Syst. Biol. 59(4):384–399, 2010

DOI:10.1093/sysbio/syq027

Using Fossils to Break Long Branches in Molecular Dating: A Comparison of Relaxed Clocks Applied to the Origin of Angiosperms

SUSANA MAGALLÓN*

Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 3er Circuito de Ciudad Universitaria, Del. Coyoacán, México D.F. 04510, México;
**Correspondence to be sent to: Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 3er Circuito de Ciudad Universitaria, Del. Coyoacán, México D.F. 04510, México;*
E-mail: s.magallon@ibiologia.unam.mx.

Received 14 February 2009; reviews returned 26 May 2009; accepted 16 February 2010
Associate Editor: Mark Fishbein

Abstract.—Long branches are potentially problematic in molecular dating because they can encompass a vast number of combinations of substitution rate and time. A long branch is suspected to have biased molecular clock estimates of the age of flowering plants (angiosperms) to be much older than their earliest fossils. This study explores the effect of the long branch subtending angiosperms in molecular dating and how different relaxed clocks react to it. Fossil angiosperm relatives, identified through a combined morphological and molecular phylogenetic analysis for living and fossil seed plants, were used to break the long angiosperm stem branch. Nucleotide sequences of angiosperm fossil relatives were simulated using a phylogeny and model parameters from living taxa and incorporated in molecular dating. Three relaxed clocks, which implement among-lineage rate heterogeneity differently, were used: penalized likelihood (using 2 different rate smoothing optimization criteria), a Bayesian rate-autocorrelated method, and a Bayesian uncorrelated method. Different clocks provided highly correlated ages across the tree. Breaking the angiosperm stem branch did not result in major age differences, except for a few sensitive nodes. Breaking the angiosperm stem branch resulted in a substantially younger age for crown angiosperms only with 1 of the 4 methods, but, nevertheless, the obtained age is considerably older than the oldest angiosperm fossils. The origin of crown angiosperms is estimated between the Upper Triassic and the early Permian. The difficulty in estimating crown angiosperm age probably lies in a combination of intrinsic and extrinsic complicating factors, including substantial molecular rate heterogeneity among lineages and through time. A more adequate molecular dating approach might combine moderate background rate heterogeneity with large changes in rate at particular points in the tree. [Autocorrelation; fossil constraints; molecular clock; rate smoothing; seed plant phylogeny; total evidence; rate heterogeneity.]

Long branches, separated by short ones, are famously problematic in phylogeny estimation (e.g., Felsenstein 1978; Hendy and Penny 1989; Anderson and Swofford 2004). Branches estimated from DNA sequences in a phylogenetic tree represent the inseparable combination of substitution rate and time. Therefore, long branches are likely to represent a challenge to molecular dating as well because they can potentially encompass a vast number of combinations of rate and time. This challenge may be especially severe in the absence of independent information of substitution rates or absolute times to guide molecular dating. Yet, surprisingly little is known about the effect of long branches in molecular dating, either theoretically or empirically.

Flowering plants (angiosperms) are an extraordinary evolutionary success. They encompass a vast species diversity and extensive morphological and functional innovation, and significantly, they constitute the structural and energetic basis of most living terrestrial ecosystems. The origin and diversification of angiosperms determinantly influenced the composition and function of modern terrestrial life. Molecular dates of the onset of living angiosperm diversification (crown angiosperms) almost ubiquitously predate, usually substantially so (e.g., Wikström et al. 2001; Bell et al. 2005; Magallón and Sanderson 2005; Moore et al. 2007), the oldest unequivocal angiosperm fossils (reviewed in Friis et al. 2006). Implicitly or explicitly, all molecular estimates of crown angiosperm age have used phylogenetic hypotheses in which angiosperms are separated from their closest living relatives by a very long branch. Although there is no theoretical or empirical basis for expecting long branches to bias dating toward older (or younger) ages, it has been informally suggested that this long branch may be responsible for the very old molecular estimates of angiosperm age. Temporal constraints on this branch are unavailable because apparently all living gymnosperms are distantly related to angiosperms, all close evolutionary relatives to angiosperms are extinct, and a clear understanding of phylogenetic relationships among major seed plant lineages, living and extinct, is still pending (e.g., Mathews 2009). The discrepancies between molecular and fossil crown angiosperm ages, which in some cases exceed 100 Myr (e.g., Sanderson and Doyle 2001; Magallón and Sanderson 2005), could be the consequence of an incomplete fossil record. The fossil duration of a lineage represents only a fraction of its phylogenetic history because the stratigraphic record does not preserve phylogenetic speciation events, but rather, the appearance of distinctive morphological characters and character sets. The magnitude of missing fossil histories can range from very short to very long, depending on morphological and life history traits and on the nature of the stratigraphic record. The probability of a missing fossil history of the magnitude implied by molecular dates might be estimated (Foote et al. 1999). Nevertheless, the nature and pattern of the early angiosperm fossil record suggest that an extensive missing fossil history is unlikely.

Angiosperm fossil history begins in the Lower Cretaceous (130–140 Ma; Friis et al. 2006) with monoaperturate pollen grains that closely resemble those of living angiosperms in details of their wall structure. This type of pollen first occurs in a few late Hauterivian (130–136 Ma) localities (Hughes and McDougall 1987; Hughes et al. 1991; Hughes 1994), where it is very scarce and has a limited morphological diversity. In younger Lower Cretaceous sediments, there is a continuous increase in the number of localities at a global scale that contain angiosperm pollen, and in its local abundance and morphological diversity, until angiosperm pollen becomes ubiquitous at a worldwide scale. Late Barremian to early Aptian (125 Ma) sediments contain unequivocal angiosperm macroscopic remains. The oldest angiosperm fossils whose systematic affinity can be reliably identified belong to the earliest phylogenetic branches, including Nymphaeales (Friis et al. 2001), Austrobaileyales (Friis et al. 2006), Chloranthales (Friis et al. 1997; Eklund et al. 2004), magnoliids (Doyle 2000), early monocots (Friis et al. 2004, 2006), and early eudicots (Hughes and McDougall 1987; Doyle 1992; Leng and Friis 2003; Friis et al. 2006). In contrast, molecular estimates indicate the origin of crown angiosperms in the Lower Jurassic (175–200 Ma, Sanderson 1997; Sanderson and Doyle 2001; Wikström et al. 2001; Bell et al. 2005; Magallón and Sanderson 2005), in the Triassic (200–250 Ma; Sanderson and Doyle 2001; Magallón and Sanderson 2005), or in the Paleozoic (300 or 350 Ma; e.g., Ramshaw et al. 1972; Martin et al. 1989). In view of the substantial congruence between the fossil record and molecular phylograms regarding the sequence and tempo of early angiosperm diversification, the disagreement between fossil and molecular estimates on crown angiosperm age is unexpected.

Currently available relaxed molecular clocks are increasingly powerful approaches to estimate divergence times, but there is still a limited understanding of their accuracy and reliability under different estimation conditions. In this study, the possibility that the long angiosperm stem branch misguides molecular clocks into estimating an excessively old crown angiosperm age is empirically evaluated. The main objectives of this study are to investigate the role of a long branch in molecular dating and the response of different relaxed clocks to it. These questions will be addressed by 1) evaluating the effect of breaking the long angiosperm stem branch by including angiosperm stem lineage relatives and 2) using different relaxed clocks to estimate ages across the tree, with a special focus on the angiosperm crown node.

Materials and Methods

Taxa, Data, and Phylogenetic Analysis

The taxonomic sample consists of 40 angiosperms (Angiospermae sensu Cantino et al. 2007), 3 gnetophytes (Gnetophyta), 14 conifers (5 Pinaceae and 9 Cupressophyta), *Ginkgo biloba* (Ginkgophyta), 3 cycads (Cycadophyta), 7 ferns (Monilophyta, including whisk ferns, horsetails, eusporangiate, and leptosporangiate ferns), 1 lycophyte (Lycopodiophyta), and 1 liverwort (Marchantiophyta), for a total of 70 species in 69 genera. The data are the nucleotide sequences of 4 highly conserved plastid protein-coding genes: *atpB* (e.g., Hoot et al. 1995; Savolainen et al. 2000), *psaA*, *psbB* (e.g., Graham and Olmstead 2000; Sanderson et al. 2000), and *rbcL* (e.g., Chase et al. 1993; Soltis et al. 1999). Each gene was aligned manually, and uneven segments at the 5′ and 3′ end were removed. The data set is based on Magallón and Sanderson (2005). Newly added sequences were downloaded from GenBank or obtained for this study following the molecular protocols in Magallón and Sanderson (2002, 2005). Included taxa and GenBank accession numbers are listed in online Appendix 1 (available from http://www.sysbio.oxfordjournals.org). The data set is deposited in TreeBase (S2634).

The best-fit model for each codon position of each gene was identified with the Akaike information criterion using Modeltest (Posada and Crandall 1998; Posada and Buckley 2004). A comparison of estimated model parameter values indicated similarities in the same position among different genes but substantial differences between first–second codon positions and third codon position. The data were therefore partitioned into first–second and third codon positions.

Phylogenetic relationships were estimated with Bayesian analysis using MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001). Best-fit models were applied to each data partition with unlinked parameters and allowing different rate variation. The Metropolis-coupled Markov chain Monte Carlo (MC^3) analysis consisted of 2 independent runs of 5×10^6 generations in which 1 of every 200 trees was sampled. The outputs of MrBayes were examined with Tracer v1.4 (Rambaut and Drummond 2007) to check for convergence of different parameters, to determine the approximate number of generations at which log-likelihood values stabilized, and to identify the effective sample size (ESS) for each parameter and the estimated magnitude of model parameters. Posterior probabilities (PPs) of clades were obtained from the 50% majority-rule (MR) consensus of sampled trees, after excluding the initial 10% as burn-in. Topological convergence between the 2 Markov chain Monte Carlo (MCMC) runs was evaluated through a bivariate plot of PPs of splits (i.e., bipartitions), using the program AWTY (http://ceb.csit.fsu.edu/awty; Wilgenbusch et al. 2004; Nylander et al. 2008). The maximum a posteriori (MAP) tree topology was used as a working hypothesis of phylogenetic relationships. The 50% MR consensus of sampled trees is fully resolved and topologically identical to the MAP tree. Trees were rooted by selecting *Marchantia*, the single nonvascular plant in the sample, as an outgroup.

Breaking the Long Angiosperm Stem Branch

Identifying angiosperm stem relatives.—A usual approach to break a long branch is to increase the taxonomic sample, so that newly added taxa may transform the long branch into several shorter ones (Hendy and Penny 1989). In this case, taxa that can break the long angiosperm stem branch are extinct. To break the long angiosperm stem branch, an experimental strategy that involved simulating sequences for angiosperm stem relatives was implemented.

Angiosperm stem relatives were identified through a parsimony analysis for living and fossil seed plants using a combined morphological and molecular data set. The taxonomic sample consisted of 46 seed plants including 13 extinct gymnosperms, 21 living gymnosperms, and 12 living angiosperms. Morphological data were the 121 characters scored by Doyle (2006), and molecular data were the nucleotide sequences of *atpB*, *psaA*, *psbB*, and *rbcL*. To avoid well-documented artifacts in parsimony seed plant phylogeny estimation (e.g., Sanderson et al. 2000; Magallón and Sanderson 2002; Soltis et al. 2002; Aris-Brosou 2003; Burleigh and Mathews 2004; Mathews 2009), only first–second codon positions were included. Molecular characters of fossil taxa were scored as question marks. The total evidence data set is deposited in TreeBase (S2634). The parsimony analysis was conducted with PAUP* v4.0b 10 (Swofford 2002) using a heuristic search with 100 random addition replicates with stepwise addition of taxa and tree bisection and reconnection branch swapping. *Elkinsia*, the oldest known fossil seed plant (Rothwell and Scheckler 1988), was specified as an outgroup. Clade support was obtained through a bootstrap (BS) analysis consisting of 500 replicates equal to the parsimony searches, except for having 10 random addition replicates.

The total evidence analysis identified 3 extinct branches between angiosperms and any living gymnosperm (see Results section). The first branch to diverge from the angiosperm stem lineage includes Glossopteridales and *Pentoxylon*, the second is Bennettitales, and the third, representing the sister group of angiosperms, is *Caytonia*. These angiosperm stem relatives were inserted in the Bayesian MAP tree along the branch subtending angiosperms.

Assigning branch lengths to angiosperm stem relatives.—Branch lengths for angiosperm stem relatives were postulated considering the mean path length between the seed plant crown node and all living terminals, the age of the seed plant crown node, and stratigraphic ranges of angiosperm stem relatives. The mean path length between the seed plant crown node and all living terminals was calculated from the branch lengths in the 50% MR consensus of sampled trees in the MC^3. The age of the seed plant crown node was postulated as 350 Ma old because this age is younger than the oldest fossil seeds (*Elkinsia polymorpha*, Famennian [Upper Devonian]; Rothwell et al. 1989) and older than the oldest presumed crown seed plants (Cordaitales, Namurian [Lower-Upper Carboniferous]; Taylor T.N. and Taylor E.L. 1993) and falls within the interval estimated for this node by different relaxed clocks (including only living taxa, see Results section). Stratigraphic ranges of angiosperm stem relatives are as follows: Glossopteridales: Permian (299 to 251 Ma); Pentoxylales: Carnian (Upper Triassic) to middle Upper Cretaceous (228 to 85.8 Ma); Bennettitales: Upper Triassic to end Cretaceous (228 to 65.5 Ma); and Caytoniales: Jurassic to end Cretaceous (199.6 to 65.5 Ma). The total path length between the seed plant crown node and each of the 4 angiosperm stem relatives was proportional to the mean path length to living terminals, trimmed considering the age of the upper boundary of the youngest stratigraphic interval from which each fossil group is known. Internal branch lengths were assigned by postulating the age of divergence of angiosperm stem relatives as the earliest appearance of the lineage in the fossil record plus 10% of its stratigraphic range and assuming that branch length is directly proportional to time. Finally, the divergence of the branch leading to Glossopteridales plus Pentoxylales from the angiosperm stem lineage was placed at the midpoint between the seed plant crown node and the Glossopteridales–Pentoxylales split.

Sequences of angiosperm stem relatives.—Nucleotide sequences for angiosperm stem relatives were simulated using Seq-Gen v1.3.2 (Rambaut and Grassly 1997). The input phylogram was the 50% MR consensus of sampled Bayesian trees with inserted angiosperm stem relatives (as described above). Model parameters were obtained from maximum likelihood (ML) optimization of sequence data for living taxa on the MAP topology (using PAUP*). One hundred data sets of the same size as the original were generated. Each was used to obtain a phylogram by optimizing branch lengths and model parameter values with ML (using PAUP*) on the MAP topology with inserted angiosperm stem relatives.

Phylogenetic analysis including angiosperm stem relatives.—One simulated data set selected at random was used to estimate phylogenetic relationships among living and extinct taxa. Phylogenetic estimation was conducted with MrBayes as described above, except for implementing the best-fit model for the unpartitioned data. Bayesian outputs were examined with Tracer. Topological convergence between the 2 MCMC runs was evaluated with AWTY. PPs of clades were obtained from the 50% MR consensus of sampled trees (excluding burn-in). The MAP tree was used as a working hypothesis of phylogenetic relationships.

Age Estimation

Ages were estimated with an autocorrelated (Gillespie 1991) semiparametric relaxed clock (penalized likelihood [PL]; Sanderson 2002), an autocorrelated Bayesian relaxed clock (multidivtime; Thorne et al. 1998; Kishino et al. 2001; Thorne and Kishino 2002), and an uncorrelated Bayesian relaxed clock (uncorrelated lognormal relaxed molecular clock; Drummond et al. 2006; Drummond and Rambaut 2007). Two PL analyses were conducted, each using the optimal rate smoothing derived from alternative cross-validation procedures. A set of analyses was conducted on phylograms with only living taxa and including the long angiosperm stem branch, and another set was conducted on phylograms in which the angiosperm stem branch was broken by including stem relatives. All analyses implemented temporal calibrations and constraints derived from critically evaluated fossil information.

Calibration and constraints.—Trees were calibrated by fixing the age of the vascular plant crown node (Tracheophyta) at 421 Ma. This age is derived from the Upper Silurian (Ludlow) age of zosterophyllophytes and *Baragwanathia* (Tims and Chambers 1984; Garrat and Rickards 1987; Hueber 1992), the oldest fossils that can be securely assigned to either of the lineages diverging from the node (Kenrick and Crane 1997). Nineteen internal nodes were constrained with minimal ages and one, the tricolpate angiosperm (Eudicotyledoneae; eudicots) crown node, with a maximal age (see ahead). When included, angiosperm stem relatives were specified as nonextant terminals, and their ages were fixed at 65.5 Ma for Caytoniales and Bennettitales, at 85.8 Ma for Pentoxylales, and at 251 Ma for Glossopteridales (see above). Three additional internal nodes, resulting from the inclusion of extinct stem relatives, were constrained with minimal ages (Fig. 1). Stratigraphic occurrences of fossils were transformed into absolute ages using the upper boundary of the narrowest stratigraphic interval to which they are assigned, according to Gradstein and Ogg (2004). Fossils used for calibration and constraints, stratigraphic ranges, and absolute ages are listed in online Appendix 2.

Test of the molecular clock.—To test for among-lineage substitution rate constancy, a likelihood ratio test of the molecular clock (Felsenstein 2004) and the Langley-Fitch test (r8s v1.71, Sanderson 2003, 2006) were implemented. In the Langley–Fitch test, ML branch lengths are compared with and without assuming rate constancy.

PL dating.—Autocorrelated semiparametric dating was conducted with PL (Sanderson 2002), implemented in r8s v1.71 (Sanderson 2003, 2006). Analyses including the angiosperm stem branch were conducted on 100 randomly chosen phylograms topologically identical to the MAP tree among those sampled by the MC^3 during Bayesian phylogeny estimation. Phylograms topologically identical to the MAP among those sampled in the MC^3 were identified with PAUP* using the FILTER command and specifying the MAP topology as constraint. This strategy allowed to estimate dates on a fully resolved topology with branch lengths derived from optimization of the 2 data partitions and to use the mean and standard deviation of ages from the 100 phylograms as age estimate and associated error, respectively, for each node in the tree.

Optimal magnitudes for the rate smoothing parameter (λ) were estimated with 2 types of data-driven cross-validations (Sanderson 2002, 2006): A branch-pruning cross-validation sequentially removes terminal phylogram branches to calculate prediction error (Sanderson 2002). A more recently implemented fossil-based cross-validation sequentially unconstrains nodes constrained with a minimal or a maximal age and estimates rates and times across the full tree under a given rate smoothing magnitude. A fossil-derived age is violated if the estimate is younger than a minimal age constraint or older than a maximal age constraint (Sanderson 2006). Each cross-validation was conducted on a single phylogram chosen at random from among the 100 used for dating, testing 40 smoothing magnitudes from $\log_{10}\lambda = -2.0$ to 5.8 at 0.2 intervals. PL analyses were conducted using the branch-pruning (PLBP) and the fossil-based (PLFB) smoothing magnitudes. These analyses encompassed moderate and high levels of among-lineage rate heterogeneity, respectively (see Results section). PL analyses used a truncated Newton algorithm with bound constraints with 5 initial restarts and 3 perturbed restarts of magnitude 0.05 in random directions. Comparable analyses were conducted on data sets including angiosperm stem relatives (PLBP* and PLFB*).

Multidivtime dating.—Autocorrelated Bayesian dating analysis was conducted with multidivtime (Multidistribute package; Thorne and Kishino 2002; http://statgen.ncsu.edu/thorne/multidivtime.html), a relaxed clock in which the rate of molecular evolution is modeled through time, assuming that different genes may have different substitution rates but share divergence times (Thorne et al. 1998; Kishino et al. 2001; Thorne and Kishino 2002). The multidivtime analysis including the long angiosperm stem branch (MD) was based on the MAP tree and considered the 4 genes simultaneously as unlinked data partitions with different substitution parameters. After the initial 500,000 cycles were discarded as burn-in, the MCMC was sampled 25,000 times with 200 cycles between samples.

Priors and standard deviations for analysis parameters were based on recommendations in multidivtime documentation. Time units equaled 100 Myr, and an age prior of 4.21 time units was assigned to the root. The mean of the prior of the rate at the root node was 0.04, derived from the constant rate estimated with the molecular clock method of Langley–Fitch (Langley and Fitch 1974; implemented in r8s; results not shown) and converted to the time units used in multidivtime. The mean of the prior for hyperparameter ν, which controls rate variation, was assigned a value of 0.36, so that the product of this value multiplied by the prior of the root node age fell between 1 and 2. Hyperparameter ν

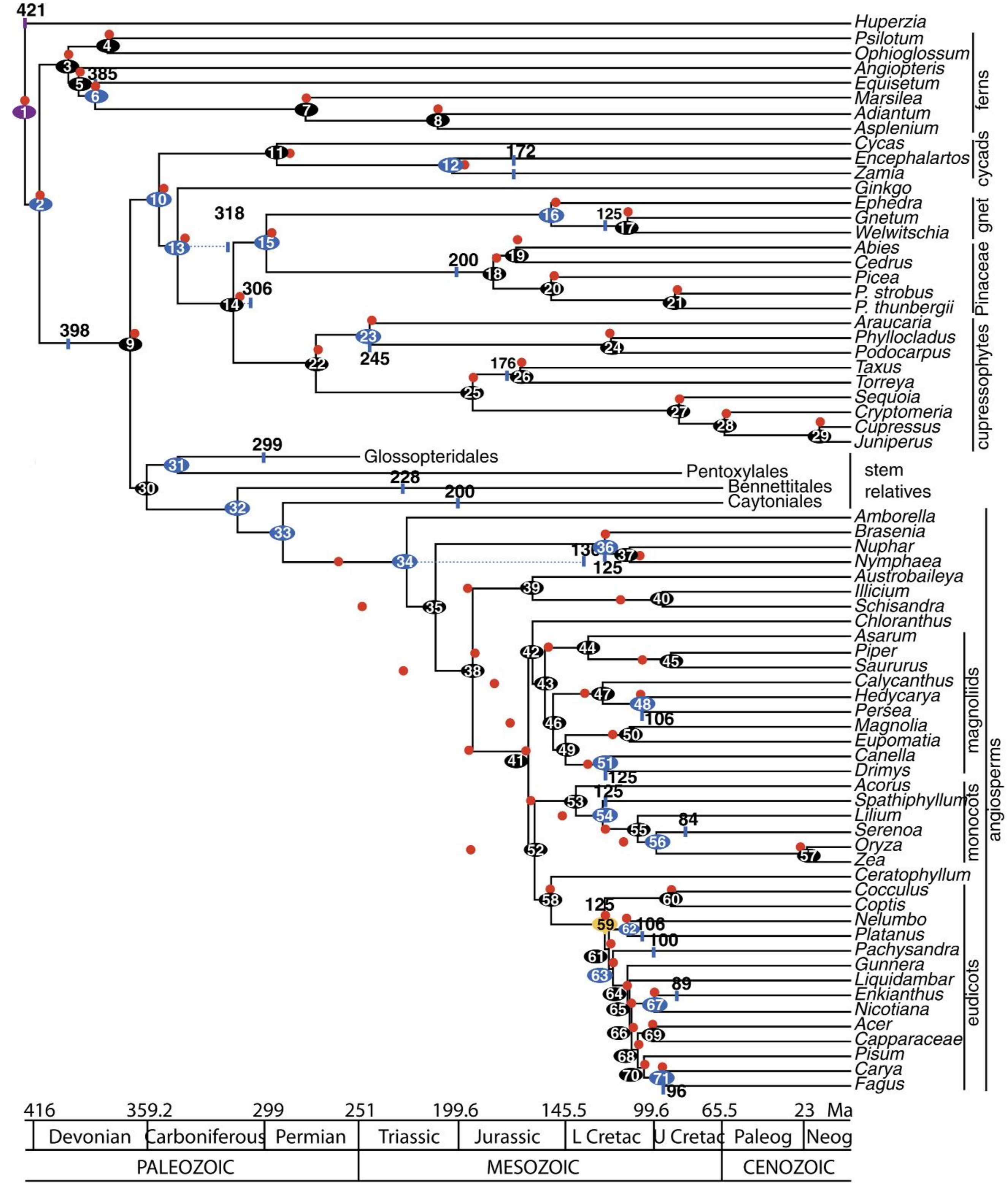

FIGURE 1. Dated tree. Chronogram with numbered nodes, obtained with PL using fossil-based rate smoothing, and breaking the angiosperm stem branch (PLFB*). Angiosperm stem relative branches terminate before the present. The calibration node is shown in purple, a maximal age–constrained node is in orange, and minimal age–constrained nodes are in blue. Blue bars indicate the phylogenetic placement and age of the fossils used to constrain the immediately subtending node or the second subtending node (dashed lines). Red dots indicate ages obtained with the same dating method but without breaking the angiosperm stem branch (PLFB). In this pair of analyses (PLFB and PLFB*), breaking the angiosperm stem branch provided the largest age difference for the crown angiosperm node and for deep backbone angiosperm nodes. All branches are supported by ≥0.95 PP, except for the branches subtending nodes 5, 6, 42, 52, 58, and 66.

was modeled independently for the 4 genes, allowing them to vary in different ways from a molecular clock. Standard deviations around the age and rate at the root and hyperparameter ν were each made equal to their respective magnitudes (multidivtime documentation). The bigtime parameter was assigned 4.3 time units, considering the Lower Devonian (Ludlow; 422.9–418.7 Ma) age of the oldest vascular plant fossils (Tracheophyta; Garrat and Rickards 1987; Hueber 1992; Kenrick and Crane 1997). Nineteen nodes were constrained with lower age bounds and one with an upper age bound (Fig. 1; online Appendix 2). The analysis was performed 4 times under equal conditions (except for the seed) to check for convergence.

The multidivtime analysis breaking the angiosperm stem branch (MD*) was based on the MAP topology and one randomly chosen simulated data set for living taxa and angiosperm stem relatives. The analysis was conducted as described above, except for using unpartitioned data and including 3 additional lower bound age constraints, corresponding to angiosperm stem relatives (Fig. 1; online Appendix 2). In the MD and MD* analyses, ages of nodes and associated errors were derived from the mean and standard deviation, respectively, of trees sampled in the MCMCs.

Uncorrelated lognormal dating.—Uncorrelated Bayesian dating was conducted with the relaxed molecular clock model implemented in BEAST v1.4.7 (Drummond et al. 2006; Drummond and Rambaut 2007). An uncorrelated lognormal (UCLN) relaxed-clock model was chosen based on results from simulated and empirical data (Drummond et al. 2006). In the analysis including the angiosperm stem branch (UCLN), the data set was partitioned into first–second and third codon positions. Each partition was modeled with a general time reversible model, accounting for among-site rate variation and a proportion of invariable sites (GTR+I+G) but substitution parameters, rate heterogeneity, and base frequencies were unlinked across the 2. A Yule prior was assigned to the branching process. Twenty taxon sets were specified with the MRCA command, and the ingroup (ferns plus seed plants; Euphyllophyta) was constrained to be monophyletic. The tree root height was assigned a uniform prior bounded between 420 and 422 Ma before the present. Lognormal priors were assigned to the height of 19 internal nodes, based on fossil-derived minimal ages of clades (online Appendix 2). In each case, the lognormal mean was equal to (fossil age + 10 Myr), to accommodate for the fact that a lineage's divergence is older than its first fossil appearance, and the zero offset was equal to (fossil age – 5 Myr), to ensure that the minimal age fell within the distribution. The height prior for the eudicot crown node was given a uniform distribution bounded between 122 and 128 Ma. The dated tree obtained from the PLBP analysis (see above) was used as a starting tree in MCMC searches. A preliminary MCMC (1×10^6 steps) estimated a topology equal to the input topology; hence, in subsequent MCMCs, tree topology was fixed by deselecting the 4 tree operators. Eight independent MCMCs of different lengths were run for a combined total of 50×10^6 steps, in which 1 of every 200 trees was sampled. The program Tracer was used to visually evaluate the behavior of each chain, determine appropriate burn-in cutoff (10% of sampled trees), and obtain the ESS for each parameter. The 8 chains were combined to obtain estimates of posterior distributions for all parameters.

Analyses breaking the angiosperm stem branch (UCLN*) used the same simulated data set as the MD* analysis, to which an unpartitioned GTR+I+G substitution model was applied. Priors were assigned as described above, and also included height priors of nodes associated to angiosperm stem relatives (online Appendix 2). The dated tree obtained in the PLBP* analysis was used as a starting tree in MCMC runs. After a preliminary run indicated that estimated tree topology was equal to the input topology, this parameter was no longer estimated. Several independent MCMCs of variable lengths and sampling intervals were conducted, for a total of 60×10^6 steps and 27,000 sampled trees. After visual examination of the chains and estimation of parameter ESS (with Tracer), the initial 10% of sampled trees in each run was excluded as burn-in and the remaining trees were combined to obtain posterior distributions of all parameters. In the UCLN and UCLN* analyses, the mean and the 95% highest posterior density (HPD) interval of ages in all sampled trees were used as age and associated error, respectively, for each node.

Results

Phylogenetic Analysis

Sequences of the 4 genes were available for all sampled taxa, except for the *psbB* sequence of *Brasenia* (Nymphaeales; online Appendix 1). Each of the 4 genes could be aligned unambiguously. Concatenated alignments were 6717 bp long, of which uneven segments at the 5′ and 3′ ends of each gene were removed, leaving 6315 bp that were used for model estimation and phylogenetic and dating analyses.

Best-fit models for first–second and third codon positions are of the GTR+I+G type, but parameter values differ substantially between the 2 partitions. After 5×10^6 generations, the average standard deviation of split frequencies between the 2 MCMCs was 0.002. The initial 10% (500,000 generations; 2500 sampled trees) were excluded as burn-in. Tree lnL and other parameters stabilized before the burn-in cutoff. The combined runs reached an ESS of 1717.14 for tree lnL, and the ESSs of all other parameters were over 200, in most cases substantially so. Graphical examination of topological convergence with AWTY showed a high correlation of PPs of splits in the 2 MCMC runs (graphs not shown).

The MAP tree accounted for 0.157 of the total PP. The 50% MR consensus of the combined sample of trees was fully resolved and topologically identical to the MAP tree. Phylogenetic relationships in the MAP tree are congruent with the most frequently obtained results among major groups of seed plants (Fig. 1), including the sister relationship between ferns and seed plants, and the rooting of seed plants on the branch between angiosperms and all living gymnosperms. *Amborella* (Amborellales) is the sister to all other angiosperms, and within core angiosperms (Mesangiospermae), *Chloranthus* (Chloranthales) and magnoliids are sister taxa, and monocots are the sister to *Ceratophyllum* (Ceratophyllales) plus eudicots (Fig. 1). Most branches are supported by PPs ≥0.95 (Fig. 1).

Breaking the Long Angiosperm Stem Branch

Identifying angiosperm stem relatives.—The total evidence data set included 4331 characters (121 morphological, 4210 first–second DNA codon positions) of which 544 (12.56%) were parsimony informative. Considering morphological characters alone, the per-taxon percentage of missing characters ranged from 7.4% to 74.4% (median: 34.7; standard deviation: 17.63). The total evidence analysis resulted in 30 most parsimonious trees (MPTs) of 1358 steps (confidence interval = 0.503, retention index = 0.781, rescaled consistency index = 0.393; all scores based on parsimony-informative characters). The strict consensus of MPTs is shown in Figure 2.

The strict consensus tree is well resolved, but few clades are supported by high BS values (Fig. 2). *Elkinsia*, *Lyginopteris*, and medullosans (Paleozoic "seed ferns") form a grade outside the seed plant crown group (93% BS). In contrast with the most frequent result in molecular analyses, the root of crown seed plants separates a clade that includes *Ginkgo biloba*, gnetophytes, and conifers, from a clade that includes cycads and angiosperms (see Discussion section). These 2 clades also contain several extinct gymnosperms, and neither is supported by ≥50% BS. Three extinct branches are placed between cycads and angiosperms: the first contains *Glossopteris* (Glossopteridales; Permian Gondwanan "seed ferns") plus *Pentoxylon* (Pentoxylales, Jurassic–Cretaceous gymnosperms); the second corresponds to Bennettitales (Triassic–Cretaceous gymnosperms similar to cycads but with different reproductive organs); and the third is *Caytonia* (Caytoniales; Upper Triassic–Cretaceous "seed ferns"), which is the sister group of angiosperms (66% BS). The finding that Glossopteridales plus Pentoxylales, Bennettitales, and Caytoniales are angiosperm stem relatives is congruent with some previously published results (e.g., Albert et al. 1994; Doyle 2006, 2008; Hilton and Bateman 2006; see Discussion section). These fossil gymnosperms are here postulated as angiosperm stem relatives (Figs. 1–2).

Assigning branch lengths to angiosperm stem relatives.—The mean path length between the seed plant crown node and all terminals in the 50% MR consensus of sampled Bayesian trees is 0.1428 substitutions per site/unit time. Postulated total lengths between the seed plant crown node and angiosperm stem relatives are 0.0404 for Glossopteridales, 0.1078 for *Pentoxylon*, and 0.1161 for each Bennettitales and *Caytonia*. Internal splits along the angiosperm stem lineage were placed at 326.9 Ma for the branch leading to Glossopteridales plus *Pentoxylon*, at 244.2 Ma for Bennettitales, and at 213.0 Ma for *Caytonia*. The split between Glossopteridales and *Pentoxylon* was placed at 303.8 Ma.

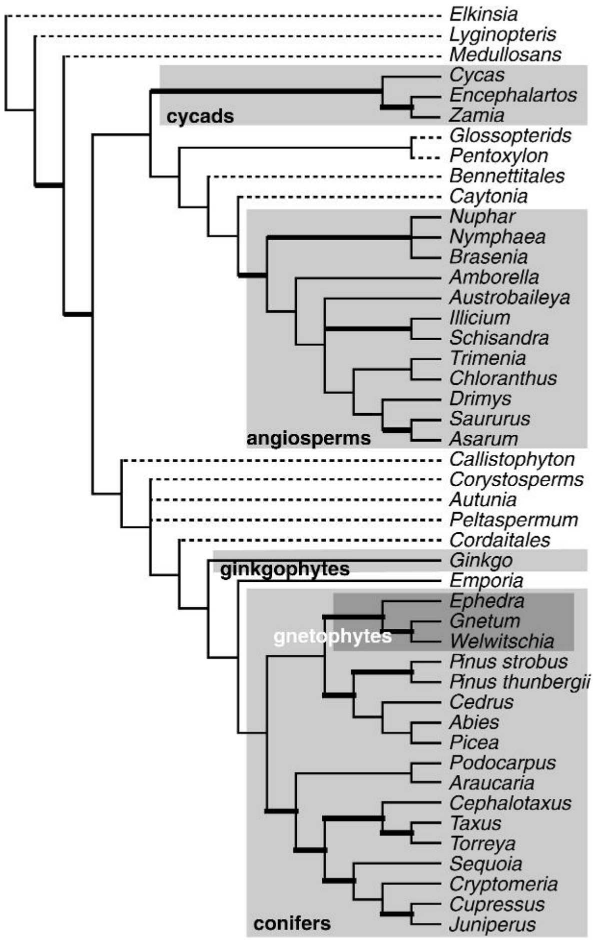

FIGURE 2. Total evidence analysis of living and fossil seed plants. Strict consensus of 30 MPTs obtained from morphological and molecular data for living and fossil seed plants. Branches in bold are supported by ≥85% BS. Dashed lines indicate a fossil branch.

A simulated data set was used for Bayesian phylogenetic analysis breaking the angiosperm stem branch. Examination of combined trace files indicated that all parameters rapidly converged and reached appropriate ESSs (well over 200). The initial 10% of sampled trees in each run was excluded as burn-in. The MAP tree accounted for 0.466 of the total PP. Graphical examination with AWTY showed a high correlation of PPs of splits, indicating topological convergence between the 2 MCMC runs (graphs not shown). The 50% MR consensus of the trees sampled in the 2 runs was fully resolved and identical to the MAP tree. Both are equal to the MAP tree for living taxa with angiosperm stem relatives inserted, except for relationships among core angiosperms: magnoliids are the sister of a (monocots (Ceratophyllales, eudicots)) clade (0.54 PP), instead of sister of Chloranthales. All but 2 of the branches in the tree are supported by ≥0.95 PP.

Age Estimation

The calibration node and minimal and maximal age constraints used in divergence time estimation are shown in Figure 1. The phylograms including and breaking the long angiosperm stem branch were found to be significantly unclocklike by likelihood ratio tests (likelihood ratio = 3723.47, $\chi^2_{0.05[\mathrm{DF}=67]} = 87.11$, $P \ll 0.001$; and likelihood ratio = 4876.26, $\chi^2_{0.05[\mathrm{DF}=71]} = 91.67$, $P \ll 0.001$, respectively) and Langley–Fitch tests ($P \ll 0.001$ for both phylograms).

PL cross-validations.—Breaking the angiosperm stem branch had negligible effects on the outcome of cross-validations. However, using the branch-pruning or the fossil-based approaches resulted in pronounced differences on the magnitude of the respective optimal rate smoothing (λ) and on the pattern of the error versus $\log_{10}\lambda$ plot (Fig. 3). Branch-pruning cross-validations resulted in chi-squared error versus $\log_{10}\lambda$ plots in which a smoothing magnitude of $\log_{10}\lambda = 2.0$, implying moderate rate heterogeneity, was unambiguously identified as optimal (Fig. 3a). Fossil-based cross-validations resulted in raw error versus $\log_{10}\lambda$ plots, where, as smoothing decreased, the associated raw error became smaller (although slightly; Fig. 3b). The smoothing magnitude with the lowest raw error was the smallest of the examined range, i.e., $\log_{10}\lambda = -2.0$; however, it is likely that a lower smoothing magnitude would have returned a slightly lower raw error. This level of smoothing implies high molecular rate heterogeneity.

PL dating with branch-pruning rate smoothing.—Among the implemented methods, PLBP and PLPB* tended to provide the youngest ages across the tree (Figs. 4–5, online Appendix 3). Except for the UCLN and UCLN* methods, associated errors around the mean age of most nodes are narrow (Fig. 5). The widest errors are found within ferns, gnetophytes, Pinaceae, and cycads, whereas the narrowest belong to nodes within core eudicots.

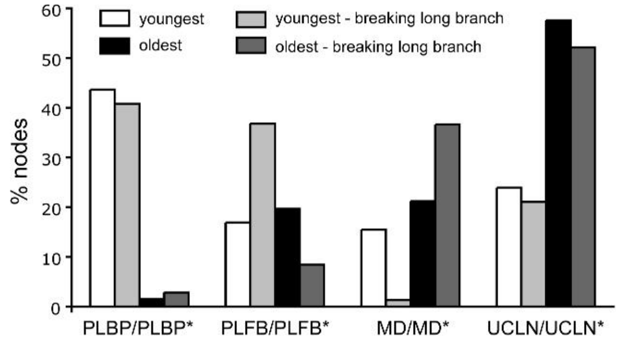

FIGURE 4. Oldest and youngest ages obtained by different dating methods. Percentage of nodes for which each method provided the youngest or oldest age. PLBP = penalized likelihood with branch-pruning rate smoothing; PLFB = penalized likelihood with fossil-based rate smoothing; MD = multidivtime; UCLN = uncorrelated lognormal; * = breaking the angiosperm stem branch.

Breaking the angiosperm stem branch (PLBP*) resulted in small changes in mean ages across the tree (online Appendix 3) and in younger estimates for slightly less than half of the nodes (42.6%). Only for 2 nodes did the absolute difference between PLBP* and PLBP mean ages exceed 10 My (Table 1). Except for 2 nodes (within ferns), mean ages estimated with PLBP fall within the associated error of the PLBP* age, and vice versa. The angiosperm crown node exhibited one of the largest differences between the 2 analyses (Table 1). The PLBP* mean age is 5.9 Myr older than the PLBP mean age (215.6 and 221.5 Ma old, respectively), representing the single instance in which breaking the angiosperm stem

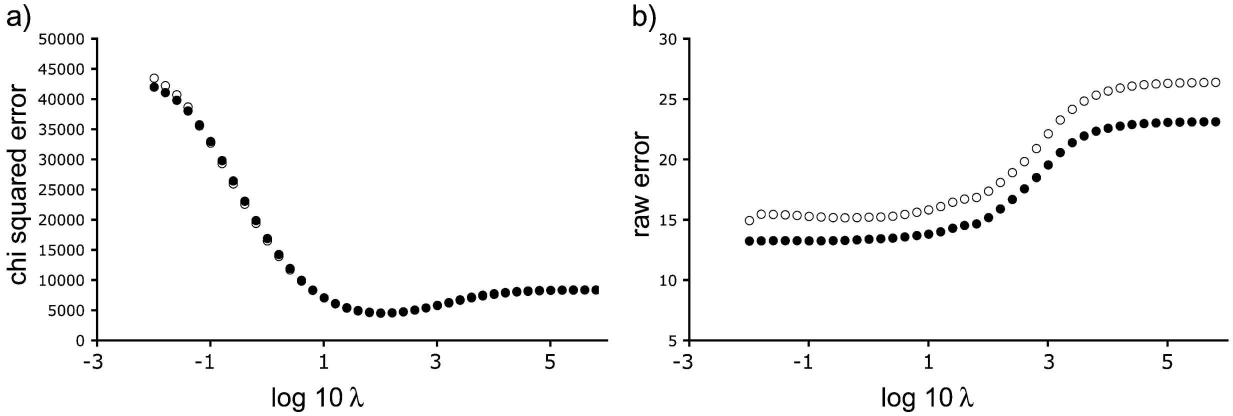

FIGURE 3. Cross-validation error versus smoothing magnitude. a) Branch-pruning cross-validations. Cross-validations with only living taxa (solid circles) and breaking the angiosperm stem branch (empty circles) are almost totally overlapped. In both, the lowest chi-squared error was associated with a $\log_{10}$ smoothing magnitude (λ) = 2.0. b) Fossil-based cross-validation. Cross-validations with only living taxa (solid circles) and breaking the angiosperm stem branch (empty circles) display the same pattern but different raw error magnitudes. Their respective lowest raw error was associated with a $\log_{10}\lambda = -2.0$.

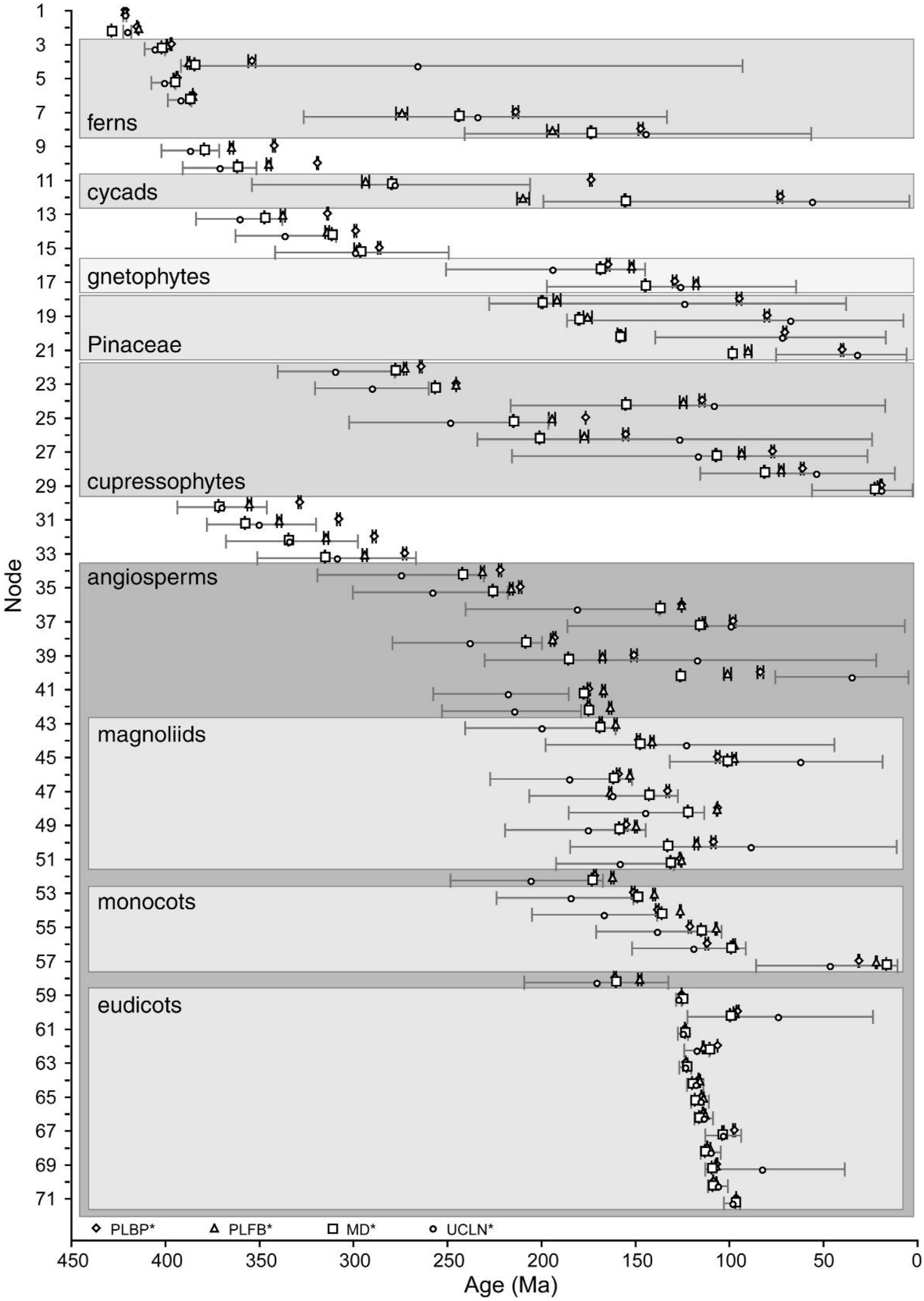

FIGURE 5. Ages obtained with different dating methods. Numbers on vertical axis correspond to nodes in the tree, as in Figure 1. For each node, the age obtained by each of the 4 methods, breaking the angiosperm stem branch, is shown (PLBP [diamonds] = penalized likelihood with branch-pruning rate smoothing; PLFB [triangles] = penalized likelihood with fossil-based rate smoothing; MD [squares] = multidivtime; UCLN [circles] = uncorrelated lognormal; * = breaking the angiosperm stem branch). For PLBP*, PLFB*, and MD*, the "point" estimate of age represents a mean, and the associated error is its standard deviation. For UCLN*, the "point" estimate of age represents a mean and the associated error is the 95% HPD.

TABLE 1. Effect of breaking the long angiosperm stem branch on estimated ages, with different clocks

PLBP–PLBP*		PLFB–PLFB*		MD–MD*		UCLN–UCLN*	
Node	Difference	Node	Difference	Node	Difference	Node	Difference
4	−27.57	35	31.64	18	−43.68	12	61.63
8	13.19	38	31.58	19	−35	4	−43.06
69	−8.35	34	29.35	17	−32.46	18	−36.24
26	−6.27	47	−25.27	20	−30.8	45	23.89
34	−5.87	42	24.32	4	−30.77	20	−23.41
39	−5.23	52	24.29	11	−29.82	69	22.46
28	−5.13	39	24.23	16	−27.21	42	−19.53
7	5.07	41	23.79	26	−26.86	19	−18.08
56	4.46	53	21.55	27	−24.87	7	−18.03
47	−4.15	54	20.08	21	−23.41	43	−16.71

Note: The 10 largest differences for each clock are listed. Positive differences indicate that breaking the long branch resulted in a younger age. Node numbers correspond to Figure 1. PLBP = penalized likelihood with branch-pruning rate smoothing; PLFB = penalized likelihood with fossil-based rate smoothing; MD = multidivtime; UCLN = uncorrelated lognormal; * = breaking the angiosperm stem branch.

branch resulted in an older mean age for the angiosperm crown node.

PL dating with fossil-based rate smoothing.—The PLFB* analysis provided the youngest mean age for about half of the nodes in the tree, however, the PLFB analysis usually produced ages intermediate among those of other methods (Figs. 4–5). Associated errors range from narrow, particularly in nodes within core eudicots, to wide, especially in nodes within ferns and Pinaceae (Fig. 5).

Breaking the angiosperm stem branch (PLFB*) caused pronounced mean age differences for many nodes. Twenty nodes, including angiosperm backbone nodes, exhibit an absolute difference greater than 10 Myr (Table 1). Slightly over half of the nodes (53.2%) obtained younger mean ages. In spite of the large absolute differences between PLFB and PLFB* ages, the mean age estimated by one analysis always fell within the associated error of the other. Breaking the angiosperm stem branch provided a mean age 29.3 Myr younger for the angiosperm crown node (231.0 and 260.3 Ma old with PLFB* and PLFB, respectively). This is the largest difference obtained with the examined methods.

Multidivtime dating.—Mean ages estimated in MD and MD* analyses are usually intermediate among those estimated by other methods (Figs. 4–5). Associated errors range from wide, especially in nodes within conifers and ferns, to narrow, particularly within core eudicots (Fig. 5).

Breaking the angiosperm stem branch (MD*) resulted in mean age differences greater than 10 Myr for 23 nodes and in younger mean ages for 30.3% of the nodes. Nodes showing the largest absolute differences belong to Pinaceae and gnetophytes (Table 1). Associated errors from the 2 analyses overlap. Breaking the angiosperm stem branch resulted in a mean age 11.5 Myr younger for the angiosperm crown node (241.7 and 253.2 Ma old for MD* and MD, respectively).

Uncorrelated lognormal dating.—The UCLN and UCLN* analyses usually provided the oldest mean ages among the examined methods (Figs. 4–5). Associated errors are very large, except for nodes within core eudicots (Fig. 5). Breaking the angiosperm stem branch resulted in an absolute age difference greater than 10 Myr for 21 nodes and provided younger mean ages for 31.3% of the nodes. Nodes with the largest observed differences belong to ferns, conifers, and angiosperms (Table 1). Associated errors from the 2 analyses are widely overlapping. Breaking the angiosperm stem branch (UCLN*) resulted in a negligibly younger mean age (0.6 Myr) for the angiosperm crown node (274.4 and 275.0 Ma old for UCLN* and UCLN, respectively).

DISCUSSION

Seed Plant Rooting, Angiosperm Stem Relatives, and Branch Lengths

Seed plant relationships have proven a difficult phylogeny estimation problem due, at least, to frequent extinction and possibly to ancient rapid diversifications. The tree obtained in the total evidence analysis is similar to molecular phylogenies in finding a close relationship between conifers and gnetophytes but differs in placing the root of crown seed plants on the branch between cycads and a (*Ginkgo*, conifers, gnetophytes) clade, resulting in the sister group relationship of cycads and angiosperms.

Different data and methods of analysis have provided substantially different relationships among seed plants (for recent reviews, see Doyle 2008; Mathews 2009). Although many analyses of nucleotide sequences root crown seed plants on the branch between angiosperms and living gymnosperms, it seems far from clear that an accurate picture of seed plant relationships has been obtained (e.g., Burleigh and Mathews 2007, p. 134; Doyle 2008, p. 818). A rooting of crown seed plants as found here in the total evidence analysis is unusual but not unknown. It was previously obtained in ML analyses of *psaA* plus *psbB* (all codon positions and third codon positions; Magallón and Sanderson 2002, Fig. 4b,c); in experimental analyses using PHYN/A duplicated genes (Mathews 2009, Fig. 4c, pp. 233–234); and in ML analyses of amino acid sequences of 3 paralog sets of PHY genes (Mathews et al. 2010, Fig. 2). Parsimony morphological analyses of living and extinct taxa have also rooted crown seed plants between (cycads and angiosperms) and (*Ginkgo*, conifers and gnetophytes) in trees one step longer than the MPTs (Doyle 2006, Fig. 7) and in one island of MPTs (Doyle 2008, Fig. 3c). Cycads and angiosperms were strongly supported as sister taxa in an ML analysis of 15–17 plastid genes and associated noncoding regions (Rai et al. 2008, Fig. 2). The rooting of seed plants, and the identity of the closest relatives of angiosperms, will very likely remain difficult questions (Burleigh and Mathews 2004). Previous authors have discussed that a rooting of crown seed plants along the internode separating cycads from the (*Ginkgo*, conifers, gnetophytes) clade may be difficult to detect because this branch is very short, especially in comparison with the long branches subtending angiosperms and gnetophytes (Donoghue and Doyle 2000, p. R108; Burleigh and Mathews 2004, p. 1611).

The fossils here identified as angiosperm stem relatives have been found to be phylogenetically close to angiosperms in previous analyses of living and extinct seed plants. Most analyses have suggested a close relationship between angiosperms and gnetophytes, within an anthophyte clade that also includes the extinct Bennettitales and *Pentoxylon* (e.g., Crane 1985; Doyle and Donoghue 1986; see review in Donoghue and Doyle 2000). Several analyses have found a close relationship of *Caytonia* and Glossopteridales with the anthophytes (e.g., Crane 1985, Fig. 22; Albert et al. 1994, Fig. 6a; Doyle 1996, Fig. 5). More recently, the placement of *Caytonia*, Bennettitales, and Glossopteridales plus *Pentoxylon* as successive sister groups of angiosperms, with gnetophytes distantly related, was found by Doyle (2006, trees one step longer than the MPTs; Fig. 7), Hilton and Bateman (2006, Fig. 4), and Doyle (2008, one island of MPTs; Fig. 3c). Nevertheless, these analyses, and the total evidence analysis presented here, are not entirely independent because all were based on the same morphological data matrix (Doyle 2006), modified to different degrees.

Branch lengths of angiosperm stem relatives were based on the mean path length between the root and terminals of crown seed plants, trimmed considering extinction times. This strategy requires at least the following assumptions: 1) the probability of molecular change during the evolution of angiosperm stem relatives was average among seed plants; 2) substitution rate was constant along branches leading to stem relatives; 3) an age of 350 Ma is an accurate estimate for the seed plant crown node; and 4) the times of origin and extinction of angiosperm stem relatives correspond approximately to their stratigraphic ranges. These assumptions are highly tentative. Experiments allowing different molecular rates and extending the time of existence of angiosperm stem relatives beyond their stratigraphic ranges should provide a broader perspective about the utility of stem relatives in divergence time estimation.

Rate Smoothing Model Selection

Cross-validation methods to identify the optimal rate smoothing in PL dating analysis rely on the difference between observed and estimated lengths of terminal branches or on the summed magnitude of violations to fossil-derived age constraints (Sanderson 2006). The fossil-based cross-validation is attractive because it relies on independent temporal information to estimate the amount of rate heterogeneity in the data. However, the number of fossils and the accurate inference of their phylogenetic position and age become crucial in cross-validation. The sensitivity of fossil-based cross-validations to erroneous fossil phylogenetic placement and age estimation requires explicit evaluation. The utility of different phylogenetic reconstruction methods for living and extinct taxa (e.g., Magallón 2007; Manos et al. 2007) becomes relevant. Because of the burden imposed on the correct phylogenetic placement of fossils in the tree, criteria to assign the position of a fossil in a molecular phylogeny should be critically revised and, preferably, avoided in favor of explicit phylogenetic analysis.

In this study, the branch-pruning and fossil-based cross-validations provided substantially different optimal smoothing magnitudes for the same data. In spite of the empirical ambiguity in identifying the optimal smoothing in the fossil-based cross-validation, it became clear that substitution rate is highly heterogeneous and that data are best explained by very low smoothing. The net difference in errors associated with smoothing magnitudes lower than $\log_{10}\lambda = 2$ is small, and particularly, differences below $\log_{10}\lambda = -0.8$ are very small (Fig. 3b). Regardless of the exact magnitude of the best-fitting smoothing parameter, using fossil-based constraints as a cross-validation criterion requires a much greater amount of rate heterogeneity to explain the data, in comparison with the branch-pruning cross-validation (see also Magallón and Sanderson 2005).

Contrasting behaviors and different optimal rate smoothing magnitudes derived from branch-pruning and fossil-based cross-validations were also found by Near and Sanderson (2004). Experimental simulations and further empirical studies may provide insights about which cross-validation procedure, and under what circumstances, offers the best chance to correctly identify the best-fitting smoothing magnitude. The need for fossil-derived multiple calibrations and constraints in molecular dating is well understood and has been abundantly justified (e.g., Lee 1999; Smith and Peterson 2002; Thorne and Kishino 2002; Near and Sanderson 2004; Magallón and Sanderson 2005; Rannala and Yang 2007). Fossil-based model selection requires a discerning choice of the fossils to include in molecular dating. A fossil should ideally be a structure with distinctive attributes that allow unequivocal phylogenetic placement and, by being abundantly produced and easily preserved, with good chances of becoming fossilized soon after phylogenetic differentiation (or at least, after the origin of a distinctive morphological attribute; Magallón

2004). Additionally, accurate information about its absolute age, derived from reliable stratigraphic correlations or radiometric dating, is necessary.

Different Relaxed-Clock Methods

PL, MD, and UCLN are among the recently available approaches to estimate divergence times and molecular rates without assuming molecular rate constancy. The 3 clocks encompass different statistical foundations and implementations of among-lineage rate heterogeneity. PL and MD rely on the assumption that rates across the tree are autocorrelated (Gillespie 1991); therefore, large differences among adjacent branches receive low probabilities. The UCLN method does not rely on autocorrelation. Instead, the rate of each branch is drawn independently from a given distribution (Drummond et al. 2006), in this case, a lognormal distribution. Multidivtime and UCLN are strongly parametric implementations that use MCMCs to model parameters on the basis of prior assumptions, expressed as probability distributions. PL is a semiparametric method that combines a model that optimizes rates on phylogram branches, with a numerical function that penalizes large rate changes. Therefore, MD shares with UCLN a strongly parametric Bayesian foundation and with PL, rate autocorrelation. However, in this study, MD used gene-partitioned data, whereas PL and UCLN used codon position–partitioned data. The UCLN method differs from the other 2 in representing an implementation of "relaxed" phylogenetics, where phylogeny, divergence times, and molecular rates are simultaneously estimated (Drummond et al. 2006).

The type of age estimate and associated error that each of the 3 clocks provides are not equivalent. A PL analysis provides a point estimate of the age of each node in the tree and no direct measure of its associated error. In this study, PL analyses (PLBP and PLFB) were performed on 100 Bayesian topologically identical phylograms, from which the mean and standard deviation of the age of each node were obtained. Multidivtime provides the mean and standard deviation of age obtained from trees sampled in MCMCs. The UCLN method provides the mean and the minimal and maximal limits of the 95% HPD of MCMC sampled trees, which resulted in associated errors much larger than those provided by the other methods. Only errors obtained from PLBP and PLFB can be directly compared. Errors from PLFB are appreciably larger than those from PLBP, possibly as a consequence of the greater rate heterogeneity allowed in the former. Some nodes, however, consistently display the largest associated errors, regardless of the method. These include the MRCA of *Marsilea* and *Asplenium* (Node 7), the MRCA of *Adiantum* and *Asplenium* (Node 8), and the MRCA of *Ophioglossum* and *Psilotum* (Node 4), within ferns; the MRCA of *Encephalartos* and *Zamia* (Node 12), within cycads; and the Pinaceae crown node (Node 18) and the MRCA of *Abies* and *Cedrus* (Node 19), within conifers.

Although some nodes received substantially different ages from different clocks, a remarkable positive correlation among ages derived from the 4 methods was found, whether including or breaking the angiosperm stem branch (Fig. 6). All pairwise comparisons resulted in a correlation coefficient (r) >0.90, and >0.95 in 7 of 12 comparisons. The slopes of regression lines (s) in pairwise comparisons involving PLBP, PLFB, and MD analyses are close to 1.0 (0.93–1.0), but regression lines of analyses involving UCLN analyses show a greater departure (0.85–0.88). The UCLN method appears as the most deviant clock, usually estimating ages older than those provided by the other methods. The PLFB–MD comparisons showed the highest correlations and regression line slopes closest to 1.0 (Fig. 6). The MD and MD* analyses, although providing the largest differences when the angiosperm stem branch was broken, appeared as least deviant, as they exhibited the highest correlations with other methods.

Breaking the Long Angiosperm Stem Branch and Angiosperm Age

Breaking the angiosperm stem branch resulted in younger ages in 39% of the comparisons (100 nodes) and in older ages in 41% of the comparisons (156 nodes; 11 comparisons had a difference of zero). The mean magnitudes of positive differences and negative differences (i.e., for a given node, breaking the long branch resulted in a younger age or in an older age, respectively) across all comparisons are very similar (7.68 and −7.93 Myr, respectively). Nodes that were particularly sensitive to breaking the angiosperm stem branch included the angiosperm crown node (with some methods), the MRCA of *Psilotum* and *Ophioglossum* (Node 4) within ferns; and the Pinaceae crown node (Node 18) and the MRCA of *Picea* and *Pinus thunbergii* (Node 20) within conifers (Table 1).

Estimates of angiosperm crown node age range from 215.6 to 275.0 Ma old (including the long stem branch) and from 221.47 to 274.43 Ma old (breaking the stem branch). The youngest estimates were provided by PLBP and PLBP* and the oldest by UCLN and UCLN*. These estimates imply that the diversification that lead to living angiosperm species began sometime between the Upper Triassic and the early Permian.

In 3 of 4 comparisons, breaking the angiosperm stem branch yielded younger crown angiosperm ages, but the magnitude of the difference ranged from substantial (29.4 Myr in PLFB–PLFB*) to small (0.56 Myr in UCLN–UCLN*). Breaking the angiosperm stem branch with PLFB* also resulted in substantially younger ages for angiosperm backbone nodes. Although breaking the angiosperm stem branch with PLBP* resulted in an older age for the angiosperm crown node, PLBP and PLBP* yielded the youngest ages overall for this node.

Summary and Conclusions

The effect of a long branch in molecular dating was here explored by introducing simulated sequences of

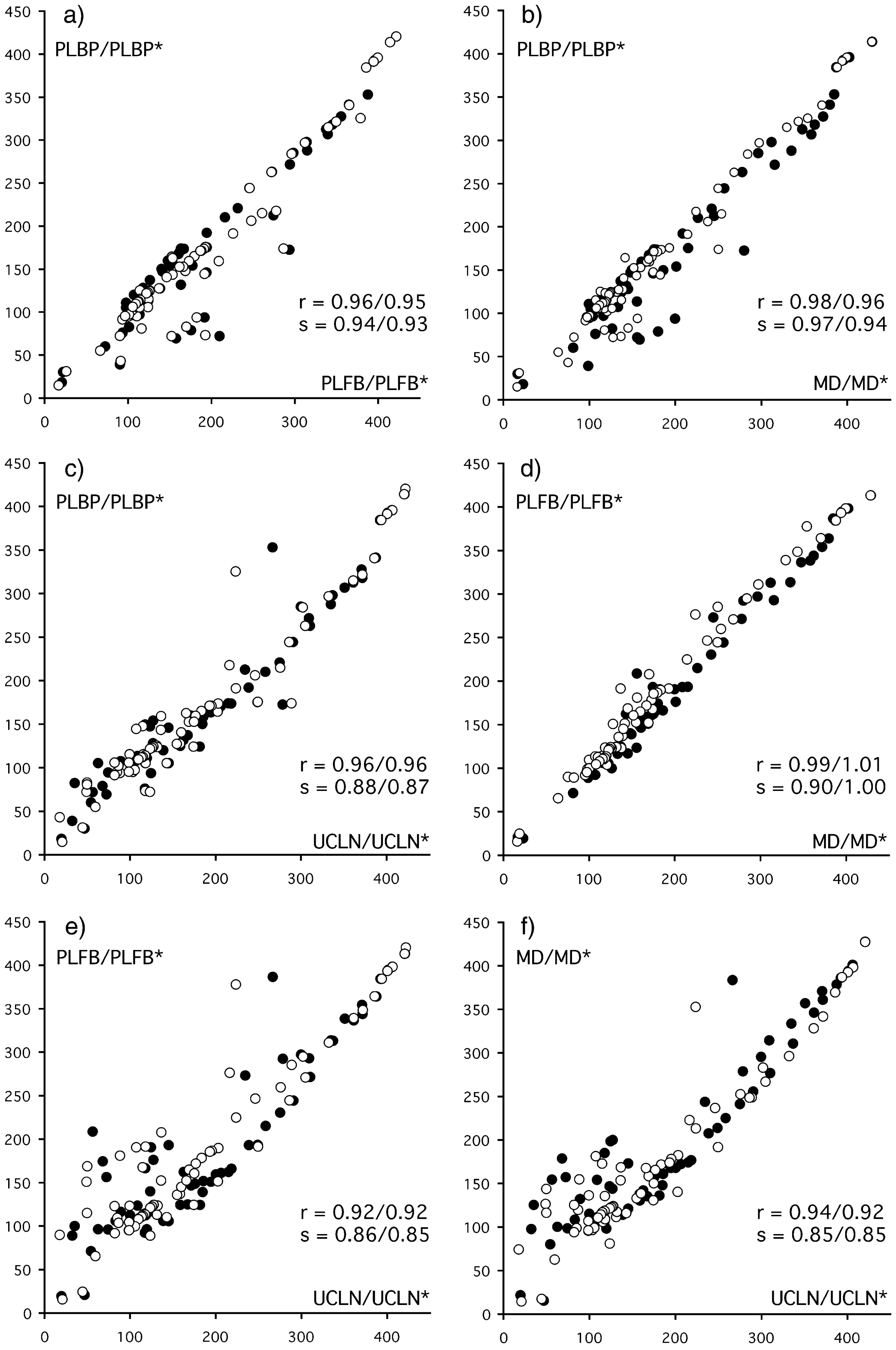

FIGURE 6. Correlation among dating methods for a) PLBP versus PLFB; b) PLBP versus MD; c) PLBP versus UCLN; d) PLFB versus MD; e) PLFB versus UCLN; and f) MD versus UCLN. Pairwise correlations between ages obtained by different dating methods including the angiosperm stem branch (white dots) and breaking the angiosperm stem branch (black dots). For each comparison, Pearson's correlation coefficient (*r*) and slope of the regression line (*s*) are indicated. PLBP = penalized likelihood with branch-pruning rate smoothing; PLFB = penalized likelihood with fossil-based rate smoothing; MD = multidivtime; UCLN = uncorrelated lognormal; * = breaking the angiosperm stem branch.

fossils in analyses using different relaxed clocks. Breaking the long angiosperm stem branch resulted in a substantially younger age for the angiosperm crown node only with 1 of the 4 dating methods. Hence, these results do not support the hypothesis that the long angiosperm stem branch is causally responsible for the old molecular estimates of crown angiosperm age or for its discrepancy with fossil dates. In general, breaking the angiosperm stem branch did not cause substantial age differences across the tree, except for a few particularly sensitive nodes, some of which are phylogenetically distant from the angiosperm crown node.

The angiosperm crown node is sensitive to different relaxed clocks and, with some clocks, to breaking its long subtending branch. There is a 60 Myr difference between the oldest and the youngest estimates, spanning from the Upper Triassic to the early Permian. All these estimates are much older than the oldest angiosperm fossils.

The fact that a strong age correlation among different methods was found is an interesting result, especially considering that the clocks are based on different statistical foundations and methodological implementations. Most notably, they rely on different sources of data and assumptions to assign among-lineage rate heterogeneity: prediction of length of terminal branches (PLBP); violations to fossil ages (PLFB); PP estimation under rate autocorrelation (MD); and uncorrelated PP estimation (UCLN). A possible interpretation of the observed agreement is that the different methods converged on the same (presumably correct) ages. An alternative is that minimal and maximal age constraints bind different methods into providing similar ages (discussed below). This question needs to be explicitly explored by using different clocks without imposing internal temporal constraints.

Correctly accounting for among-lineage rate heterogeneity is crucially important in molecular dating. The 2 procedures for identifying the best-fitting rate smoothing for PL (Sanderson 2006) behaved differently and provided substantially different best-fit values for the same data. The 2 procedures use different parts of the data, each subject to particular types of biases and errors. Which cross-validation procedure, and under what circumstances, has greater chances of accurately identifying the best-fitting rate smoothing needs to be examined through simulations and rigorous experimentation.

Age estimation for several nodes was sensitive to different dating methods and to breaking the angiosperm stem branch, even though some of these nodes were phylogenetically distant from angiosperms. These nodes also exhibit the widest associated errors. These sensitive nodes are weakly constrained by hard temporal bounds (i.e., the calibration node and the present) and by minimal or maximal age constraints. The contrary occurs in nodes within the eudicot clade, which are strongly constrained by the maximal age imposed to crown eudicots, by several minimal ages, and by the present. The ages of these nodes have narrow associated errors and are immune to different dating methods or to the length of the angiosperm stem branch. These observations circumstantially support the possibility that temporal constraints greatly influenced different clocks into estimating similar ages across the tree.

Calculating divergence times in the absence of independent information about the timing of phylogenetic branching or rates of molecular substitution is an especially complex problem. Welch and Bromham (2005) considered that most currently available relaxed clocks assume either a few large changes of rate, for example, ML-based local clocks (e.g., Yoder and Yang 2000; Yang 2004), or many small changes across the tree, for example, the relaxed clocks used in this study. The difficulty in estimating the age of crown angiosperms probably lies in the combination of several complicating factors, including a large amount of extinct historical diversity, rapid radiations during seed plant and angiosperm evolution, and possibly substantial molecular rate heterogeneity among lineages and through time. Allowing relatively small changes in rate probably accounts for only a fraction of the rate heterogeneity existing in this phylogenetic tree. A more appropriate model might combine a background moderate rate heterogeneity with large changes in rate at particular points in the tree. Using the fossil record to introduce temporal constraints on nodes has been recognized as crucially important toward better dating. Analogously to the use of independent information to indicate sites in a tree where rate changes have occurred (e.g., Uyenoyama 1995), paleobiology could perhaps provide macroevolutionary guidelines to inform where in a tree large substitutional changes might have occurred. How exactly paleobiology can inform about molecular rates is an open question.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Acknowledgments

I thank J. Doyle, G. Upchurch, S. McLoughlin, and P. Crane for providing information and references about stratigraphic ranges of fossil seed plants; H. Sauquet for suggestions about the use of BEAST; A. de Nova for processing GenBank submissions; S. Mathews, S. Smith and collaborators, and C. Bell and collaborators, for sharing unpublished manuscripts; and M. Sanderson for confirming the impression that very little is known about the effect of branch lengths on molecular dating. M. Fishbein, A. Mast, and 2 anonymous reviewers are thanked for many relevant observations that substantially improved the final manuscript.

References

Albert V.A., Backlund A., Bremer K., Chase M.W., Manhart J.R., Mishler B.D., Nixon K.C. 1994. Functional constraints and *rbcL* evidence for land plant phylogeny. Ann. Mo. Bot. Gard. 81:534–567.

Anderson F.E., Swofford D.L. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol. Phylogenet. Evol. 33:440–451.

Aris-Brosou S. 2003. Least and most powerful phylogenetic test to elucidate the origin of seed plants in the presence of conflicting signals under misspecified models. Syst. Biol. 52:781–793.

Bell C.D., Soltis D.E., Soltis P.S. 2005. The age of the angiosperms: a molecular timescale without a clock. Evolution. 59:1245–1258.

Burleigh J.G., Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Burleigh J.G., Mathews S. 2007. Assessing systematic error in the inference of seed plant phylogeny. Int. J. Plant Sci. 168:125–135.

Cantino P.D., Doyle J.A., Graham S.W., Judd W.S., Olmstead R.G., Soltis D.E., Soltis P.S., Donoghue M.J. 2007. Towards a phylogenetic nomenclature of Tracheophyta. Taxon. 56:822–846.

Chase M.W., Soltis D.E., Olmstead R.G., Morgan D., Les R.H., Mishler B.D., Duvall M.R., Price R.A., Hills H.G., Qiu Y.L., Kron K.A., Rettig J.H., Conti E., Palmer J.D., Manhart J.R., Sytsma K.J., Michaels H.J., Kress W.J., Karol K.G., Clark W.D., Hedrén M., Gaut B.S., Jansen R.K., Kim K.J., Wimpee C.F., Smith J.F., Furnier G.R., Strauss S.H., Xiang Q.Y., Plunkett G.M., Soltis P.S., Swensen S.M., Williams S.E., Gadek P.A., Quinn C.J., Eguiarte L.E., Golenberg E., Learn G.H. Jr., Graham S.W., Barrett S.C.H., Dayanandan S., Albert V.A. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann. Mo. Bot. Gard. 80:528–580.

Crane P.R. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann. Mo. Bot. Gard. 72:716–793.

Donoghue M.J., Doyle J.A. 2000. Seed plant phylogeny: demise of the anthophyte hypothesis? Curr. Biol. 10:R106–R109.

Doyle J.A. 1992. Revised palynological correlations of the lower Potomac Group (USA) and the Cocobeach sequence of Gabon (Barremian-Aptian). Cret. Res. 13:337–349.

Doyle J.A. 1996. Seed plant phylogeny and the relationships of Gnetales. Int. J. Plant Sci. 157(Suppl):S3–S39.

Doyle J.A. 2000. Paleobotany, relationships, and geographic history of Winteraceae. Ann. Mo. Bot. Gard. 87:303–316.

Doyle J.A. 2006. Seed ferns and the origin of angiosperms. J. Torrey Bot. Soc. 133:169–209.

Doyle J.A. 2008. Integrating molecular phylogenetic and paleobotanical evidence on origin of the flower. Int. J. Plant Sci. 169:816–843.

Doyle J.A., Donoghue M.J. 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. Bot. Rev. 52:321–431.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4(5):e88.

Eklund H., Doyle J.A., Herendeen P.S. 2004. Morphological phylogenetic analysis of living and fossil Chloranthaceae. Int. J. Plant Sci. 165:107–151.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Foote M., Hunter J.P., Janis C.M., Sepkoski J.J. Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of Eutherian mammals. Science. 283:1310–1314.

Friis E.M., Crane P.R., Pedersen K.R. 1997. Fossil history of magnoliid angiosperms. In: Iwatsuki K., Raven P.H., editors. Evolution and diversification of land plants. Tokyo (Japan): Springer-Verlag. p. 121–156.

Friis E.M., Pedersen K.R., Crane P.R. 2001. Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. Nature. 410:357–360.

Friis E.M., Pedersen K.R., Crane P.R. 2004. Araceae from the Early Cretaceous of Portugal: evidence on the emergence of monocotyledons. Proc. Natl. Acad. Sci. USA. 101:16565–16570.

Friis E.M., Pedersen K.R., Crane P.R. 2006. Cretaceous angiosperm flowers: innovation and evolution in plant reproduction. Palaeogeogr. Palaeoclimatol. Palaeoecol. 232:251–293.

Garrat M.J., Rickards R.B. 1987. Pridoli (Silurian) graptolites in association with *Baragwanathia* (Lycophytina). Bull. Geol. Soc. Denmark. 35:135–139.

Gillespie J.H. 1991. The causes of molecular evolution. Oxford: Oxford University Press.

Gradstein F.M., Ogg J.G. 2004. Geologic time scale 2004—why, how and where next. Lethaia. 37:175–181.

Graham S.W., Olmstead R.G. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. 87:1712–1730.

Hendy M.D., Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hilton J., Bateman R.M. 2006. Pteridosperms are the backbone of seed plant evolution. J. Torrey Bot. Soc. 133:119–168.

Hoot S.B., Culham A., Crane P.R. 1995. The utility of *atpB* gene sequences in resolving phylogenetic relationships: comparison with *rbcL* and 18S ribosomal DNA sequences in the Lardizabalaceae. Ann. Mo. Bot. Gard. 82:194–207.

Hueber F.M. 1992. Thoughts on the early lycopsids and zosterophylls. Ann. Mo. Bot. Gard. 79:474–499.

Huelsenbeck J.P., Ronquist F.R. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Hughes N.F. 1994. The enigma of angiosperm origins. Cambridge: Cambridge University Press. p. 1–303.

Hughes N.F., McDougall A.B. 1987. Records of angiospermid pollen entry into the English Early Cretaceous succession. Rev. Palaeobot. Palynol. 50:255–272.

Hughes N.F., McDougall A.B., Chapman J.L. 1991. Exceptional new record of Cretaceous Hauterivian angiospermid pollen from southern England. J. Micropalaeontol. 10:75–82.

Kenrick P., Crane P.R. 1997. The origin and early diversification of land plants—a cladistic study. Washington (DC): Smithsonian Institution Press.

Kishino H., Thorne J.L., Bruno W.J. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18:352–361.

Langley C.L., Fitch W.M. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Lee M.S.Y. 1999. Molecular clock calibrations and metazoan divergence dates. J. Mol. Evol. 49:385–391.

Leng Q., Friis E.M. 2003. *Sinocarpus decussatus* gen. et sp. nov., a new angiosperm with basally syncarpous fruits from the Yixian Formation of northeast China. Plant Syst. Evol. 241:77–88.

Magallón S. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. Int. J. Plant Sci. 165: S7–S21.

Magallón S. 2007. From fossils to molecules: phylogeny and the core eudicot floral groundplan in Hamamelidoideae (Hamamelidaceae, Saxifragales). Syst. Bot. 32:317–347.

Magallón S., Sanderson M.J. 2002. Relationships among seed plant inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. Am. J. Bot. 89:1991–2006.

Magallón S., Sanderson M.J. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution. 59:1653–1670.

Manos S., Soltis P.S., Soltis D.E., Manchester S.R., Oh S.-H., Bell C.D., Dilcher D.L., Stone D.E. 2007. Phylogeny of extant and fossil Juglandaceae inferred from the integration of molecular and morphological data sets. Syst. Biol. 56:412–430.

Martin W., Gierl A., Saedler H. 1989. Molecular evidence for pre-Cretaceous angiosperm origins. Nature. 339:46–48.

Mathews S. 2009. Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. Am. J. Bot. 96: 228–236.

Mathews S., Clements M.D., Beilstein M.A. 2010. A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants. Philos. Trans. R. Soc. B. 365:383–395.

Moore M.J., Bell C.D., Soltis P.S., Soltis D.E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. USA. 104:19363–19368.

Near T.J., Sanderson M.J. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. Philos. Trans. R. Soc. Lond. B. 359:1477–1483.

Nylander A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 24:581–583.

Posada D., Buckley T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Posada D., Crandall K.A. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Rai H.S., Reeves P.A., Peakall R., Olmstead R.G., Graham S.W. 2008. Inference of higher-order conifer relationships from a multilocus plastid data set. Can. J. Bot. 86:658–669.

Rambaut A., Drummond A.J. 2007. Tracer v1.4 2003–2007. MCMC Trace Analysis Package. Available from: http://tree.bio.ed.ac.uk/software/tracer/.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Ramshaw J.A.M., Richardson D.L., Meatyard B.T., Brown R.H., Richardson M., Thompson E.W., Boulter D. 1972. The time of origin of the flowering plants determined by using amino acid sequence data of cytochrome c. New Phytol. 71:773–779.

Rannala B., Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56:453–466.

Rothwell G.W., Scheckler S.E. 1988. Biology of ancestral gymnosperms. In: Beck C.B., editor. Origin and evolution of gymnosperms. New York: Columbia University Press. p. 85–134.

Rothwell G.W., Scheckler S.E., Gillespie W.H. 1989. *Elkinsia* gen. nov., a late Devonian gymnosperm with cupulate ovules. Bot. Gaz. 150:170–189.

Sanderson M.J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.

Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Sanderson M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 19:301–302.

Sanderson M.J. 2006. r8s version 1.71. Analysis of rates (“r8s”) of evolution. Section of Evolution and Ecology, University of California, Davis. Available from: http://loco.biosci.arizona.edu/r8s/.

Sanderson M.J., Doyle J.A. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. Am. J. Bot. 88:1499–1516.

Sanderson M.J., Wojciechowski M.F., Hu J.M., Sher Khan T., Brady S.G. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17:782–797.

Savolainen V., Chase M.W., Hoot S.B., Morton C.M., Soltis D.E., Beyer C., Fay M.F., De Bruijn A.Y., Sullivan S., Qiu Y.L. 2000. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. Syst. Biol. 49:306–362.

Smith A.B., Peterson K.J. 2002. Dating the time of origin of major clades: molecular clocks and fossil record. Annu. Rev. Earth Planet. Sci. 30:65–88.

Soltis D.E., Soltis P.S., Zanis M.J. 2002. Phylogeny of seed plants based on evidence from eight genes. Am. J. Bot. 89:1670–1681.

Soltis P.S., Soltis D.E., Chase M.W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature. 402:402–404.

Swofford D.L. 2002. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0b 10 for Macintosh and 4.0b 10 for UNIX. Sunderland (MA): Sinauer Associates.

Taylor T.N., Taylor E.L. 1993. The biology and evolution of fossil plants. Englewood Cliffs (NJ): Prentice Hall.

Thorne J.L., Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689–702.

Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Tims J.D., Chambers T.C. 1984. Rhyniophytina and Trimerophytina from the early land flora of Victoria, Australia. Palaeontology. 27:265–279.

Uyenoyama M. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics. 139:975–992.

Welch J.J., Bromham L. 2005. Molecular dating when rates vary. Trends Ecol. Evol. 20:320–327.

Wikström N., Savolainen V., Chase M.W. 2001. Evolution of the angiosperms: calibrating the family tree. Proc. R. Soc. Lond. B. 268:1–10.

Wilgenbusch J.C., Warren D.L., Swofford D.L. 2004. AWTY: a system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. Available from: http://ceb.csit.fsu.edu/awty.

Yang Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. Acta Zoologica Sinica. 50:645–656.

Yoder A.D., Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.