Heterogeneous rates of molecular evolution and diversification could explain the
Triassic age estimate for angiosperms

Jeremy M. Beaulieu\textsuperscript{1,*}, Brian O'Meara\textsuperscript{2}, Peter Crane\textsuperscript{3}, and Michael J. Donoghue\textsuperscript{4}

\textsuperscript{1}National Institute for Mathematical & Biological Synthesis (NIMBioS), University of Tennessee, 1122 Volunteer Boulevard, Knoxville, TN 37996 USA; \textsuperscript{2}Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996-1610, USA; \textsuperscript{3}School of Forestry and Environmental Studies, Yale University, 195 Prospect Street, New Haven, CT 06511, USA; and \textsuperscript{4}Department of Ecology and Evolutionary Biology, Yale University, P.O. Box 208106, New Haven, CT 10620 USA

*Correspondence to be sent to: jbeaulieu@nimbios.org
Abstract – Dating analyses based on molecular data imply that crown angiosperms existed in the Triassic, long before their undisputed appearance in the fossil record in the Early Cretaceous. Following a re-analysis of the age of angiosperms using updated sequences and fossil calibrations, we use a series of simulations to explore the possibility that the older age estimates are a consequence of (i) major shifts in the rate of sequence evolution near the base of the angiosperms and/or (ii) the representative taxon sampling strategy employed in such studies. We show that both of these factors do tend to yield substantially older age estimates. These analyses do not prove that younger age estimates based on the fossil record are correct, but they do suggest caution in accepting the older age estimates obtained using current relaxed-clock methods. Although we have focused here on the angiosperms, we suspect that these results will shed light on dating discrepancies in other major clades.
Controversy surrounds the sometimes major differences between age estimates for clades based on the fossil record versus molecular clock methods (e.g., metazoans: Peterson et al. 2004; eukaryotes, Parfrey et al. 2011; mammals: Meredith et al. 2011; dos Reis et al. 2012; O’Leary et al. 2013). Flowering plants (angiosperms) provide a classic example. The generally accepted fossil evidence for the existence of the angiosperm crown clade dates to ca. 140 Ma (Valanginian to Hauterivian stages of the Early Cretaceous; Hughes 1994; Brenner 1996; see also reviews reviewed by Friis et al. 2006; 2011), whereas, unless the age of the crown is fixed to reflect the fossil age (Magallón et al. 2015), recent molecular phylogenetic estimates mostly imply that the angiosperm crown existed in the Triassic, some 200 Ma, or even earlier (e.g., Smith et al. 2010; Magallón 2010; Bell et al. 2010; Clark et al. 2011; Zeng et al. 2014). Even though molecular dating methods have steadily increased in complexity – further relaxing the assumption of substitution rate inheritance (e.g., uncorrelated relaxed-clocks; Drummond et al. 2006; Drummond and Rambaut 2007) and treating fossil calibrations as probabilistic priors – the gap between the stratigraphic record and molecular age estimates for angiosperms has remained stubbornly persistent.

Consistency among the several molecular age estimates of a Triassic origin of flowering plants requires explaining a gap in the fossil record of at least 60 Myr. One possibility is that crown angiosperms existed during that time interval, but were not ecologically dominant and/or were living in environments where fossilization was unlikely (c.f., Feild et al. 2004; Smith et al. 2010). Given repeated claims of angiosperm fossils from the Triassic (e.g. Seward 1904; Cornet 1986; Hochuli and Feist-Burkhardt 2004; 2013) such an explanation might seem plausible, but all such claims are disputed.
(although some could represent lineages along the angiosperm stem; Doyle 2012). Of course, if any fossil from the Triassic proved to be a legitimate member of the angiosperm crown, the debate would instantly be settled in favor of the older molecular dates.

On the other hand, it is possible that the truth is more in line with the much younger fossil-based dates, in which case there must be serious methodological issues with current molecular dating methods that have not yet been addressed. This is the topic that we investigate here. Specifically, we focus on the possibility that clade-specific heterogeneity in rates of molecular evolution and the nature of taxon sampling could cause a systematic bias in age estimation using certain relaxed-clock methods. We re-analyze the age of angiosperms using updated sequences and fossil calibrations and then explore potential sources of error in a series of simulations designed to test whether methodological biases might partially explain why molecular-clock studies have consistently yielded much older ages than those suggested by the fossil record.

A RE-ANALYSIS OF THE AGE OF ANGIOSPERMS

We developed a set of 24 fossil calibrations mainly selected from those used by Smith et al. (2010), but supplemented with several additional calibrations proposed by Doyle and Endress (2010). Fifteen of these fossils calibrate nodes within flowering plants, with the remaining fossils providing temporal information in the other land plant clades (see Fig. S1). We also assembled a molecular dataset of published sequences of nuclear 18S and chloroplast $atpB$, $psbB$, and $rbcL$ across land plants. We estimated divergence times using the uncorrelated lognormal (UCLN) clock model implemented in BEAST (Drummond et al. 2006; Drummond and Rambaut 2007). Our dataset includes
representatives of every angiosperm order (*sensu* Angiosperm Phylogeny Group, 2009) and an expanded sample for Nymphaeales, Austrobaileyales, and Magnoliidae to permit a more precise placement of several fossils. It also includes representatives of every major clade of acrogymnosperms (the clade containing the four major extant lineages of non-angiosperm seed plants: conifers, gnetophytes, cycads, ginkgos), seven taxa representing monilophytes (ferns and allies), and a single lycophyte (club moss) to root the tree.

Analytical details, including justification for the placement of each fossil calibration point, and our rationale for assigning prior probabilities, are provided in the Supplemental Materials (http://dx.doi.org/10.5061/dryad.629sc). However, our aim is not to refine previous studies. For present purposes, the important point is that our age estimate for crown angiosperms is consistent with previous results, with a median age of 232 Ma and a 95% HPD of 210-256 Ma (Fig.1). This suggests that the angiosperm crown clade existed in the Triassic, and implies a gap in the fossil record of around 90 million years. Age estimates within several subgroups of angiosperms (e.g., campanulids – a clade that contains the sunflowers and their relatives [Asterales], carrots and their relatives [Apiales], and honeysuckles and relatives [Dipsacales]) are also consistent with previous estimates (Beaulieu et al. 2013).

The potential impact of clade-specific rate heterogeneity

In angiosperms, significant differences in the rate of molecular evolution are correlated with growth habit, with herbaceous clades exhibiting higher and more variable molecular rates than related woody clades (Gaut et al. 1992, 1996; Laroche et al. 1997; Kay et al. 2006; Smith and Donoghue 2008; Lanfear et al. 2013). There is some evidence that relaxed clock methods can perform poorly in the presence of significant among-
lineage rate variation (Wertheim et al. 2012), or when sister clades differ substantially in substitution rates (Dornburg et al. 2012), both of which can lead to overestimation of ages, although in the latter case the effects can be alleviated (at a cost of precision) when there are many calibration points spanning both clades. In the case of angiosperms our worry is that rapidly evolving herbaceous lineages will tend to appear older than they really are, and that this has the potential to "trickle-down" and bias age estimates for nodes below and some distance from the inferred shift in growth habit (also see Smith et al. 2010). Specifically, we wonder whether multiple shifts to the herbaceous habit nested not too far within the angiosperms might have the effect of pushing back the estimated age of the angiosperm crown (Fig. 1).

To explore this possibility, we conducted a set of simulations to test the impact of shifts in molecular substitution rates not far from a node of interest. Using the tree in Figure 1, we fixed the crown age of angiosperms at 140 Ma, based on the fossil record, and then asked how well we recovered this age as we varied the difference in rate of evolution between herbaceous clades and woody ones. Specifically, we supposed for this purpose that four herbaceous angiosperm clades whose stem connect nearer to the root – Nymphaeales, Piperales, Monocotyledonae, and Ceratophyllum – had variously elevated rates of molecular evolution. The “true” ages of nodes in the seed plant portion of our tree for which we have calibrations were set by treating as fixed ages the median value of the prior distribution applied to each fossil calibration (see Supplemental Materials); the ages of all other nodes were obtained by smoothing the molecular rates using r8s, which assumes an autocorrelated model of rate variation (Sanderson, 2002).
Shifts in molecular rate due to life history differences were simulated by independently drawing rates for branches within Nymphaeales, Piperales, monocotyledons, and *Ceratophyllum* from a lognormal distribution that differed from all other branches in the tree (Fig. 2). For this purpose we used the inferred parametric shape of the lognormal distribution of rates from our reanalysis of land plants (mean=5e-4, sd=0.75) as a baseline for increasing the rate of the herbaceous clades to a mean that was initially three times (3x) higher than all other branches in the tree. The 3x rate difference roughly corresponds to the average difference between the woody and herbaceous clades examined by Smith and Donoghue (2008).

Using the fixed tree topology shown in Figure 1, and the associated ages (see above), we generated a set of 100 molecular trees (differing in branch lengths) that were used to simulate gene alignments of 1000 sites in length using SeqGen (Rambaut and Grassly 1997). We assumed a general time reversible model (GTR) of nucleotide substitution using the inferred parameters from our land plant study. Each simulated data set was then analyzed using the UCLN model implemented in BEAST with 7,500,000 generations following the removal of the first 2,500,000 steps as burn-in. We also conducted a complimentary set of analyses that incorporated a calibration prior for the age of crown seed plants (age = 317.0 Ma) from our r8s analysis. We focus our discussion on simulations that exclude this calibration point, but the results were consistent between the two sets of analyses (see Supplemental Materials). All trees were summarized with TreeAnnotator with the consensus ages representing the median estimate of the posterior distribution.
Our primary finding is that when we simulated significant clade-specific rate heterogeneity located not far from the origin of angiosperms, we obtained age estimates for the crown node that are much older than the age assumed in the simulation, despite use of an uncorrelated rate model. As shown in Figure 3, across 100 randomly generated data sets the age of angiosperms was estimated to be, on average, 209.5 Ma, a difference of about 70 Myr from our fixed age of 140 Ma. It is important to note that the impact of rate heterogeneity does not appear to be widely scattered throughout our seed plant tree. Instead, it appears to strongly affect only a few nodes that occur in the general vicinity of the simulated rate shifts (Fig. 3), but this includes the angiosperm crown node. Increasing the rate difference between woody and herbaceous lineages overestimates the age of angiosperms even further. When we repeated the simulations assuming a 6x rate difference (Fig. 4), we estimated the crown age of angiosperms to be 244.3 Ma, on average, with even the youngest ages never falling outside of the Triassic (2.5% quantile=213.9 Ma).

Of course, the situation is likely to be far more complicated than our simple two-rate scenario. For instance, there is evidence that the substitution rate in woody angiosperms is nearly 4x higher than that in woody acrogymnosperms (Buschiazzo et al. 2012), due, at least in part, to the generally shorter times to first reproduction in woody angiosperms (Verdú 2002). Simulated data sets under such a three-rate scenario resulted in a distribution of the inferred age of angiosperms very similar (Fig. 4) to that for the 6x rate difference (Fig. 4)(average at 244.9 Ma [95% CI = 208.7-275.2]).

We performed additional simulations that applied an angiosperm molecular rate distribution to the Gnetales, a small clade of acrogymnosperms that consistently exhibits
THE POTENTIAL IMPACT OF REPRESENTATIVE SAMPLING

The simulations described above implicitly assumed that if there were no clade-specific rate variation the relaxed-clock method would perform well. In other words, if all branches were randomly drawn from a single lognormal distribution of molecular rates, then the age estimate for crown angiosperms would be centered on the “true” age of 140-130 Ma. As a post-hoc check, we ran a set of simulations that assumed the same baseline lognormal distribution of rates for all branches in our seed plant tree. We were surprised to still find a considerable bias, with the age of crown angiosperms estimated to be 188 Ma on average. This is 20 to 60 Ma younger than estimates incorporating rate heterogeneity (see above), but still nearly 50 Myr older than the fixed age used in the simulation. In other words, rate heterogeneity may explain part of the discrepancy between fossil-based and molecule-based estimates, but by no means all of it.

To explore this further we first compared the lengths of each branch in the “true” tree to the inferred branch lengths across the 100 randomly generated datasets that contain no lineage-specific rate heterogeneity (Fig. 5). This revealed that internal branch lengths were consistently inferred to be longer than the “true” branch lengths, while terminal branches were generally shorter. This overestimation and underestimation of internal and terminal branches, respectively, could be viewed as largely compensatory. However, in terms of percent difference from the true branch length, the bias for internal
branches is much larger and can be upwards of nearly 1000% of the true length (Fig. 5). Cumulatively, this has the effect of increasing the total length of the tree, even in the absence of rate heterogeneity.

We suspect that this tendency to push dates back is caused by the way in which extant diversity is sampled in such studies. In our representative sampling of the major land plant lineages, each terminal branch is generally a placeholder from anywhere from a handful to many thousands of species (e.g., Asterales, with ca. 27,000 species, represented here by *Helianthus annuus*; Asparagales, with ca. 26,000 species, represented by *Apostasia stylidlioides*, and Lamiales, with ca. 24,000 species, represented by *Antirrhinum majus*). Perhaps this sort of sampling creates a problem for a dating method that requires a robust estimate of the underlying birth-death process as a means of calibrating rates.

With BEAST, there is a natural tension between whether differences in branch length reflect variation in the rate of molecular evolution or differences in time, with differences in time being inferred from a combination of calibration (from the fossil record) and an estimate of the diversification process, which for the purposes of the current discussion we treat as a birth-death model. By contrast, most other dating methods, such as penalized likelihood (Sanderson 2002), do not rely upon a diversification model that provides a separate signal for the true branch lengths. Under normal circumstances, the estimates of the birth-death parameters should provide an adequate measure of the expected wait times (i.e., \( \frac{1}{b+d} \)) between successive nodes. BEAST should be able to discern whether to shorten relatively long branches to accommodate a birth-death process that expects generally shorter wait times, or to
lengthen relatively short branches to account for longer expected wait times. However, in the case of our tree, and many other studies that employ representative sampling, the wait times in the tree cannot be reflecting some common underlying birth-death process. In fact, such representative sampling insures a particular kind of heterogeneity: one process that generated the internal branch lengths, which should generally better reflect the true expected wait times (assuming, of course, extinction has been random and that all major land lineages have been sampled), and another that would be consistent with the long durations in the terminal branches where, ostensibly, nothing happened (see Fig. 1). Because terminal branches make up more than half the tree, the “average” should be biased toward the longer wait times of the terminal branches. The worry is that, as a result, the internal branches in our seed plant tree are being systematically lengthened.

To explore this issue we conducted a set of simulations where we compared age estimates obtained from a complete tree to those of a representative sample of the same tree (Fig. 6). We generated a single random birth-death tree (birth=0.08, death=0.04) that contained 100 species; the total length of the tree was rescaled to reflect 100 time units. We followed the same procedures described above: we generated a set of 100 molecular trees under our baseline UCLN distribution, brought them into SeqGen (Rambaut and Grassly 1997), and simulated gene alignments of 1000 sites assuming a general time reversible model (GTR) of nucleotide substitution. Each simulated data set was also then selectively pruned down to 10 tips to ensure a single representative of ten major lineages (shown with different colors in Fig. 6). Both the “complete” and the “sampled” data sets were analyzed in BEAST assuming a UCLN model of substitution, a fixed topology, and
calibration priors that were set such that the true ages were the median values in a lognormal distribution (mean=1.5, sd=0.75).

When the “complete” trees were analyzed there was an increase of around 7 Myr in the median age estimate across the 100 datasets. When the “sampled” trees were analyzed this increased to around 24 Myr. This, we think, reflects the great disparity in the lengths of internal and terminal branch lengths in the sampled tree. The impact of this disparity on the birth-death process can be seen in the estimates of net diversification, which were more than three times lower in the sampled (mean=0.009) than in the complete data set (mean=0.032). When the parameters of the birth-death process are converted to an estimate of the expected wait times, the sampled trees show wait times that are, on average (mean=70.0 Myr), nearly an order of magnitude longer than in the complete trees (mean=9.82 Myr).

CONCLUDING REMARKS

What can we conclude about Triassic age estimates for the flowering plants based on molecular data? Unfortunately, it appears that these could largely reflect the methodological issues highlighted here. As our simulations imply, such an age could be largely explained by the potential additive effects of not properly accounting for (1) clade-specific heterogeneity in molecular substitution rates, and (2) the use of a representative sample of the major land plant groups. It is possible, in other words, that crown angiosperms are ca. 140-130 million years old, and that we are obtaining much older age estimates using BEAST owing to these two sources of bias. We are not asserting here that the crown angiosperms are actually only ca. 140 million years old (or
perhaps slightly older), only that the molecular dating analyses conducted to date do not strongly rule out this possibility.

The shifts in plant habit and molecular rate that actually occurred in the course of angiosperm evolution will have been far more complicated than we have represented in our simulations. For example, within the species-rich eudicot clade there are many transitions in habit, but for purposes of our simulations we treated this entire clade as having only evolved under a woody rate. There is also the vexing problem of how lineages that are missing as a result of extinction might complicate the issue. It is unclear what impact having many more nested rate shifts would have on our results. Of course, the impact of rate heterogeneity may also vary depending on the number and the identity of the genes that are analyzed, and more detailed comparisons along these lines would be useful. In any event, what is clear is that even simple scenarios of rate heterogeneity can have a pronounced impact on age estimation.

Regarding representative sampling (both of extant and extinct lineages), we note that BEAST does implement a birth-death prior that assumes incomplete sampling, which could potentially alleviate the problem. With this approach, the normal birth-death parameters are estimated with the addition of a third variable, \( \rho \), that scales estimates of birth and death to compensate for incomplete sampling (Gernhard 2008). We applied this prior to our land plant data set (see Supplemental Materials), but found little change in age estimates for angiosperms. Furthermore, the mean of the posterior distribution for the sampling frequency was \( \rho=0.059 \) (95% HPD=0.001-0.117), which is nowhere near the dismal sampling frequency in our tree (ca. 0.0003% assuming 375,000 species of land plants and no extinction). This is consistent with recent theoretical work demonstrating
that $\rho$ must be considered a known quantity and cannot be estimated from a phylogeny while also estimating speciation and extinction rates (Stadler 2013), as BEAST does. Otherwise, there is a tendency to overestimate the sampling frequency ($\rho$) at the cost of underestimating both speciation and extinction rates.

Even if we could treat sampling frequency as fixed (either through direct specification or by restricting the prior on $\rho$), the assumption of a constant birth-death process would still misrepresent the heterogeneity in diversification across land plants. For example, the success of angiosperms is generally linked to higher overall rates of diversification (e.g., Sanderson and Donoghue 1994; Smith et al. 2011), which would translate into shorter expected wait times within angiosperms relative to other land plants. With the assumption of a single distribution from which waiting times are drawn, rapidly diversifying clades like the angiosperms will appear to be older than they really are when analyzed as a part of a larger clade.

Under the circumstances, it behooves us to remain humble and to honestly assess potential biases not only in the fossil record, but also in our methods for analyzing molecular data. Our simulations highlight two potentially general systematic biases introduced by (1) the phylogenetic position and the magnitude of shifts in rates of molecular evolution, and (2) a representative sampling scheme that can create difficulties for methods that rely on birth-death processes. Although we have focused here on the angiosperms, we suspect that these results will shed light on dating discrepancies in other major clades.
SUPPLEMENTAL MATERIALS

Supplemental material, including data files and online-only appendices, can be found in the Dryad data repository (http://dx.doi.org/10.5061/dryad.629sc)

ACKNOWLEDGEMENTS

We thank Mark Fishbein, Susana Magallón, and an anonymous reviewer for thoughtful comments and suggestions for improving the manuscript. We also thank Jim Doyle, Tom Near, Andrew Leslie, Stephen Smith, Alex Dornburg, and Nick Matzke for helpful discussions. Support for JMB has been provided by the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville.
REFERENCES


Figure legends

Figure 1. Divergence time estimates for land plants from a four-gene analysis (atpB, rbcL, psbA, and 18S), with the inferred Triassic age for angiosperms highlighted (star and light blue box). See Supplemental information for information on fossil calibrations. On the left, we illustrate our concern that herbaceous clades (green branches; often with elevated substitution rates relative to woody plants), positioned not far from the origin of angiosperm, may push back the age of the angiosperm crown.

Figure 2. Basic steps in our simulation of clade-specific rate heterogeneity. We isolated the seed plant portion of our land plant tree and generated a set of “true” ages by treating as fixed ages the median value of the prior distribution applied to each fossil calibration from our reanalysis of land plants, with all other node ages obtained using r8s (see main text). We fixed the age of crown angiosperms (star) to be 140 Ma to reflect the fossil record. Faster molecular rates within the angiosperms (due to the evolution of the herbaceous habit) were simulated by independently drawing rates for the green branches (Nymphaeales, Piperales, Monocotyledonae, and Ceratophyllum) from a lognormal distribution with a mean initially 3x higher than all other (woody) branches in the tree. This process was repeated 100 times, and we analyzed each data set using the uncorrelated lognormal (UCLN) model implemented in BEAST.

Figure 3. Results from the 2-rate simulation (Fig. 2). Across 100 randomly generated data sets, the age of angiosperms was overestimated on average by roughly 70 Myr.
(compare yellow versus gray star). On the right, the chronogram of a randomly selected simulated tree is overlain on the “true” tree (shown in gray).

**Figure 4.** Biases in additional simulations. Above: a 2-rate scenario, where rates for herbaceous clades were drawn from a lognormal distribution 6x higher than woody clades. Below: a 3-rate scenario in which acrogymnosperms (the clade containing the four major clades of extant “gymnosperms”), woody angiosperms, and herbaceous angiosperms have different rates of molecular evolution.

**Figure 5.** Distribution of bias between internal (orange) and terminal (blue) branches assuming that all branches are randomly drawn from a single lognormal distribution (i.e., “no-variation” scenario). Left: Relationship between “true” branch lengths and the mean of the estimated branch lengths from 100 randomly generated data sets (line shows the 95% CI). Right: When viewed as percentage difference, the bias is seen to be in internal branches (red line shows the trend in the rolling mean).

**Figure 6.** Basic steps in our simulation to test the influence of the birth-death prior. 100 sequence data sets were simulated on a random birth-death tree with 100 taxa. Each simulated data set was then pruned to ensure a single representative of the ten major lineages (taxonomic groups shown with different colors). The “complete” and “sampled” data sets were analyzed in BEAST. Yellow boxplots below show an average bias of ca. 7 Myr in the median age estimate for the complete trees, and an average bias of ca. 24 Myr for the sampled trees.