Notes in Computer Science* (eds R. Guigó & D. Gusfield) 375-391 (Springer Verlag, Berlin, Heidelberg, New York, 2002).

22 Bryant, D. & Moulton, V. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255-265 (2004).
23 Felsenstein, J. *Inferring phylogenies.*, 664 (Sinauer Associates Inc., 2004).
24 PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta (Sinauer Associates, Sunderland, MA, 2002).
25 Stamatakis, A. RAxML-VI-HPC: Maximum-Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).
26 Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791 (1985).
27 Müller, K. F. The efficiency of different search strategies for estimating parsimony, jackknife, bootstrap, and Bremer support. *BMC Evol. Biol.* **5**, 58 (2005).
28 Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **57**, 758-771 (2008).
29 Stamatakis, A. in *Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006), High Performance Computational Biology Workshop.*
30 Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* **43**, 304-311 (1996).
31 Pattengale, N. D., Masoud, A., Bininda-Emonds, O. R. P., Moret, B. M. E. & Stamatakis, A. in *RECOMB 2009.* Vol. 5541 *Lecture Notes in Computer Science* (ed S. Batzoglou) 184-200 (Springer-Verlag, 2009).
32 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).
33 Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685-695 (1997).
34 Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).
35 Rzhetsky, A. & Nei, M. A simple method for estimating and testing minimum evolution trees. *Mol. Biol. Evol.* **9**, 945-967 (1992).
36 Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and branching order in Hominoidea. *Journal of Molecular Evolution* **29**, 170-179 (1989).
37 Herbert, J., Chase, M. W. & Moller, M. Nuclear and plastid DNA sequences confirm the placement of the enigmatic *Canacomyrica moticola* in Myricaceae. (???).
38 PAUP*: Phylogenetic analysis using parsimony (* and other methods) v. 4.0 (Illinois Natural History Survey, Champaign, IL, 1998).
39 Hershkovitz, M. A. & Lewis, L. A. Deep-level diagnostic value of the rDNA-ITS region. *Mol. Biol. Evol.* **13**, 1276-1295 (1996).
40 Hershkovitz, M. A. & Zimmer, E. A. Conservation patterns in angiosperm rDNA ITS2 sequences. *Nucleic Acids Res.* **24**, 2857-2867 (1996).
41 Hershkovitz, M. A., Zimmer, E. A. & Hahn, W. J. in *Molecular Systematics and Plant Evolution* (eds P. M. Hollingsworth, R. M. Bateman, & R. J. Gornall) 268-326 (Taylor & Francis, 1999).
42 Mai, J. C. & Coleman, A. W. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *Journal of Molecular Evolution* **44**, 258-271 (1997).
43 Coleman, A. W. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics* **19**, 370-375 (2003).

44 Cannone, J. J. *et al.* The Comparative RNA Web (CRW) Site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinf.* **3**, 2, 15 (Erratum) (2002).
45 Hörandl, E. Paraphyletic versus monophyletic taxa - evolutionary versus cladistic classifications. *Taxon* **55**, 564-570 (2006).
46 Hörandl, E. Neglecting evolution is bad taxonomy. *Taxon* **56**, 1-5 (2007).
47 Mayr, E. & Bock, W. J. Classifications and other ordering systems. *J. Zool. Syst. Evol. Research* **40**, 169-194 (2002).
48 Zander, R. H. Evolutionary inferences from non-monophyly on molecular trees. *Taxon* **57**, 1182-1188 (2008).
49 Hennig, W. *Grundzüge einer Theorie der phylogenetischen Systematik*. 370 (Dt. Zentralverlag, 1950).
50 Haeckel, E. *Generelle Morphologie der Organismen*. (Georg Reiner, 1866).
51 Qiu, Y.-L. *et al.* Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. *International Journal of Plant Science* **166**, 815-842 (2005).
52 Soltis, D. E., Gitzendanner, M. A. & Soltis, P. S. A 567-taxon data set for angiosperms: The challenges posed by Bayesian analyses of large data sets. *International Journal of Plant Science* **168**, 137-157 (2007).
53 Soltis, D. E. *et al.* Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* **133**, 381-461 (2000).
54 Kim, S., Soltis, D. E., Soltis, P. S., Zanis, M. J. & Suh, Y. Phylogenetic relationships among early-diverging eudicots based on four genes: were the eudicots ancestrally woody? *Mol. Phylogenet. Evol.* **31**, 16-30 (2004).
55 Soltis, D. E. *et al.* Gunnerales are sister to other core eudicots: Implications for the evolution of pentamery. *Am. J. Bot.* **90**, 461-470 (2003).
56 Cuénoud, P. *et al.* Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *Am. J. Bot.* **89**, 132-144 (2002).
57 Doyle, J. A., ??? & ??? Intergration of morphological and ribosomal RNA data on the origin of angiosperms. *Ann. MO. Bot. Garden* **81**, 419-450 (1994).
58 Berney, C. & Pawlowski, J. Revised small subunit rRNA analysis provides further evidence that Foraminifera are related to Cercozoa. *Journal of Molecular Evolution* **57**, S120-S127 (2003).
59 Bolivar, I., Fahrni, J. F., Smirnov, A. & Pawlowski, J. SSU rRNA-based phylogenetic position of the genera *Amoeba* and *Chaos* (Lobosea, Gymnamoebia): the origin of Gymnamoebae revisited. *Mol. Biol. Evol.* **18**, 2306-2314 (2001).
60 Denk, T. Phylogeny of *Fagus* L. (Fagaceae) based on morphological data. *Plant Syst. Evol.* **240**, 55-81 (2003).
61 Lockhart, P. *et al.* Heteroachy and tree building: A case study with plastids and eubacteria. *Mol. Biol. Evol.* **23**, 40-45 (2006).
62 Whitfield, J. B. & Lockhart, P. J. Deciphering ancient rapid radiations. *Trends Ecol. Evol.* **22**, 258-265 (2007).
63 Denk, T. & Grimm, G. W. The biogeographic history of beech trees. *Rev. Palaeobot. Palynol.* **158**, 83-100 (2009).
64 Spencer, M., Davidson, E. A., Barbrook, A. C. & Howe, C. J. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology* **227**, 503-511 (2004).

Table 1

	Six genes		One gene					
	n.p.	p.	18S	*atp* B	*mat* K	*mat* R	*rbc* L	*trn* L
	BS_{ML}		BS_{ML}	BS_{ML}	BS_{ML}	BS_{ML}	BS_{ML}	BS_{ML}
Fagaceae	100	100	99	100	100	100	92	100
Myricaceae	100	100	100	100	99	97	100	100
Juglandaceae	100	100	64	100	100	92	100	100
Betulaceae	100	100	71	92	100	26	25	98

	Matrix						
Topology	18S	atpB	matK	matR	rbcL	trnL	Concatenated data
18S-favoured	Best	Worse/worse	Worse/worse	Worse/worse	Worse/worse	Worse/worse	Is rejected by KH and SH test
atpB-favoured	Worse*/**fit**	Best	Worse/**fit**	Worse/**fit**	**fit/fit**	Worse/worse	With a probability of 0.2 (KH) or 0.7 (SH) as good as the best topology
matK-favoured	Worse/**fit**	**fit/fit**	Best	Worse/**fit**	**fit/fit**	Worse/worse	Best topology
matR-favoured	Worse/worse	Worse/worse	Worse/worse	Best	Worse/worse	Worse/worse	Is rejected by KH and SH test
rbcL-favoured	Worse/worse	Worse/worse	Worse/worse	Worse/worse	Best	Worse/worse	Is rejected by KH and SH test
trnL-favoured	Worse/**fit**	Worse/worse	Worse/worse	Worse/**fit**	Worse/worse	Best	Is rejected by KH and SH test
Correlation (R^2)	0.594987	0.732028	0.75687	0.617928	0.571896	0.662037	→
between single-gene BS and 5-gene BS analyses excluding the according partition							

"Worse" indicates that the input topology is significantly less optimal ($p < 0.05$) than the native topology of the according data matrix

"Fit" indicates that with this input topology the null-hypotheses (a congruent topology is optimal) has a probability > 0.05

0.963638	6x vs. 5x excl. 18S
0.991796	6x vs. 5x excl. atpB
0.922536	6x vs. 5x excl. matK
0.986968	6x vs. 5x excl. matR
0.98642	6x vs. 5x excl. rbcL
0.892622	6x vs. 5x excl. trnL

Table 2

Node	Li & al., 2004		This study, 6 genes				Five genes, excluding											
							18S		*atp* B		*mat* K		*mat* R		*rbc* L		*trn* L	
	BS_P	PP	ML tree		BS_{ML}		BS_{ML}		BS_{ML}		BS_{ML}		BS_{ML}		BS_{ML}		BS_{ML}	
			n.p.	p.	n.p.	p.	n.p.	p.	n.p.	p.	n.p.	p.	n.p.	p.	n.p.	p.	n.p.	p.
[1] Ingroup (all-Fagales) root	100	1.00	Yes	Yes	100	100	100	100	100	100	100	100	100	100	100	100	100	100
[2] *Nothofagus* first diverging branch	100	1.00	Yes	Yes	100	100	100	100	100	100	100	100	100	100	100	100	100	100
[3] Quercoid clade	100	1.00	Yes	Yes	100	100	87	100	100	100	100	100	100	100	100	100	100	100
[4] Clade comprsing *Lithocarpus, Quercus, Castanopsis, Castanea*	70	<0.95	**No**	Yes	**57**	**58**	INC	INC	INC	INC	78	74	**57**	**INC**	54	65	INC	INC
[4a] Clade comprising *Trigonobalanus, Quercus, Castanopsis, Castanea*	N/A	N/A	Yes	No	39	39	59	54	57	41	INC	INC	**INC**	**34**	INC	INC	49	60
[5] *Castanea* + *Castanopsis*	67	0.97	Yes	Yes	71	68	87	85	72	51	INC	INC	**INC**	**49**	76	68	76	78
[5a] *Castanopsis* + *Lithocarpus*	N/A	N/A	No	No	25	31	INC	INC	INC	INC	58	50	**60**	**INC**	INC	INC	INC	INC
[6] Core higher hamamelid clade	100	1.00	Yes	Yes	100	100	100	100	100	100	100	100	100	100	100	100	100	100
[7] *Casuarina-Ticodendron*- Betulaceae clade	99	1.00	Yes	Yes	100	100	98	95	99	100	100	100	100	100	100	100	99	100
[8] *Ticodendron*- Betulaceae clade	99	1.00	Yes	Yes	100	99	100	100	98	99	99	98	99	98	100	100	99	100
[9] *Betula-Carpinus-Ostrya-Corylus-Ostryopsis* clade	100	1.00	Yes	Yes	100	100	100	100	100	100	100	100	100	100	100	100	INC	INC
[9a] *Alnus* + *Betula* (Betuloideae)	N/A	N/A	No	No	0	0	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC	**60**	**57**
[10] *Carpinus-Ostrya-Corylus-Ostryopsis* clade	100	1.00	Yes	Yes	100	100	100	100	100	100	99	100	100	100	100	100	100	100
[11] *Carpinus* + *Ostrya*	99	1.00	Yes	Yes	100	100	99	100	97	99	97	99	100	97	99	99	100	100
[12] *Corylus* + *Ostryopsis*	99	1.00	Yes	Yes	95	95	98	94	90	94	INC	INC	98	100	100	100	87	88
[12a] *Corylus* sister to *Carpinus* + *Ostryopsis*	N/A	N/A	No	No	5	<5	INC	INC	INC	INC	**72**	**76**	INC	INC	INC	INC	INC	INC
[13] Myricaceae-*Rhoiptelea*- Juglandaceae clade	63	0.95	Yes	Yes	74	62	INC	INC	66	66	87	87	77	64	78	67	81	77
[13a] Myricaceae-*Casuarina-Ticodendron*- Betulaceae clade	N/A	N/A	No	No	26	38	**82**	**85**	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC
[14] *Rhoiptelea*- Juglandaceae clade	100	1.00	Yes	Yes	100	100	100	100	100	100	100	100	100	100	100	100	100	100
[15] Juglandoidea clade*	100	0.83	Yes	Yes	100	100	100	100	100	100	100	99	100	100	100	100	100	100
[16] *Carya-Juglans-Cyclocarya-Pterocarya* clade	60	0.97	**No**	**No**	<5	<5	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC	**30**	**INC**
[17] *Carya* + *Juglans*	77	1.00	**No**	**No**	<5	<5	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC	INC
[16a] *Juglans-Platycarya-Cyclocarya-Pterocarya* clade	N/A	N/A	Yes	Yes	84	79	79	75	94	86	75	70	70	71	97	85	INC	INC
[17b] *Juglans-Cyclocarya-Pterocarya* clade	N/A	N/A	Yes	Yes	**67**	**66**	69	67	83	74	65	54	87	86	INC	INC	55	56
[17c] *Platycarya-Cyclocarya-Pterocarya* clade	N/A	N/A	No	No	19	**18**	INC	INC	INC	INC	INC	INC	INC	INC		37	INC	INC
[17a] *Carya* + *Annomocarya*	N/A	N/A	Yes	Yes	92	91	88	89	93	88	75	76	80	78	98	98	**INC**	**45**

* Including *Carya, Annomocarya, Juglans, Cyclocarya, Pterocarya,* and *Platycarya.* Second subfamily Engelhardioideae only represented by *Engelhardia*

Table 3 (red: instable relatioships; green: consistently recovered

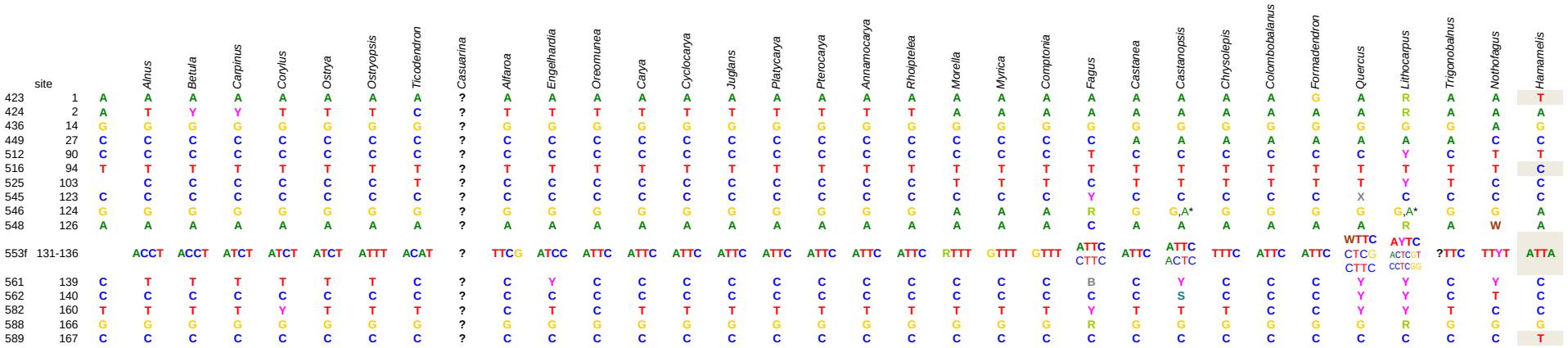
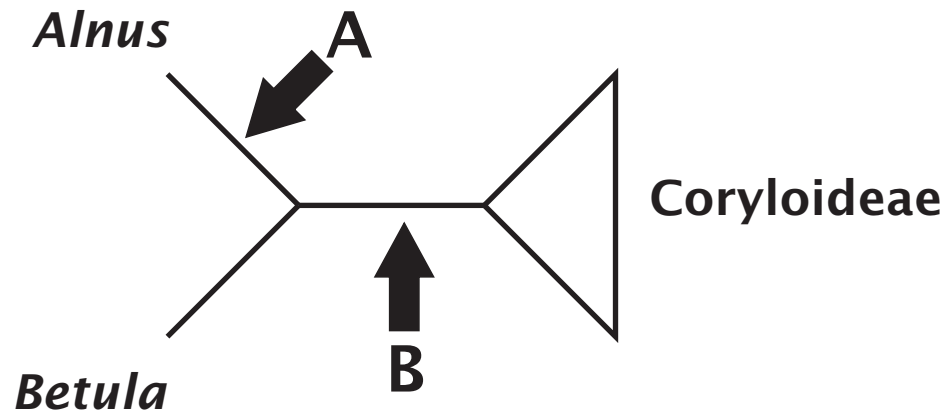
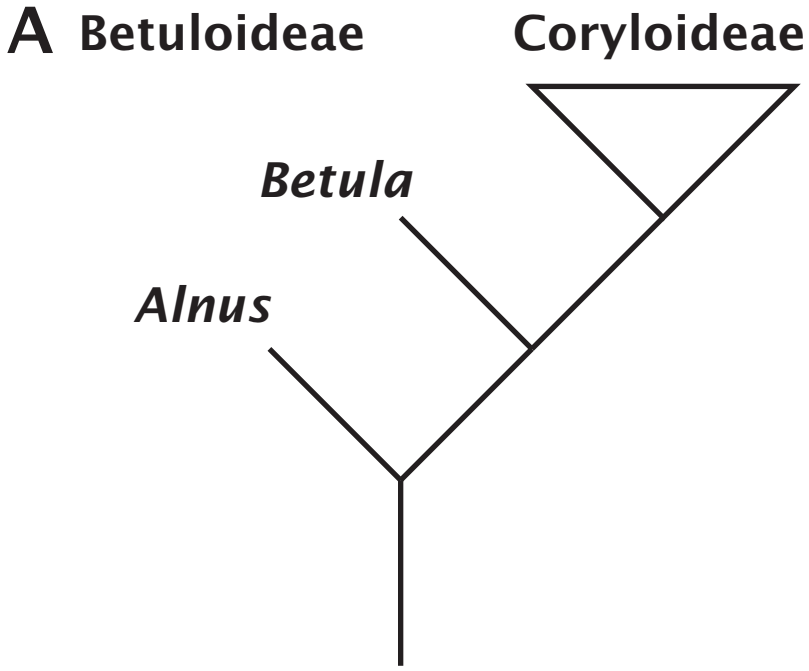

site
Alnus
Betula
Carpinus
Corylus
Ostrya
Ostryopsis
Ticodendron
Casuarina
Alfaroa
Engelhardia
Oreomunea
Carya
Cyclocarya
Juglans
Platycarya
Pterocarya
Annamocarya
Rhoiptelea
Morella
Myrica
Comptonia
Fagus
Castanea
Castanopsis
Chrysolepis
Colombobalanus
Formadendron
Quercus
Lithocarpus
Trigonobalnus
Nothofagus
Hamamelis
423 1
424 2
436 14
449 27
512 90
516 94
525 103
545 123
546 124
548 126
553f 131-136
561 139
562 140
582 160
588 166
589 167

Table 4

Figure 1


Alnus
A
Coryloideae
B
Betula


A
Betuloideae
Coryloideae
Betula
Alnus

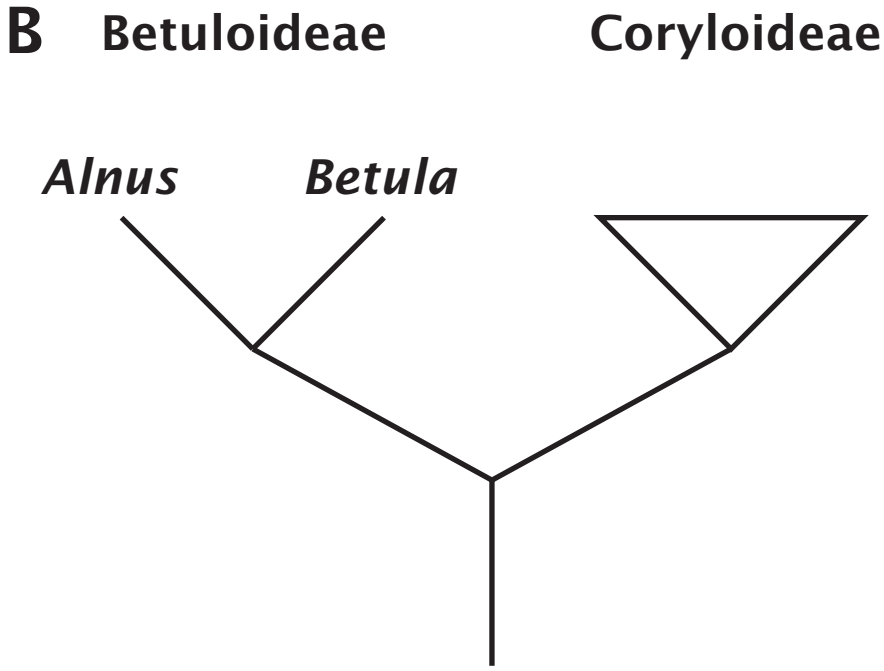
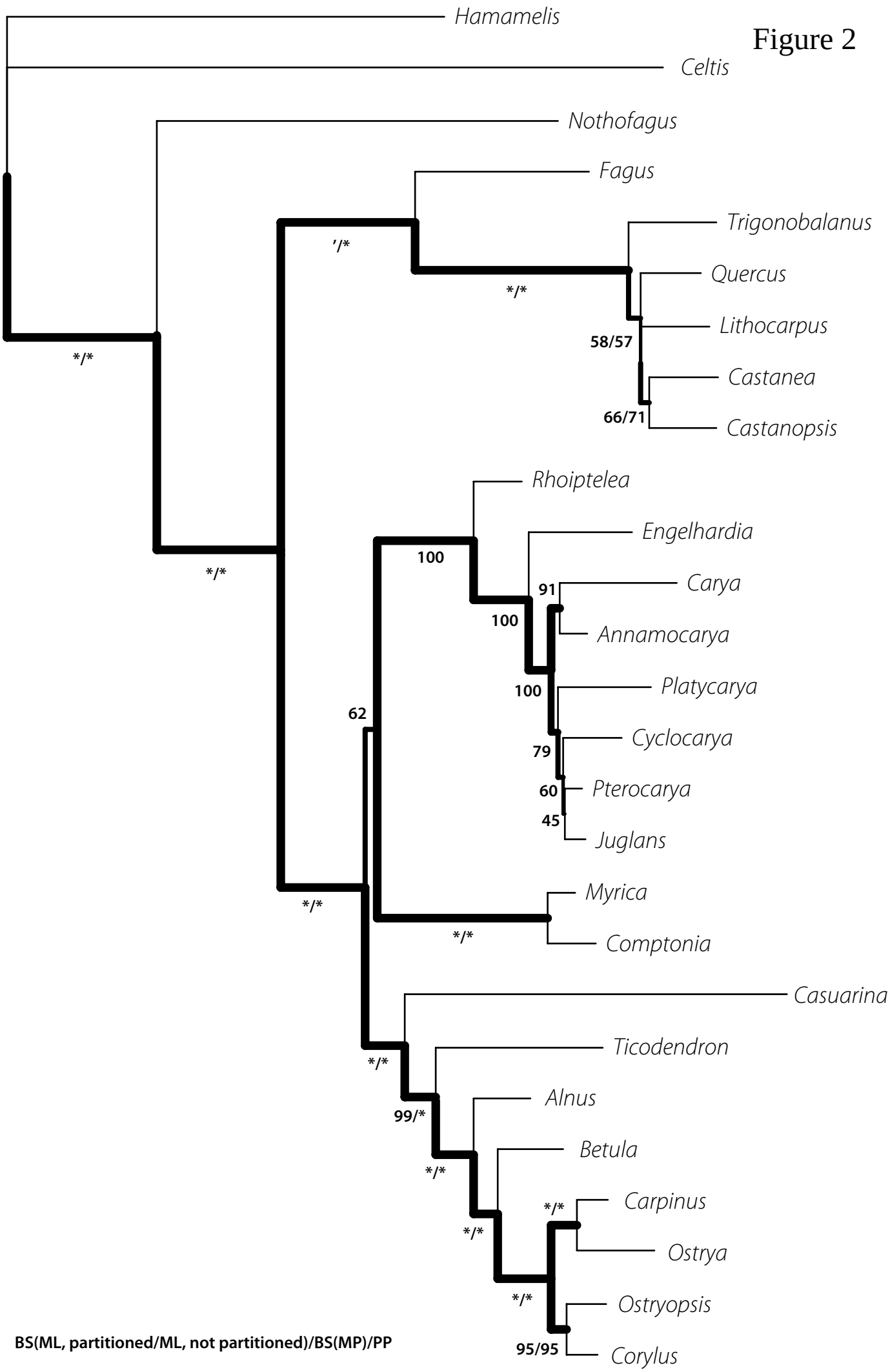

B
Betuloideae
Coryloideae
Alnus
Betula

Figure 2



BS(ML, partitioned/ML, not partitioned)/BS(MP)/PP

Figure 3

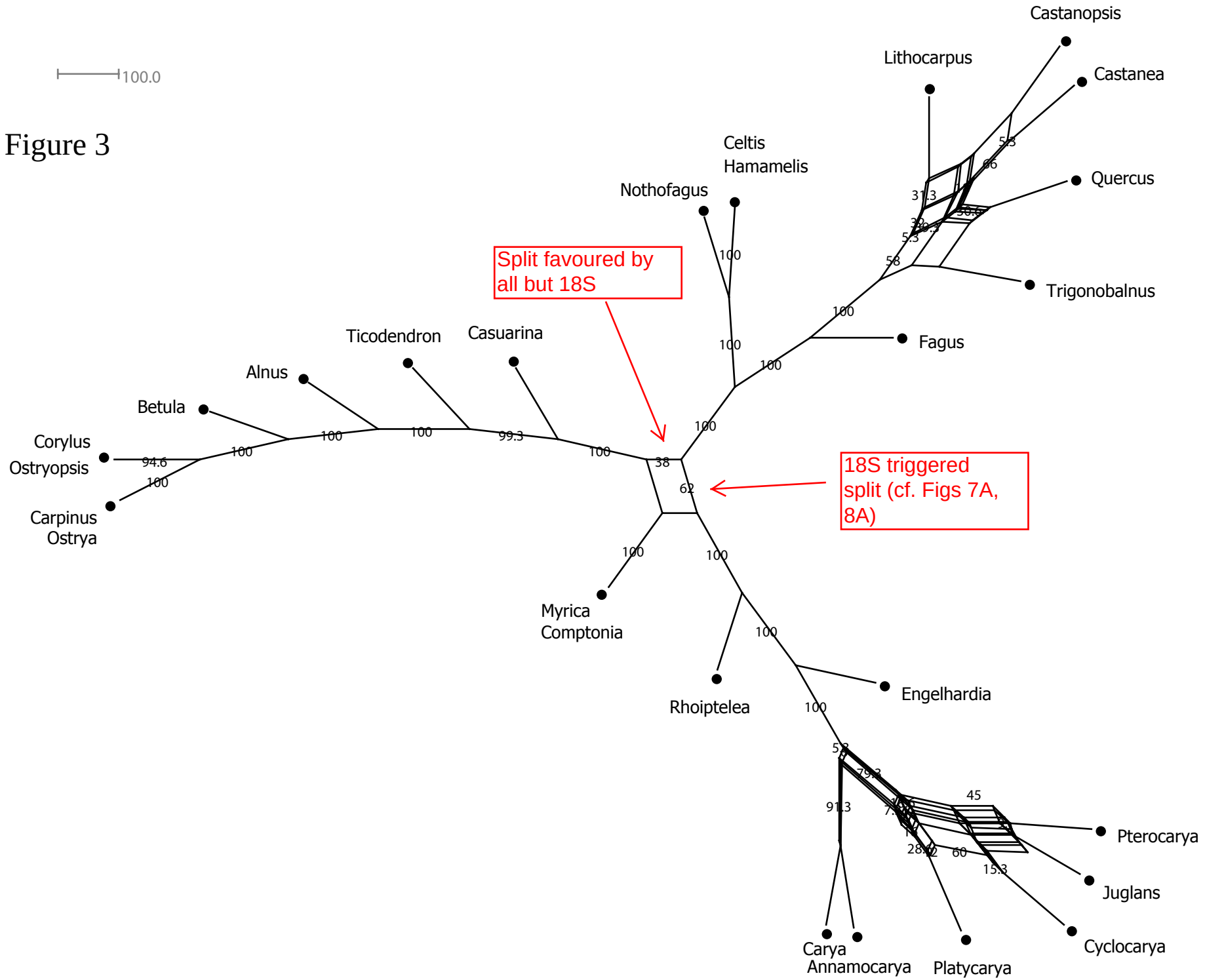

100.0
Castanopsis
Lithocarpus
Castanea
Quercus
Trigonobalnus
Fagus
Celtis
Hamamelis
Nothofagus
Split favoured by all but 18S
Ticodendron
Casuarina
Alnus
Betula
Corylus
Ostryopsis
Carpinus
Ostrya
18S triggered split (cf. Figs 7A, 8A)
Myrica
Comptonia
Rhoiptelea
Engelhardia
Pterocarya
Juglans
Cyclocarya
Platycarya
Carya
Annamocarya

Figure 4a

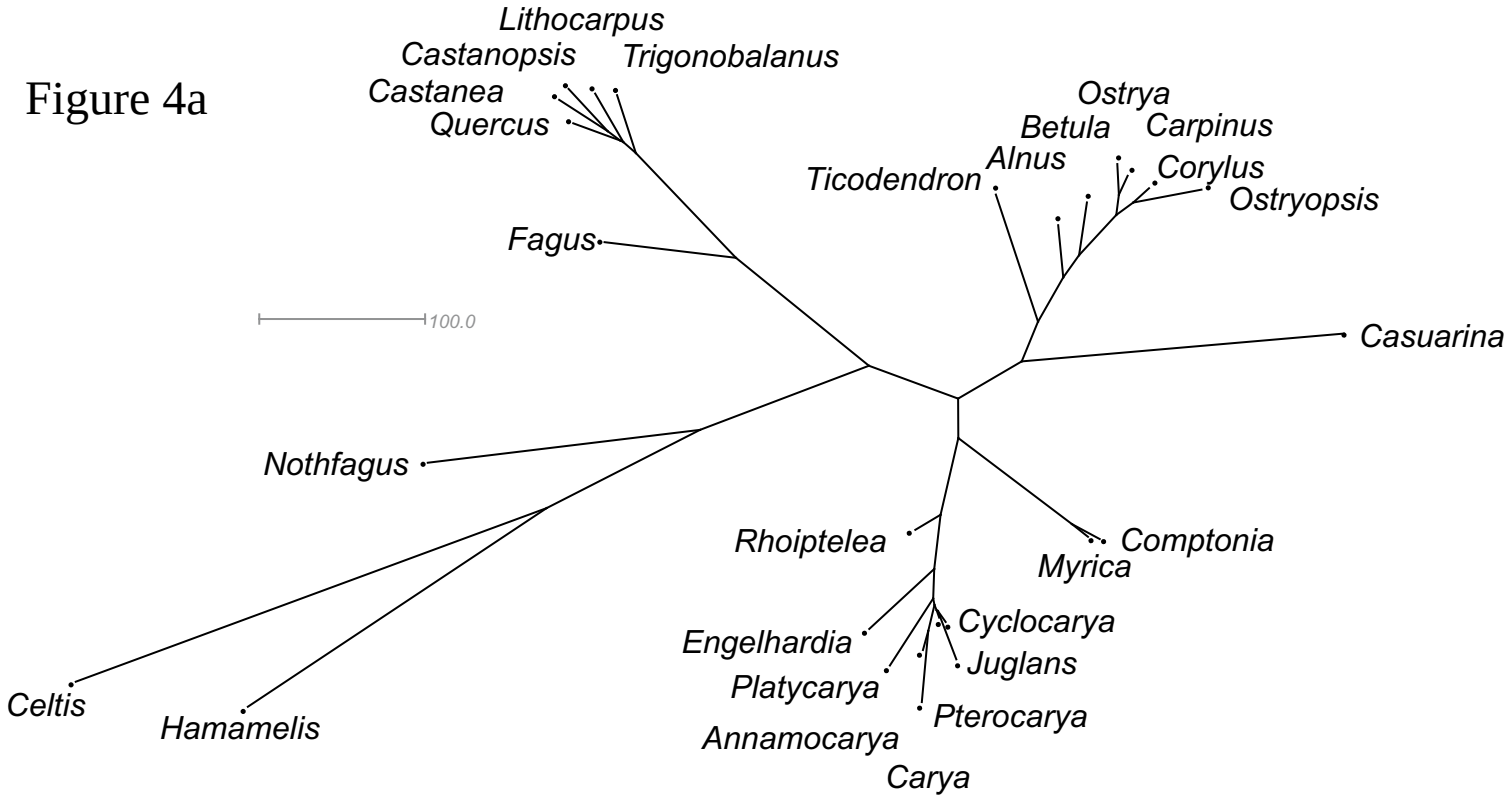

Lithocarpus
Castanopsis
Trigonobalanus
Castanea
Quercus
Ostrya
Betula
Carpinus
Alnus
Corylus
Ticodendron
Ostryopsis
Fagus
100.0
Casuarina
Nothfagus
Rhoiptelea
Comptonia
Myrica
Cyclocarya
Engelhardia
Juglans
Platycarya
Pterocarya
Annamocarya
Carya
Celtis
Hamamelis

Figure 4B

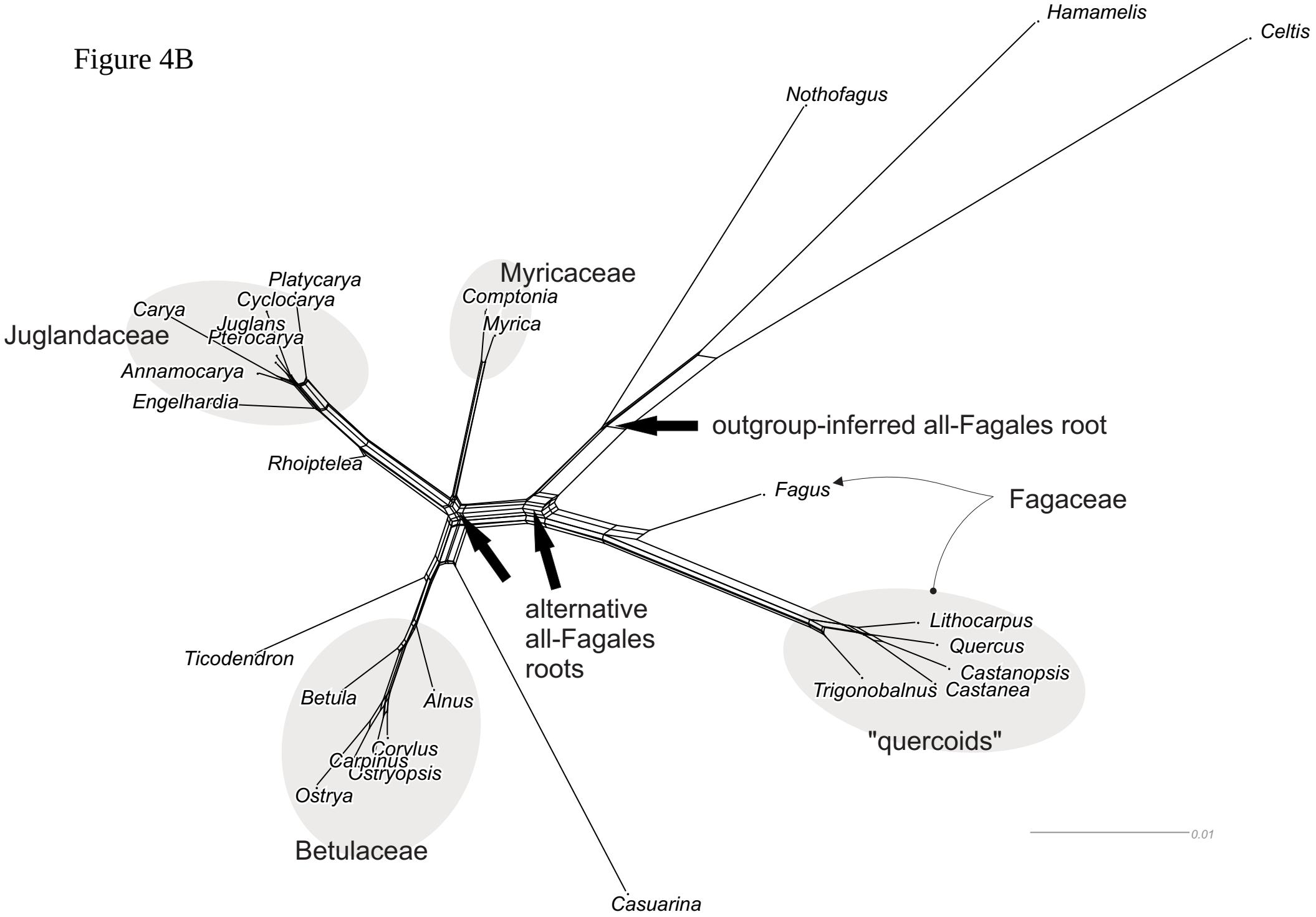

Hamamelis
Celtis
Nothofagus
Myricaceae
Comptonia
Myrica
Platycarya
Cyclocarya
Carya
Juglans
Pterocarya
Juglandaceae
Annamocarya
Engelhardia
Rhoiptelea
outgroup-inferred all-Fagales root
Fagus
Fagaceae
alternative
all-Fagales
roots
Ticodendron
Betula
Alnus
Corylus
Carpinus
Ostryopsis
Ostrya
Betulaceae
Lithocarpus
Quercus
Castanopsis
Trigonobalnus
Castanea
"quercoids"
Casuarina
0.01

Figure 5

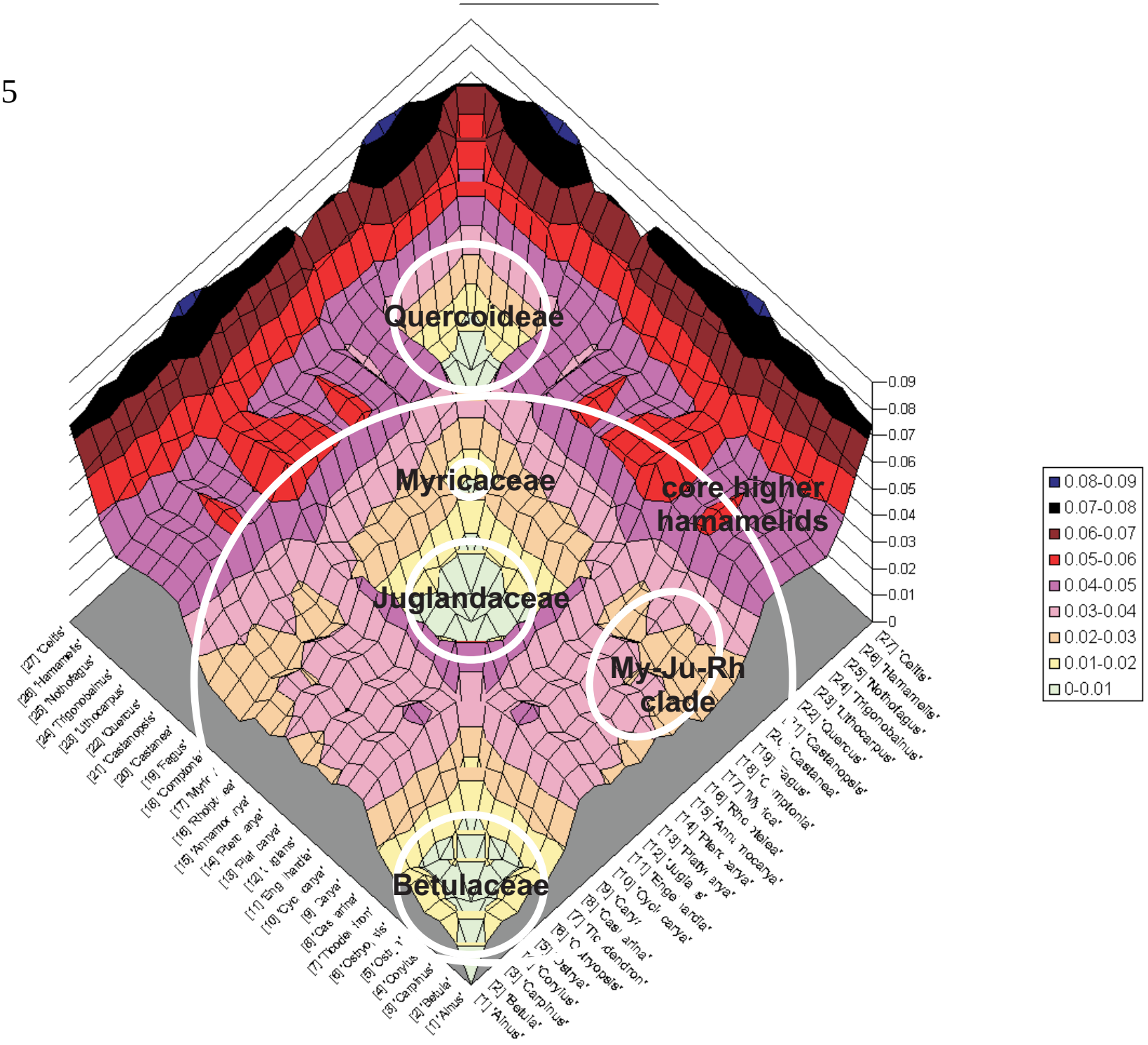

Quercoideae
Myricaceae
core higher hamamelids
Juglandaceae
My-Ju-Rh clade
Betulaceae
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0.08-0.09
0.07-0.08
0.06-0.07
0.05-0.06
0.04-0.05
0.03-0.04
0.02-0.03
0.01-0.02
0-0.01

Figure 6

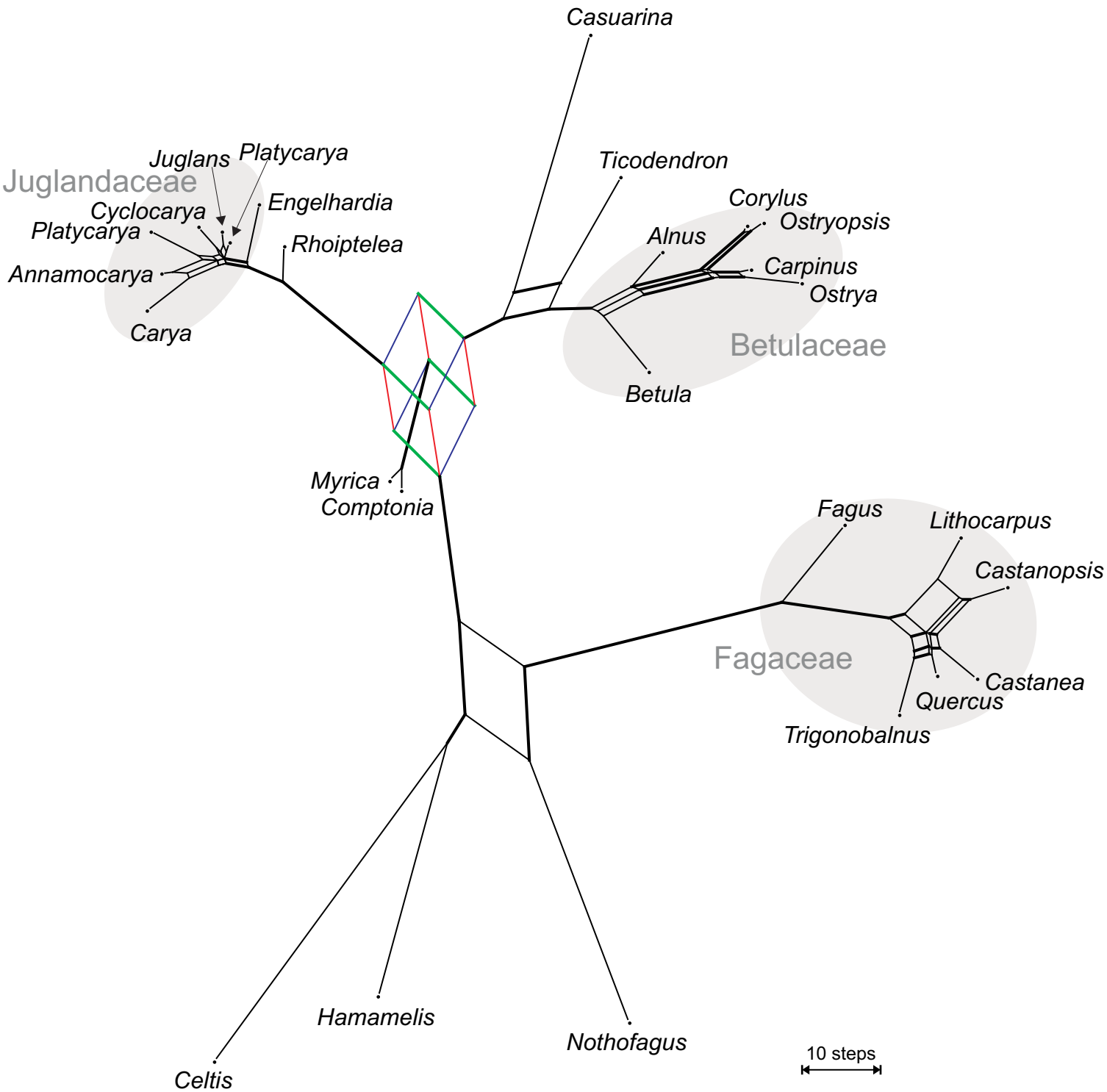

Casuarina
Juglans
Platycarya
Juglandaceae
Cyclocarya
Platycarya
Engelhardia
Rhoiptelea
Annamocarya
Carya
Ticodendron
Corylus
Ostryopsis
Alnus
Carpinus
Ostrya
Betulaceae
Betula
Myrica
Comptonia
Fagus
Lithocarpus
Castanopsis
Fagaceae
Castanea
Quercus
Trigonobalnus
Hamamelis
Nothofagus
Celtis
10 steps

Figure 7

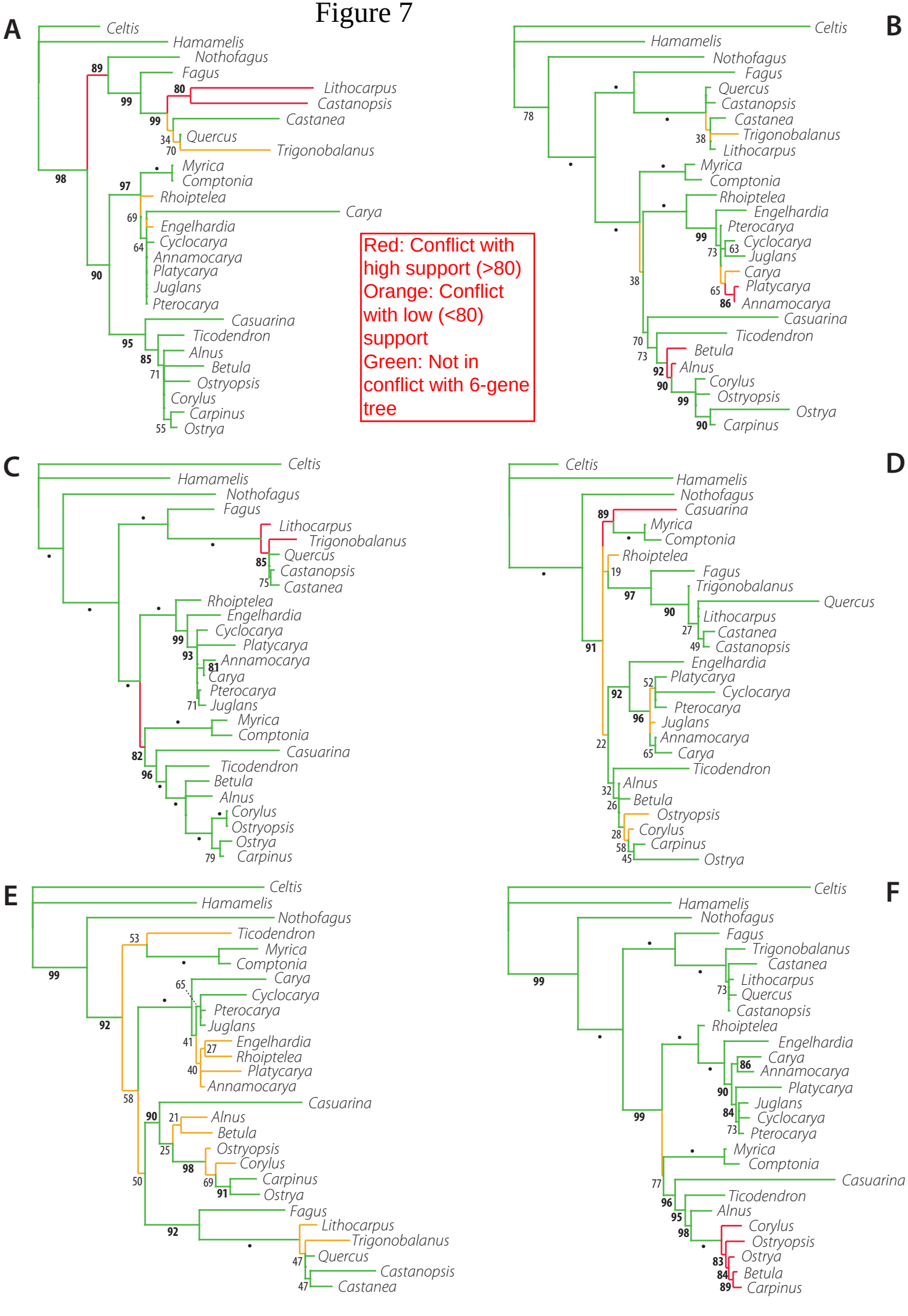
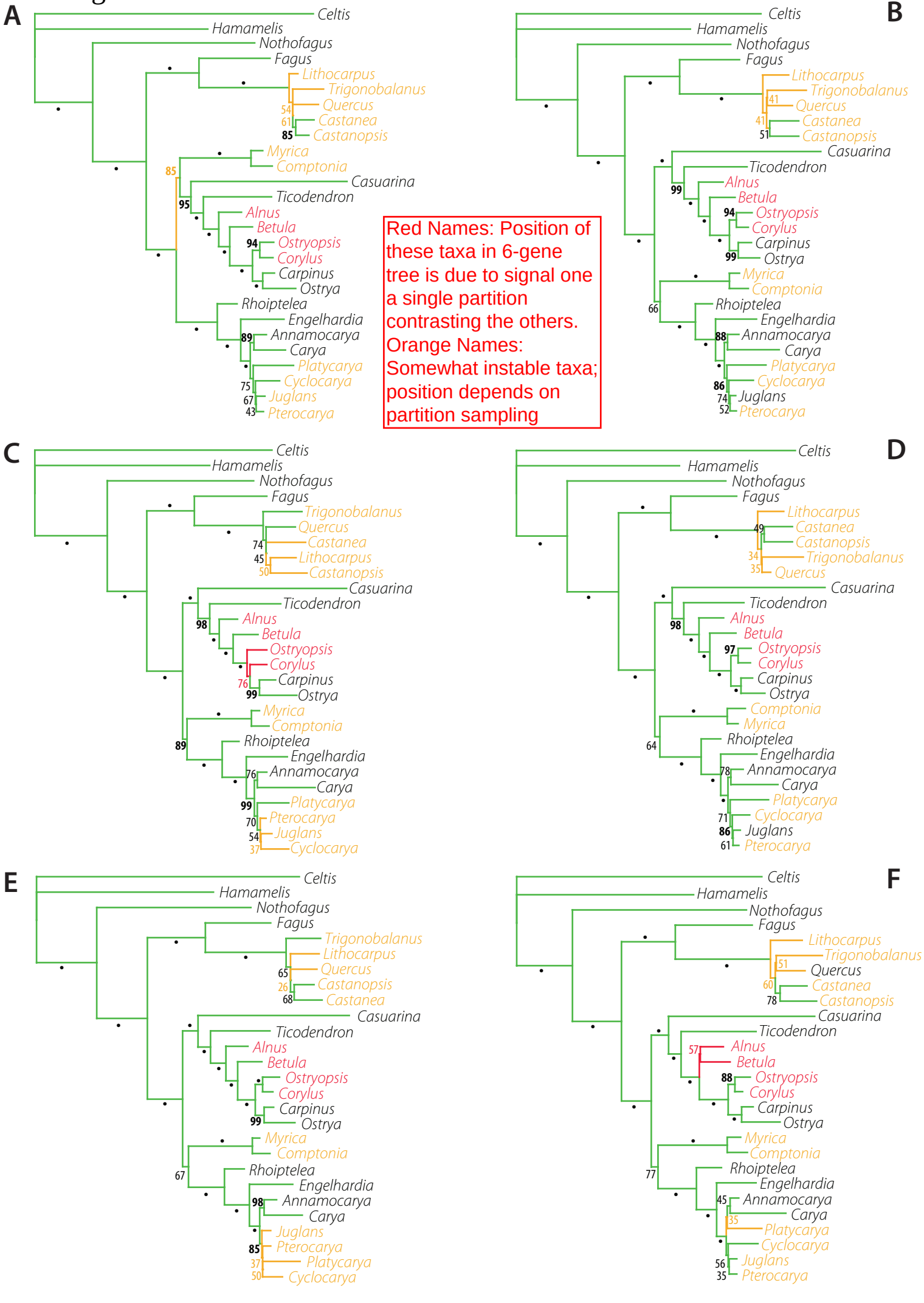
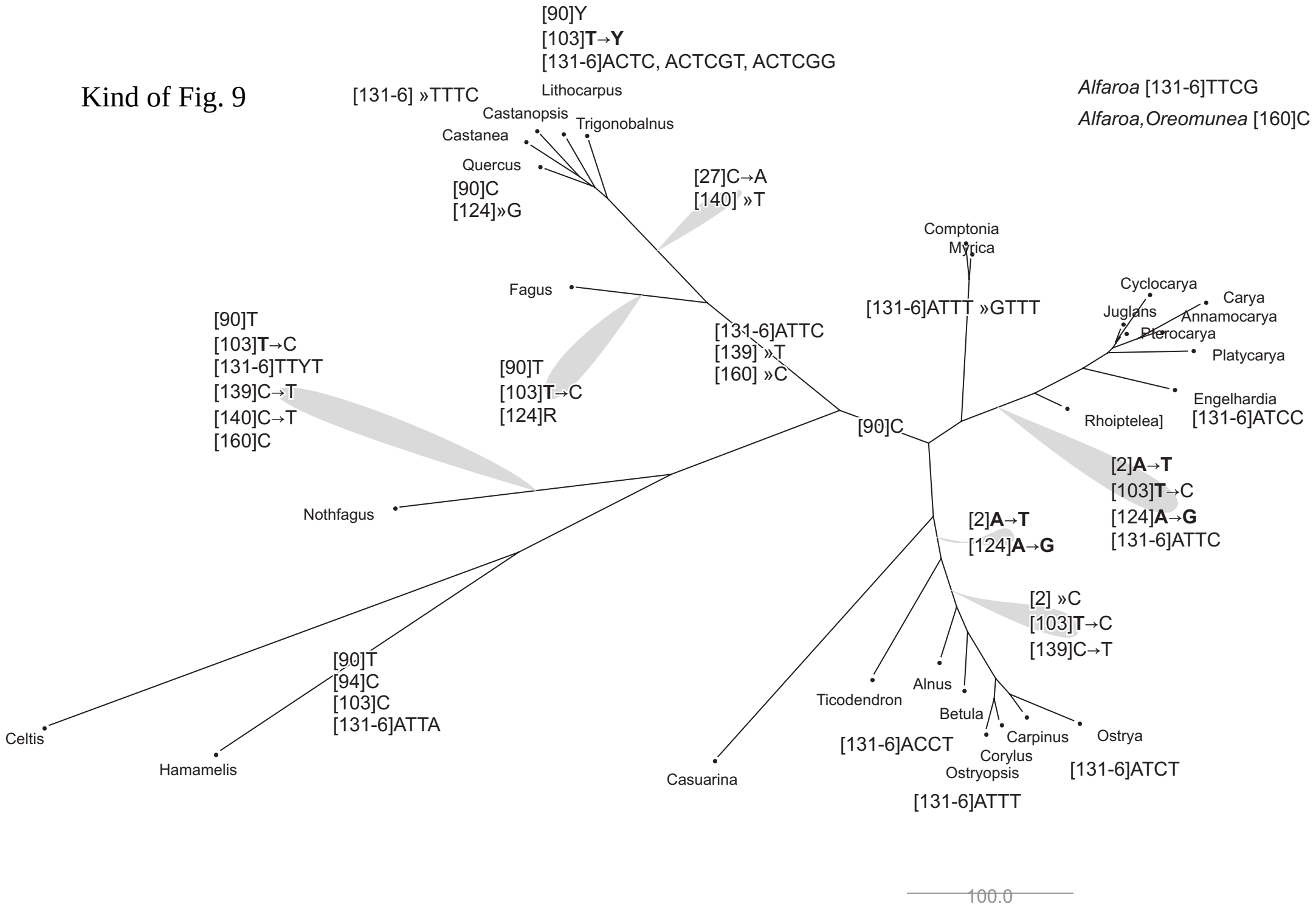

A
Celtis
Hamamelis
Nothofagus
Fagus
Lithocarpus
Castanopsis
Castanea
Quercus
Trigonobalanus
Myrica
Comptonia
Rhoiptelea
Carya
Engelhardia
Cyclocarya
Annamocarya
Platycarya
Juglans
Pterocarya
Casuarina
Ticodendron
Alnus
Betula
Ostryopsis
Corylus
Carpinus
Ostrya
Red: Conflict with high support (>80)
Orange: Conflict with low (<80) support
Green: Not in conflict with 6-gene tree
B
C
D
E
F

Figure 8


A
Celtis
Hamamelis
Nothofagus
Fagus
Lithocarpus
Trigonobalanus
Quercus
Castanea
Castanopsis
Myrica
Comptonia
Casuarina
Ticodendron
Alnus
Betula
Ostryopsis
Corylus
Carpinus
Ostrya
Rhoiptelea
Engelhardia
Annamocarya
Carya
Platycarya
Cyclocarya
Juglans
Pterocarya
B
Red Names: Position of these taxa in 6-gene tree is due to signal one a single partition contrasting the others.
Orange Names: Somewhat instable taxa; position depends on partition sampling
C
D
E
F

Kind of Fig. 9


[90]Y
[103]T→Y
[131-6]ACTC, ACTCGT, ACTCGG
Lithocarpus
[131-6] »TTTC
Castanopsis
Trigonobalnus
Castanea
Quercus
[90]C
[124]»G
[27]C→A
[140] »T
Alfaroa [131-6]TTCG
Alfaroa,Oreomunea [160]C
Comptonia
Myrica
[131-6]ATTT »GTTT
Cyclocarya
Carya
Juglans
Annamocarya
Pterocarya
Platycarya
Engelhardia
[131-6]ATCC
Rhoiptelea]
Fagus
[131-6]ATTC
[139] »T
[160] »C
[90]T
[103]T→C
[124]R
[90]T
[103]T→C
[131-6]TTYT
[139]C→T
[140]C→T
[160]C
[90]C
[2]A→T
[103]T→C
[124]A→G
[131-6]ATTC
Nothfagus
[2]A→T
[124]A→G
[2] »C
[103]T→C
[139]C→T
[90]T
[94]C
[103]C
[131-6]ATTA
Alnus
Ticodendron
Betula
Carpinus
Ostrya
[131-6]ACCT
Corylus
Ostryopsis
[131-6]ATCT
Celtis
Hamamelis
Casuarina
[131-6]ATTT
100.0