Strategies for Partitioning Clock Models in Phylogenomic Dating: Application to the Angiosperm Evolutionary Timescale

Charles S.P. Foster* and Simon Y.W. Ho

School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales 2006, Australia

***Corresponding author:** E-mail: charles.foster@sydney.edu.au.

Accepted: September 25, 2017

Abstract

Evolutionary timescales can be inferred from molecular sequence data using a Bayesian phylogenetic approach. In these methods, the molecular clock is often calibrated using fossil data. The uncertainty in these fossil calibrations is important because it determines the limiting posterior distribution for divergence-time estimates as the sequence length tends to infinity. Here, we investigate how the accuracy and precision of Bayesian divergence-time estimates improve with the increased clock-partitioning of genome-scale data into clock-subsets. We focus on a data set comprising plastome-scale sequences of 52 angiosperm taxa. There was little difference among the Bayesian date estimates whether we chose clock-subsets based on patterns of among-lineage rate heterogeneity or relative rates across genes, or by random assignment. Increasing the degree of clock-partitioning usually led to an improvement in the precision of divergence-time estimates, but this increase was asymptotic to a limit presumably imposed by fossil calibrations. Our clock-partitioning approaches yielded highly precise age estimates for several key nodes in the angiosperm phylogeny. For example, when partitioning the data into 20 clock-subsets based on patterns of among-lineage rate heterogeneity, we inferred crown angiosperms to have arisen 198–178 Ma. This demonstrates that judicious clock-partitioning can improve the precision of molecular dating based on phylogenomic data, but the meaning of this increased precision should be considered critically.

Key words: Angiospermae, molecular dating, phylogenomics, infinite-sites theory, calibration, rate heterogeneity.

Introduction

Evolutionary timescales can be estimated from molecular sequence data using phylogenetic methods based on the molecular clock. In practice, most data sets exhibit substantial rate heterogeneity among lineages. These "lineage effects" can be caused by variation in life-history traits, generation time, or exposure to mutagens (Smith and Donoghue 2008; Gaut et al. 2011; Lanfear et al. 2013). Among-lineage rate variation can be taken into account using Bayesian relaxed-clock models, in which the rates can be assumed to be either correlated between neighboring branches (Thorne et al. 1998; Kishino et al. 2001) or drawn independently from a chosen distribution (Drummond et al. 2006; Rannala and Yang 2007).

A number of factors can cause rates to vary across loci in the genome (Wolfe et al. 1987). These "gene effects" can be taken into account by allowing each locus to have a distinct relative rate. Less certain is the best way to deal with interactions between gene effects and lineage effects, which can be caused by differences in selective pressure and other processes (Gaut et al. 2011). In this case, the extent and patterns of among-lineage rate heterogeneity vary across genes or other subsets of the data. This form of rate variation can be captured by assigning separate clock models to different subsets of the data (Ho and Duchêne 2014), a process that we refer to here as clock-partitioning.

Appropriate clock-partitioning can improve the precision of Bayesian date estimates (as measured by the associated 95% credibility intervals), but it is rarely done in practice. This is also despite widespread adoption of partitioning schemes for substitution models (Lanfear et al. 2012). The most likely explanation is that the use of clock-partitioning in Bayesian phylogenetics greatly increases the risk of overparameterization, and thus to reduced Markov chain Monte Carlo performance. Overparameterization has been previously addressed in light of the bias-variance trade-off, which is well established in statistical theory (Burnham and Anderson 2003). Compared with a complex, parameter-rich model, a simple model that

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (]>http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

underfits data is expected to have low accuracy (high bias) but high precision (low variance). Conversely, a parameter-rich model that overfits the data is likely to have higher accuracy, but this comes at the cost of reduced precision. The best model is an intermediate one that simultaneously maximizes accuracy and precision (Wertheim et al. 2010)

It is useful to consider the bias-variance trade-off in the context of molecular dating with partitioned clock models. Patterns of among-lineage rate variation are likely to differ across genes (Muse and Gaut 1994), so increasing the number of relaxed clocks will better capture these patterns of rate heterogeneity and should lead to more accurate age estimates (Duchêne and Ho 2014). However, each clock-subset has parameters that need to be estimated, including a distinct set of branch rates. As a consequence, increasing the degree of clock-partitioning should lead to a widening of the posterior distributions of parameters.

Contrary to the expectations of the bias-variance trade-off, increasing the degree of clock-partitioning tends to improve the precision of Bayesian age estimates (Zhu et al. 2015). One possible explanation for this lies in the treatment of the uncertainty in the estimates of genetic branch lengths. The accuracy and precision of evolutionary rate estimates depend on the accurate inference of branch lengths (in substitutions per site). In the case of molecular dating, branch rates for each clock-subset are combined with node times to give the branch lengths. Therefore, as the number of clock-subsets increases, the node times in the chronogram are estimated from an increasing number of data points, leading to increasing precision. Although branch-length estimation generally improves as the amount of sequence data increases, branch lengths can be estimated with reasonable accuracy even with fairly small amounts of sequence data (Yang and Rannala 2006). This suggests that for a data set of a (large) fixed size, increasing the number of clock-subsets should lead to improved precision in divergence-time estimates until the amount of sequence data in each clock-subset decreases to a critical point.

Zhu et al. (2015) explain this phenomenon in their "finite sites" theory, although they use the term "loci" to refer to clock-subsets. Even with sequences of infinite length, there will still be uncertainty in the age estimates, corresponding to the uncertainty in the fossil calibrations ("infinite-data limit"; Yang and Rannala 2006; dos Reis and Yang 2013). As the number of clock-subsets (L) increases, the finite-sites theory suggests that the uncertainty in age estimates decreases to the infinite-data limit at the rate of $1/L$ (Zhu et al. 2015). This property has important consequences for analyses of genome-scale data sets, whereby many genes are analyzed concurrently. Therefore, it is important that both the finite-sites theory and the bias-variance trade-off are tested comprehensively on a genome-scale data set with clock-partitioning.

Persistent uncertainty in molecular date estimates is perhaps best exemplified by studies of the origins of flowering plants (angiosperms) (Foster 2016). The earliest unequivocal angiosperm fossils are tricolpate pollen grains from the Barremian–Aptian boundary, from approximately 125.9 Ma (Hughes 1994). Older pollen grains from the Hauterivian provide some evidence of crown-group angiosperms, and are usually accepted as belonging to this group, albeit with less confidence than for the tricolpate pollen grains (Herendeen et al. 2017). Patterns of diversification in the broader fossil record suggest that angiosperms are unlikely to have arisen much earlier than this time (Magallón et al. 2015; Sauquet et al. 2017). The majority of molecular dating analyses tell a vastly different story, with most recent analyses inferring an origin within the Triassic (Foster et al. 2017). Estimates of the angiosperm evolutionary timescale appear to be largely robust to the source of genetic markers, despite the choice between chloroplast-derived markers or nuclear-derived markers potentially affecting the deep nodes of the angiosperm phylogeny (Wickett et al. 2014; Zeng et al. 2014). However, the uncertainty surrounding the age of the angiosperm crown node is large, often spanning an interval of many tens of millions of years, unless strong age constraints are placed on the node. This uncertainty could be masking any interesting biological processes driving the age estimates for deep nodes. Improving the accuracy and precision of estimates of the age of crown angiosperms thus represents a key goal of molecular dating.

In this study, we use a Bayesian phylogenetic approach to investigate the impact of clock-partitioning on the precision of divergence-time estimates. We also investigate whether the criteria used to assign genes to different clocks has an impact on estimation error. To do so, we infer the evolutionary timescale of angiosperms using a plastome-level data set. In analyses with clock-partitioning schemes comprising up to 20 clock-subsets, we allocate genes to clock-subsets based on patterns of among-lineage rate heterogeneity or relative substitution rate, or through random assignment. In all cases, we confirm that increasing the degree of clock-partitioning can lead to vast improvements in the precision of Bayesian date estimates.

Materials and Methods

Data Sets and Clock-Partitioning

We obtained full chloroplast genome sequences for 52 angiosperm taxa and two gymnosperm outgroup taxa from GenBank (supplementary table S1, Supplementary Material online). Each angiosperm taxon was chosen to represent a different order, with our sampling designed to include as many as possible of the 63 angiosperm orders recognized by the Angiosperm Phylogeny Group (2016). We extracted all 79 protein-coding genes from the chloroplast genomes, although some genes were missing from some taxa. We initially translated all genes into amino acid sequences using VirtualRibosome (Wernersson 2006) and aligned them using

MAFFT v7.305b (Katoh and Standley 2013). We then translated the aligned amino acid sequences back into nucleotide sequence alignments using PAL2NAL (Suyama et al. 2006), made manual adjustments, and filtered out any sites in the alignment at which a gap was present in ≥80% of the taxa. Our total core data set consisted of 68,790 nucleotides, of which only 7.54% sites were gaps or missing data (see supplementary file S1, Supplementary Material online).

Our primary strategy for clock-partitioning based on patterns of among-lineage rate heterogeneity was to analyze the genes using ClockstaR v2 (Duchêne et al. 2014). ClockstaR takes predefined subsets of the data, along with the estimated gene tree for each subset, and determines the optimal clock-partitioning scheme for the data set. This involves identifying the optimal number of clock-subsets (k), as well as the optimal assignment of the data subsets to each of these clock-subsets. We used the partitioning around medoids (PAM) algorithm within ClockstaR for this purpose, which identifies k objects (medoids) that are centrally located within clusters (Kaufman and Rousseeuw 2009). In our case, this strategy identifies groups of genes that have the most similar patterns of among-lineage rate heterogeneity for increasing numbers of clusters (clock-subsets). Comparison of clock-partitioning schemes is done by comparing the patterns of among-lineage rate heterogeneity across the gene trees and clustering the gene trees according to the gap statistic (Gap_k) (Tibshirani et al. 2001). The gap statistic method evaluates the goodness-of-clustering for each value of k by comparing the mean within-cluster dispersion of the data with that of bootstrap reference data sets. Higher values for Gap_k indicate a better statistical fit, and the optimal number of clusters (clock-subsets) is selected as the smallest value of k that yields a peak in Gap_k (Tibshirani et al. 2001). ClockstaR can also determine the optimal clock-partitioning scheme for any value of k. In our case, each of the 79 protein-coding genes was considered as a separate data subset for the ClockstaR analysis.

ClockstaR requires all data subsets to share the same tree topology. Since the chloroplast genome does not typically undergo recombination (Birky 1995), all of its genes should share the same topology. Therefore, we first inferred the phylogeny for the concatenated data set using maximum-likelihood analysis in IQ-TREE v1.50a (Nguyen et al. 2015), with node support estimated using 1000 bootstrap replicates with the ultrafast bootstrapping algorithm (Minh et al. 2013). We partitioned the data set by codon position using the edge-linked partition model (Chernomor et al. 2016), and implemented the $\text{GTR}+\Gamma_4$ model of nucleotide substitution for each subset. The best-scoring tree was very similar to previous estimates of the angiosperm phylogeny based on chloroplast data (Moore et al. 2010; Soltis et al. 2011), and we found strong support for most nodes in the tree (supplementary fig. S1, Supplementary Material online). We used this tree for ClockstaR and optimized the branch lengths for each gene alignment. Finally, we determined the optimal value of k, and then created 12 clock-partitioning schemes using the optimal assignment of genes to clock-subsets for values of k from 1 to 10, 15, and 20 ("P_{CSTAR}" schemes).

As a means of comparison with the ClockstaR partitioning schemes, we also chose clock-partitioning schemes based on relative substitution rates across genes (dos Reis et al. 2012). To do so, we focused on a subset of 20 taxa for which sequences of all 79 protein-coding genes were available (supplementary table S1, Supplementary Material online). We then analyzed each gene using maximum likelihood in IQ-TREE, in each case partitioning by codon position and implementing the $\text{GTR}+\Gamma_4$ model of nucleotide substitution for each codon position. Using the tree lengths as a proxy for the overall substitution rate of each gene, we created 11 partitioning schemes based on relative rates of substitution ("P_{RATE}" schemes), in which we assigned genes to clock-subsets for values of k from 2 to 10, 15, and 20.

For an additional form of comparison, we generated clock-partitioning schemes with genes randomly allocated to clock-subsets. Genes were randomly sampled without replacement in R v3.3.2 (R Core Team 2016) and assigned to clock-subsets for values of k from 2 to 10, 15, and 20. We repeated this process three times, resulting in a total of 33 clock-partitioning schemes in which genes were randomly assigned to clock-subsets ("P_{RAND}" schemes).

Molecular Dating

We inferred the evolutionary timescale using MCMCTREE in PAML v4.8 (Yang 2007) with the $\text{GTR}+\Gamma_4$ model of nucleotide substitution. A key requirement of MCMCTREE is a fixed tree topology, so we used the best-scoring tree that we estimated from the total concatenated data set using IQ-TREE. We primarily analyzed our data sets with the uncorrelated lognormal (UCLN) relaxed clock (Drummond et al. 2006; Rannala and Yang 2007), but replicated all analyses to check for any differences under the autocorrelated lognormal (ACLN) relaxed clock (Thorne et al. 1998; Kishino et al. 2001).

We estimated the overall substitution rate for each clock-partitioning scheme by running baseml under a strict clock, with a single point calibration at the root. We then used this estimate to select the shape (α) and scale (β) parameters for the gamma-Dirichlet prior on the overall substitution rate across loci in the MCMCTREE analysis according to the formulae $\alpha=(m/s)^2$ and $\beta=m/s^2$, where m and s are the mean and standard deviation of the substitution rate, respectively. For all analyses, we set the shape and scale parameters for the gamma-Dirichlet prior on rate variation across branches to 1 and 3.3, respectively. The posterior distribution of node ages was estimated with Markov chain Monto Carlo sampling, with samples drawn every 10^3 steps across a total of 10^7 steps, after a discarded burn-in of 10^6 steps. We ran all analyses in duplicate to assess convergence, and confirmed

Table 1
The Calibration Priors Used within This Study to Estimate the Angiosperm Evolutionary Timescale

Calibration Node	Uniform Priors		Gamma Priors		Fossil	References
	Min. Age Cal. (Ma)	Max. Age Cal. (Ma)	α	β		
CG Alismatales	120.7	350	4332.8	3264.4	*Mayoa portugallica*	Magallón et al. (2015)
CG Angiospermae	136	350	5245.7	3507.2	Early Cretaceous pollen grains	Magallón et al. (2015)
CG Arecales	83.6	350	1992.0	2167.7	*Sabolites carolinensis*	Iles et al. (2015)
CG Boraginales	47.8	126.7	806.6	1535.4	*Ehretia clausentia*	Martínez-Millán 2010
CG Brassicales	89.3	126.7	2530.9	2577.0	*Dressiantha bicarpelata*	Magallón et al. (2015)
CG Caryophyllales	70.6	126.7	1495.4	1926.2	*Coahuilacarpon phytolaccoides*	Magallón et al. (2015)
CG Cornales	89.3	126.7	2530.9	2577.0	*Tylerianthus crossmanensis*	Magallón et al. (2015)
CG Ericales	89.3	126.7	2530.9	2577.0	*Pentapetalum trifasciculandricus*	Magallón et al. (2015)
CG Fabales	55.8	126.7	897.6	1462.5	*Paleosecuridaca curtissi*	Magallón et al. (2015)
CG Fagales	96.6	126.7	2689.9	2532.0	*Normapolles* pollen	Magallón et al. (2015)
CG Gentianales	37.2	126.7	445.7	1086.9	*Emmenopterys dilcheri*	Magallón et al. (2015)
CG Magnoliales	112.6	350	4197.7	3390.7	*Endressinia brasiliana*	Massoni et al. (2015b)
CG Myrtales	87.5	126.7	2534.2	2632.6	*Esgueiria futabensis*	Magallón et al. (2015)
CG Oxalidales	100.1	126.7	2918.4	2651.4	*Tropidogyne pikei*	Chambers et al. (2010)
CG Pandanales	86.3	350	2289.8	2411.3	*Mabelia connatifila*	Iles et al. (2015)
CG Paracryphiales	79.2	126.7	1926.6	2209.7	*Silvianthemum suecicum*	Magallón et al. (2015)
CG Ranunculales	112.6	126.7	3867.5	3124.8	*Texeiraea lusitanica*	Magallón et al. (2015)
CG Saxifragales	89.3	126.7	2530.9	2577.0	*Microaltingia apocarpela*	Magallón et al. (2015)
CG Zingiberales	72.1	350	1663.3	2096.8	*Spirematospermum chandlerae*	Iles et al. (2015)
SG Buxales	99.6	126.7	3306.3	3019.9	*Spanomera marylandensis*	Magallón et al. (2015)
SG Cycadales	268.3	350	21939.8	7434.1	*Crossozamia*	Nagalingum et al. (2011)
SG gymnosperms	306.8	350	28377.3	8408.2	*Cordaixylon iowensis*	Clarke et al. (2011)
SG Platanaceae	107.7	126.7	3362.6	2837.3	*Sapindopsis variabilis*	Magallón et al. (2015)
SG Winteraceae	125	350	4738.5	3419.5	*Walkeripollis gabonensis*	Massoni et al. (2015b)

Note.—"CG" and "SG" refer to the crown and stem groups, respectively, of the clade of interest.

sufficient sampling by checking that the effective sample sizes of all parameters were above 200.

We repeated the MCMCTREE analysis for all P_{CSTAR}, P_{RATE}, and P_{RAND} schemes. An advantage of MCMCTREE is the option to use approximate likelihood calculation, which is much faster than full likelihood calculation (Thorne et al. 1998; dos Reis and Yang 2011). However, this precludes the calculation of marginal likelihoods using path sampling and similar methods, which require the full likelihood to be computed. Instead, we compared the means and 95% credibility intervals of the posterior estimates of divergence times across our partitioning strategies. We chose to focus on six nodes in the angiosperm phylogeny: the crown groups of all angiosperms, magnoliids, monocots, eudicots, campanulids, and Liliales. The first four of these were chosen because they define major clades in the angiosperm phylogeny. The other two nodes were chosen because they do not have explicit fossil-based calibration priors.

For each of the 12 numbers of clock-subsets, we sampled from the joint prior by running the analysis without data. This allowed us to compare the prior and posterior distributions of node ages and to observe the influence of changing the number of clock-subsets. The P_{CSTAR}, P_{RATE}, and P_{RAND} schemes are all treated as identical because the sequence data are not taken into account.

Fossil Calibrations

Calibrations are the most important component of Bayesian molecular dating, with critical impacts on posterior estimates of divergence times. Therefore, we selected a set of 23 calibration priors primarily based on recent studies that carefully considered the phylogenetic affinities of angiosperm fossils (table 1). We also applied two calibration priors to the gymnosperm outgroup. Fossils can strictly only provide a minimum age for the divergence of lineages from their common ancestor, so we chose to implement fossil calibrations primarily as uniform distributions with soft bounds. This approach assigns an equal prior probability for all ages between specified minimum and maximum ages, with a 2.5% probability that the age surpasses each bound (Yang and Rannala 2006).

We implemented two maximum age constraints: 1) 350 Ma for the divergence between angiosperms and gymnosperms (the root), a well accepted upper bound for this divergence (Foster et al. 2017); and 2) 126.7 Ma for the origin of crown eudicots, corresponding to the upper bound of the Barremian–Aptian boundary (reviewed by Massoni et al. 2015a). The latter constraint is widely used and is justified by the complete absence of tricolpate pollen before the latest Barremian, yet some molecular dating results have suggested

an earlier origin for eudicots (Smith et al. 2010; Foster et al. 2017; Zeng et al. 2017). Ranunculales, one of the earliest-diverging eudicot orders, has a fossil record dating back to the late Aptian/early Albian. Therefore, implementing the eudicot maximum constraint results in a strong prior being placed on crown-group eudicots appearing between ~126.7 and 112.6 Ma. As a result, including the eudicot maximum constraint leads to the eudicot crown node being a useful example of a heavily constrained node for downstream comparisons of the uncertainty in posterior age estimates.

For comparison, we also performed analyses with our P_{CSTAR} schemes using gamma calibration priors and the UCLN relaxed clock. In this case, the mean of each gamma prior was set to the age of each fossil +10%, with an arbitrary standard deviation of 2 (table 1). This effectively brackets the age estimates of calibrated nodes within a very narrow interval. In such a calibration scheme, the precision of age estimates is not expected to improve substantially with increased clock-partitioning.

Results

Angiosperm Evolutionary Timescale

Our ClockstaR analysis identified the optimal value of k to be 1, suggesting that a single pattern of among-lineage rate heterogeneity is shared across protein-coding genes from the chloroplast genomes. However, despite $k = 1$ being optimal, the values of the gap statistic were still higher for all values of $k > 5$ (fig. 1). Based on our analysis using the optimal clock-partitioning scheme ($k = 1$) and the UCLN relaxed clock, we estimated the time to the most recent common ancestor of angiosperms to be 196 Ma (95% credibility interval 237–161 Ma; fig. 2). We inferred that crown magnoliids first appeared 171–115 Ma, and that crown monocots arose contemporaneously, 167–120 Ma. Crown eudicots were inferred to have arisen 128–124 Ma, with this precise estimate reflecting the strong calibration prior placed upon this node. Finally, our estimates for the time to the most recent common ancestors of campanulids and Liliales were 101–91 Ma and 108–91 Ma, respectively.

The true age of crown angiosperms is unknown, so we cannot assess the absolute accuracy of our date estimates. Instead, we consider the consistency of mean age estimates across analyses (Hillis 1995). The mean age estimates for all crown angiosperms, magnoliids, and monocots varied slightly across values of k from 1 to 3, but estimates remained stable across all other values of k. Mean age estimates for crown eudicots only varied by approximately 2 Myr across all values of k. Mean age estimates for crown Liliales were stable across all clock-partitioning schemes. However, mean estimates for crown campanulids steadily declined by approximately 10–15 Myr as the number of loci increased. We observed the same broad trends in accuracy for all nodes of interest when using the ACLN relaxed clock, although mean age estimates were consistently slightly younger than in analyses with the UCLN relaxed clock. In our analyses with the P_{CSTAR} schemes and with gamma calibration priors, mean age estimates for crown angiosperms steadily increased with increasing numbers of clock-subsets, but the mean estimates were stable for all other nodes of interest.

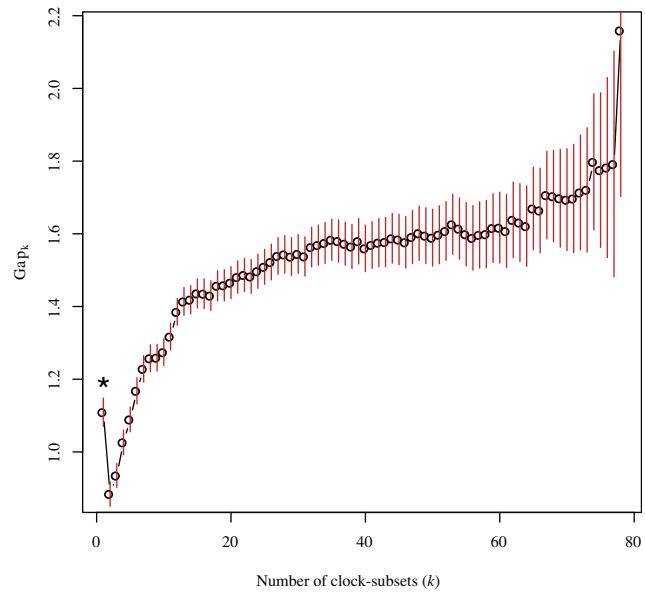

Fig. 1.—Gap statistic values for different numbers of clock-subsets (k) for the plastome-scale angiosperm data set, inferred using partitioning around medoids in ClockstaR. The asterisk indicates the optimal number of clock-subsets.

Precision in Estimates of Divergence Times

We focus first on our results when using the UCLN relaxed clock, uniform calibration priors, and with clock-partitioning according to ClockstaR. We report improvements in the precision of node-age estimates by calculating the decrease in 95% CI width, which we standardized by dividing by the posterior mean. The optimal clock-partitioning scheme was inferred to be $k = 1$, matching the results of previous analyses (Duchêne et al. 2016). However, increasing the number of clock-subsets generally led to large increases in the precision of node-age estimates. The impact of this is perhaps most striking in the inferred age of crown angiosperms. Increasing the number of clock-subsets from $k = 1$ to $k = 2$ led to a reduction in statistical fit (fig. 1), but also reduced the width of the 95% CI for the inferred age of crown angiosperms from 77 to 46 Myr (an improvement in precision of 35.4%).

Greater clock-partitioning led to further improvement in precision (fig. 3). For example, implementing a clock-partitioning scheme with $k = 20$ reduced the width of the 95% CI for the inferred age of crown angiosperms to only 20 Myr, representing a 73.1% improvement in precision.

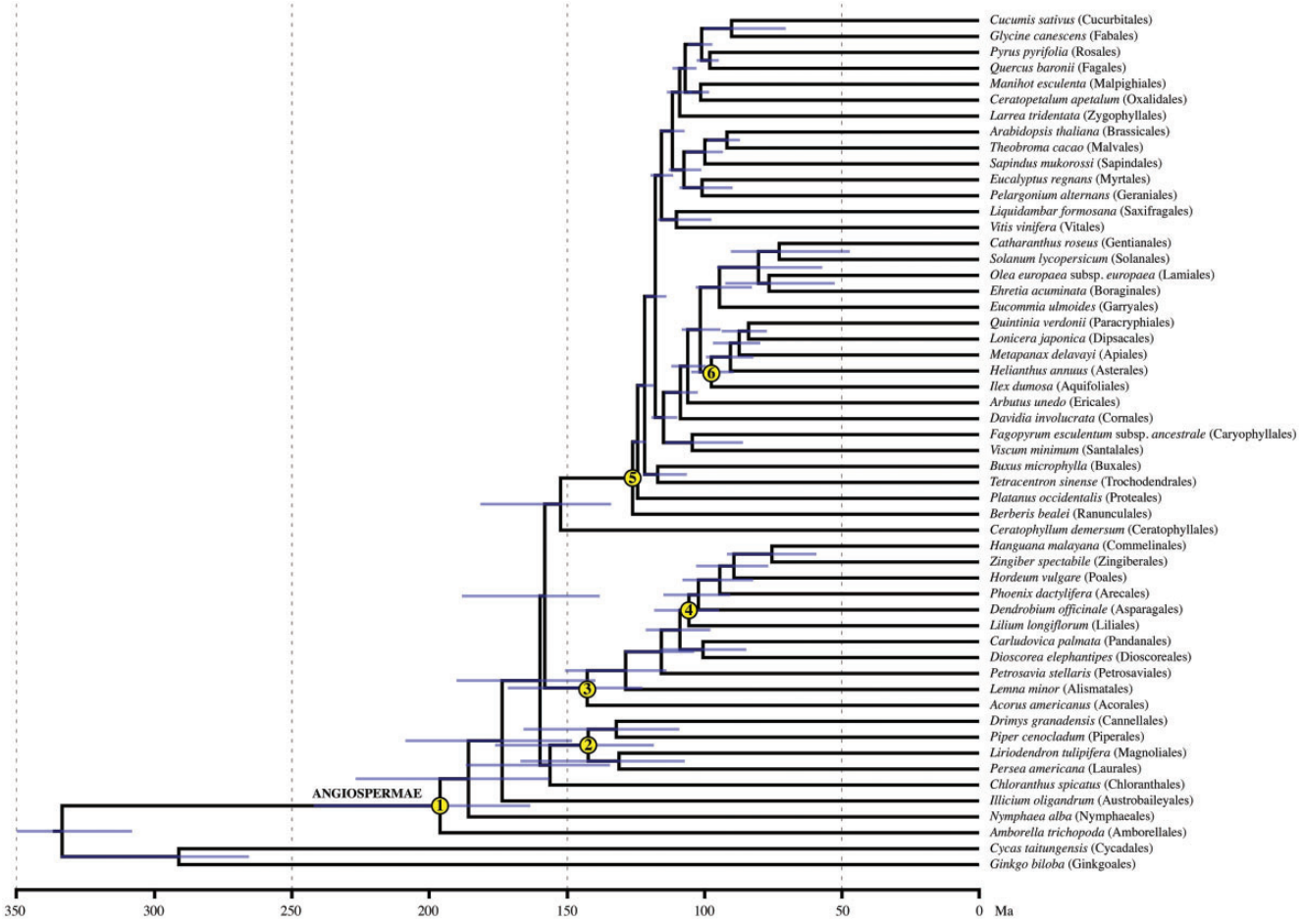

Fig. 2.—Chronogram depicting the evolutionary timescale of 52 angiosperm taxa and two gymnosperm outgroup taxa. The chronogram was estimated using Bayesian analysis of 79 genes from the 54 taxa in MCMCTREE, implementing the optimal clock-partitioning scheme ($k = 1$) and the uncorrelated lognormal relaxed clock. Tip labels indicate the taxa sampled in our study, with the orders they belong to in parentheses. Numbers in circles correspond to our six nodes of interest, as follows: 1) Angiospermae, 2) Magnoliidae, 3) Monocotyledoneae, 4) Liliales, 5) Eudicotyledoneae, and 6) Campanulidae.

However, the rate of improvement in precision declined rapidly for increasing numbers of clock-subsets (fig. 3).

An improvement in precision with the number of clock-subsets can also be observed in the age estimates for both magnoliids and monocots. For example, increasing k from 1 to 20 results in respective increases of 76.1% and 68% in precision in the age estimates for crown magnoliids and crown monocots (fig. 3). When considering the nodes corresponding to the crown groups of campanulids and Liliales, a similar trend can be observed, albeit with a less drastic increase in precision. Increasing the number of clock-subsets led to 29.7% and 37.7% increases in precision for the crown groups of campanulids and Liliales, respectively. However, there is a vastly different trend in the age estimate for crown eudicots. In this case, the age estimate for $k = 1$ is already precise (95% credibility interval: 128–124 Ma) and increasing the number of clock-subsets actually led to a slight decrease in precision of 0.02%.

Compared with the P_{CSTAR} clock-partitioning schemes, very similar trends in precision were observed for both the P_{RATE} scheme (fig. 4) and P_{RAND} scheme (fig. 5). The only differences were that there was less variation in mean age estimates for smaller values of k compared with the ClockstaR partitioning scheme, and standardized improvements in precision were consistently slightly greater (supplementary table S2, Supplementary Material online). For example, the widths of the 95% CIs, and the mean age estimates, declined monotonically in both classes of clock-partitioning schemes.

We observed the same broad trends across all clock-partitioning schemes when using the ACLN relaxed clock. With increasing numbers of clock-subsets, the uncertainty in age estimates rapidly decreased, with the exception of the age estimate for the eudicot crown node. Even with $k = 1$, however, the precision of the age estimates was much greater than in the corresponding analysis with the UCLN relaxed clock. For example, when implementing the P_{CSTAR} clock-partitioning schemes, the 95% credibility interval of the age estimate for crown angiosperms spanned 77 Myr when using the UCLN relaxed clock, but only 59 Myr when using the

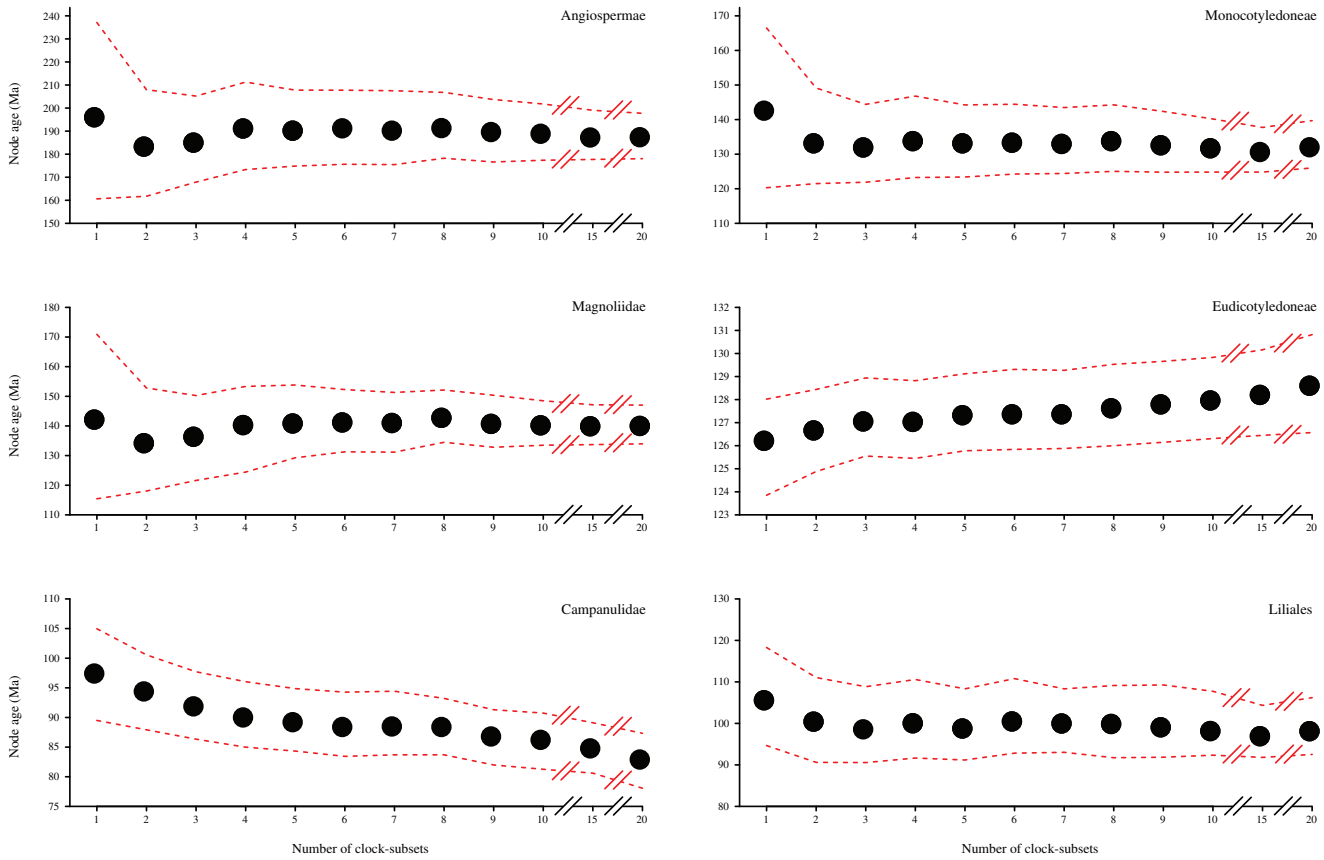

Fig. 3.—Mean posterior age estimates and associated 95% credibility intervals for six nodes in the angiosperm phylogeny with increasing numbers of clock-subsets (k), as inferred using an uncorrelated lognormal relaxed clock, clock-partitioning according to the optimal schemes identified in ClockstaR, and uniform calibration priors.

ACLN relaxed clock. Additionally, age estimates for crown eudicots became less precise as the degree of clock-partitioning increased. We observed the same trend for the other nodes of interest across analyses, and the apparent limit to uncertainty appeared to be reached much more rapidly than with the UCLN relaxed clock (supplementary figs. S2–S4 and table S2, Supplementary Material online).

When using highly informative gamma calibration priors in our additional analyses of the P_{CSTAR} schemes, we found that for the crown groups of angiosperms, monocots, and magnoliids, the increases in precision with greater clock-partitioning were much lower than with uniform calibration priors (supplementary fig. S5 and table S2, Supplementary Material online). For example, an improvement of only 18.5% occurred in the precision of the age estimate for crown angiosperms. The opposite trend occurred for the crown nodes of eudicots, campanulids, and Liliales.

When we implemented uniform calibration priors, greater clock-partitioning led to either no change or decreases in precision for age estimates of crown-group eudicots, but when using gamma calibration priors the precision improved by 36% with greater clock-partitioning. For crown-group Liliales, increasing k from 1 to 20 led to a 64.3% increase in the precision of age estimates, the greatest improvement of all six key nodes. However, it is worth noting that our age estimates for all six nodes of interest were very precise even when $k = 1$. Therefore, in terms of absolute time units, there was generally little improvement in precision with increasing numbers of clock-subsets.

In most cases, there is a clear difference between the posterior and prior distributions for our six nodes of interest (supplementary figs. S6–S8, Supplementary Material online). Additionally, while the shapes of the prior distributions are nearly identical with increasing values of k, the shapes of the posterior distributions closely mirror the trends described above based on 95% CIs, as expected.

Discussion

The goal of all molecular dating studies is to estimate the evolutionary timescale with a useful degree of precision and accuracy. We demonstrated that increasing the degree of clock-partitioning leads to increasingly precise age estimates, which has recently been shown in an independent study by Angelis et al. (2017), and is predicted by the finite-sites theory (Zhu et al. 2015). Additionally, clock-partitioning schemes

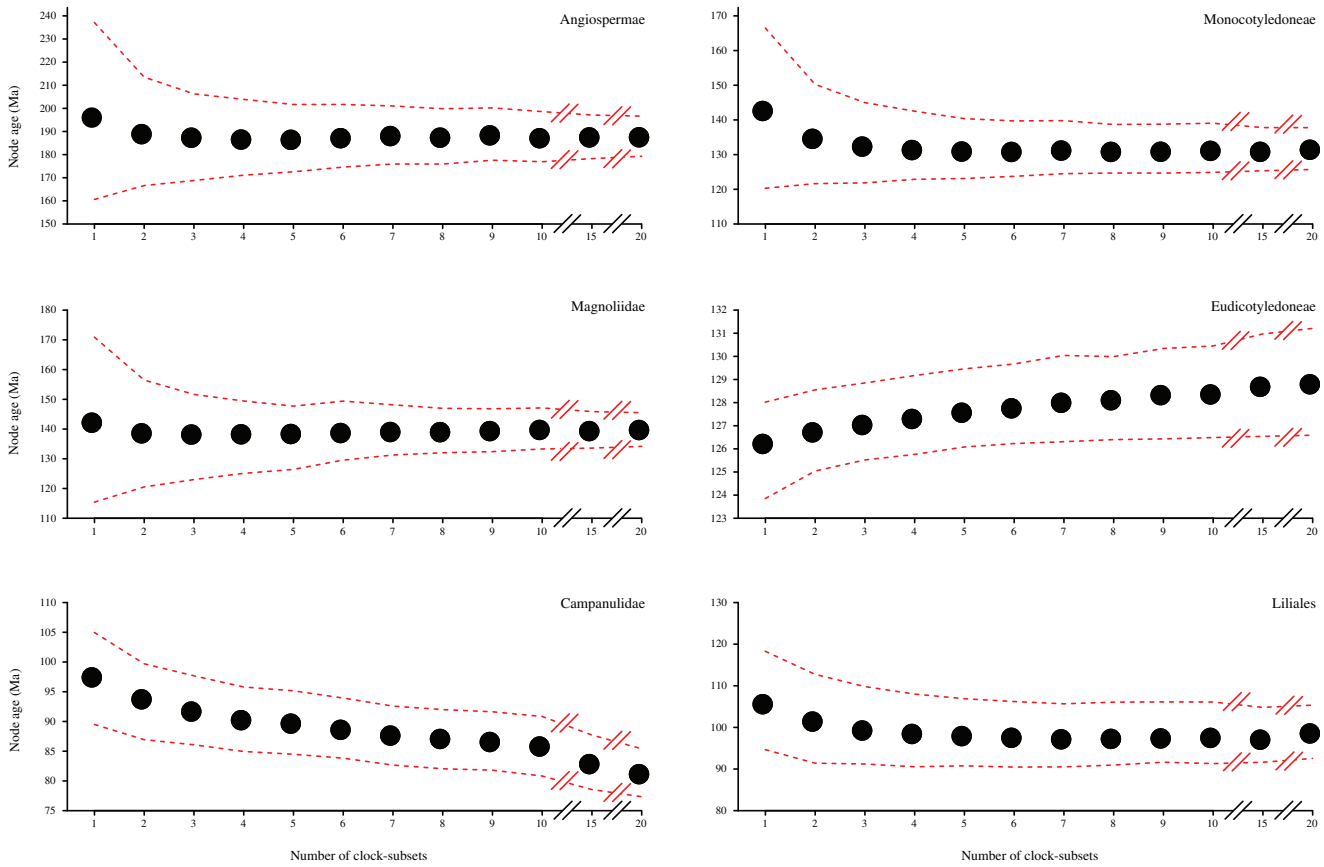

Fig. 4.—Mean posterior age estimates and associated 95% credibility intervals for six nodes in the angiosperm phylogeny with increasing numbers of clock-subsets (*k*), as inferred using an uncorrelated lognormal relaxed clock, clock-partitioning according to relative rates of substitution, and uniform calibration priors.

based on patterns of among-lineage rate heterogeneity or relative substitution rates did not have any measurable advantage over randomly assigning genes to clock-subsets, at least in terms of the accuracy and precision of the resulting estimates of divergence times.

The near-identical patterns of precision across all clock-partitioning schemes stands in contrast with some previous suggestions that the assignment of genes to clock-subsets is more important than the number of clock-subsets (Duchêne and Ho 2014). However, through simulations it has been demonstrated that different partitioning schemes only tend to have large impacts on the accuracy of posterior divergence times when the molecular clock is seriously violated, when the rate prior is misspecified, or when fossil calibrations are in conflict or incorrect (Angelis et al. 2017). In such cases, it is possible for increased clock-partitioning to yield highly precise age estimates, but for the 95% CIs of these estimates to exclude the true age. This goes some way to explaining why we observed such consistent age estimates across nearly all partitioning schemes, since we carefully chose appropriate values for the rate prior and implemented appropriate fossil calibrations that were not in conflict.

Our results demonstrate that to improve the precision of age estimates, one could simply increase the degree of clock-partitioning by assigning genes to an arbitrarily large number of clock-subsets, until the marginal benefit of increasing the number of clocks is close to zero (Zhu et al. 2015). An obvious consequence of this is that one must consider whether such an increase is desirable or biologically meaningful. If there is evidence that a data set conforms to a single pattern of rate variation among lineages, an increase in precision from clock-partitioning is not justifiable because the clock-subsets do not constitute independent realizations of the process of rate variation (Zhu et al. 2015). Our analysis using ClockstaR indicates that within our data set, all genes exhibit the same pattern of rate heterogeneity among lineages, such that they should be analysed using a single clock model. In this case, increasing the degree of clock-partitioning leads to a model that overfits the data, does not appear to accurately predict the data, and is insensitive to the sampled data. Normally this would be expected to occur when a model underfits the data, but the increasing sets of "independent" branch-rate estimates for each clock-subset ensure that estimates of node times remain precise.

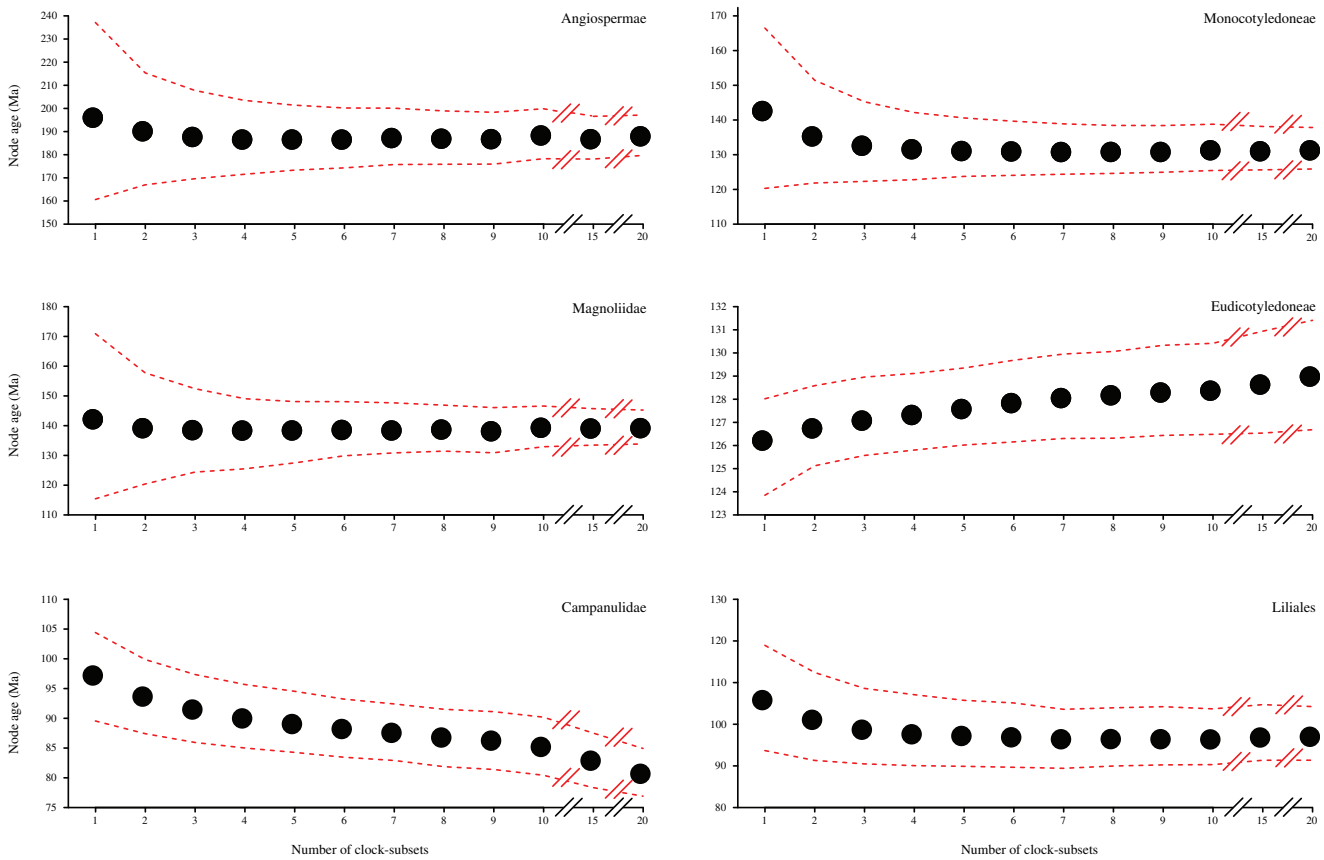

Fig. 5.—Mean posterior age estimates and associated 95% credibility intervals for six nodes in the angiosperm phylogeny with increasing numbers of clock-subsets (*k*), as inferred using an uncorrelated lognormal relaxed clock, clock-partitioning according to random assignment of genes to clock subsets, and uniform calibration priors. The estimates presented here are the averages of three random assignments of genes to clock-subsets for each value of *k*.

The uncertainty in posterior divergence times can be divided into three components: 1) uncertainty in branch lengths due to limited sequence length (*N*); 2) among-lineage rate variation for each clock-subset, as well as the evolutionary rate variation among clock-subsets; and 3) uncertainty in fossil calibrations (Zhu et al. 2015). If the number of clock subsets (*L*) is large, then the uncertainty caused by limited sequence length approaches zero at the rate of 1/*N*. Additionally, the uncertainty attributable to the second component approaches zero at the rate of 1/*L*. As $N \rightarrow \infty$ and $L \rightarrow \infty$, the uncertainty in divergence-time estimates should be wholly attributable to uncertainty in the fossil calibrations (Zhu et al. 2015). For a data set of fixed size, such as our angiosperm data set, increasing *L* will reduce *N*, and vice versa. We found that partitioning the data set into increasing numbers of clock-subsets led to improvements in precision, which implies that increasing *L* has a larger impact on precision than decreasing *N* has on reducing precision. However, it is likely that for very small values of *N*, the estimation error in branch lengths will grow rapidly.

An important exception to the overall trend was the age inferences for the crown eudicot node. The most common calibration strategy for this node has been to place a maximum bound or a highly informative prior on the age of this node, based on the absence of tricolpate pollen before the Barremian–Aptian boundary (~126 Ma) (Magallón and Castillo 2009; Sauquet et al. 2012; Massoni et al. 2015a; Foster et al. 2017). Additionally, many of the earliest-diverging eudicot lineages have relatively old fossils dating to the late Aptian (~113 Ma). These lines of evidence provide a narrow age bracket for the eudicot crown, often causing age estimates for the eudicot crown node to be necessarily highly precise. As a result, the limit in uncertainty of the fossil calibrations should be reached rapidly. Therefore, the age of the eudicot crown node is useful to evaluate in light of the finite-sites theory. We found that increasing the number of clock-subsets had essentially no effect on the uncertainty in the age estimate of this node. A very similar pattern was observed when using tightly constrained gamma calibration priors, and we expect that the general trend extends to other cases in which calibrated nodes have strongly constrained ages, for example when lognormal or exponential priors are chosen (Smith et al. 2010; Magallón et al. 2015).

Our results are especially important for analyses of genome-scale data sets. The size of phylogenomic data sets

generally precludes molecular dating with computationally intensive phylogenetic software, such as BEAST (Bouckaert et al. 2014) or MrBayes (Ronquist et al. 2012), unless workaround methods are employed (Ho 2014). For example, some researchers have chosen to analyze each gene or data subset separately and then take the average of the results (Zeng et al. 2017). However, this methodology effectively assigns to each gene its own model of nucleotide substitution and its own clock model. Not only does this run the risk of severe overparameterization, but it also raises the question of how the estimates should be combined in a way that takes full account of estimation error. Another method is to apply data filtering to select only a subset of a data set, such as those that are the most clocklike (Jarvis et al. 2014) or the most informative (Tong et al. 2016).

In cases where data-filtering approaches are not feasible, less computationally intensive methods can be employed, such as the approximate-likelihood method of MCMCTREE. There are also non-Bayesian alternatives to phylogenomic dating, such as penalized likelihood (Sanderson 2002), that have been used to analyze large data sets (Zanne et al. 2014). Additionally, a number of rapid dating methods that can account for among-lineage rate heterogeneity without an explicit statistical model of branch-rate variation have been developed specifically for phylogenomic data sets (Kumar and Hedges 2016). Although these methods appear to have accuracy comparable to that of Bayesian methods, they cannot produce reliable estimates of the uncertainty in the inferred ages (Kumar and Hedges 2016). It is also unclear how well the results of these analyses will conform to the finite-sites theory.

In the context of clock-partitioning, an important final consideration is that comparison of clock-partitioning schemes only provides an indication of relative fit. It does not indicate whether any of the partitioning schemes actually provides an adequate description of the process that generated the data (Duchêne et al. 2015). For example, even the most parameter-rich clock-partitioning scheme might be an inadequate description of the data. There have been recent developments in methods for evaluating clock-model adequacy, but these techniques involve thresholds that depend on the lengths of the sequences across the clock-subsets (Duchêne et al. 2015). Further refinement of methods for testing clock-model adequacy will be required before they can be readily applied to clock-partitioning schemes.

The primary aim of the present study was not to provide a novel estimate for the angiosperm evolutionary timescale, but it is still useful to consider our results in the context of previous estimates. Our inferred origin for crown-group angiosperms in the late Triassic to early Jurassic is consistent with most modern molecular dating estimates (Bell et al. 2010; Magallón 2010; Clarke et al. 2011; Zeng et al. 2014; Beaulieu et al. 2015; Foster et al. 2017). Similarly, our age estimate for crown magnoliids of 171–115 Ma is very similar to a previous estimate of 179–127 Ma based on the most comprehensive molecular dating analyses of Magnoliidae (Massoni et al. 2015a). Our estimate of 167–120 Ma for the age of crown monocots is compelling, because a recent study of monocots using the fossilized-birth–death model inferred a very similar age of 174–134 Ma (Eguchi and Tamura 2016). Our age estimate for crown eudicots of 128–124 Ma suggests that there was not enough signal within the data to overcome the strong calibration priors placed upon this node. Finally, although our age estimate for the appearance of crown campanulids 101–91 Ma is very similar to those of recent studies (Magallón et al. 2015; Foster et al. 2017), our age estimate of 108–91 Ma for the time to the most recent common ancestor of Liliales was slightly younger than recent estimates.

Conclusion

In this study, we have demonstrated that the finite-sites theory for molecular dating applies to a typical genome-scale data set from angiosperms, with the exception of nodes that have strong age constraints. In contrast with previous suggestions, the choice of strategy for assigning genes to clocks does not appear to be important. These results imply that the data set can be arbitrarily partitioned into a large number of clock-subsets, up to the point at which there is little marginal benefit in increasing the degree of clock-partitioning. However, we caution that all molecular date estimates should be critically interpreted to determine whether their precision is meaningful or not. To this end, the best approach is to identify the patterns of among-lineage rate heterogeneity in a data set and to apply a clock-partitioning scheme that appropriately captures this variation.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors acknowledge the facilities and the technical assistance of the Sydney Informatics Hub at the University of Sydney and, in particular, access to the high-performance computing facility Artemis. This work was supported by an Australian Government Research Training Program (RTP) Scholarship to C.S.P.F., and the Australian Research Council [grant DP110100383 to S.Y.W.H.].

Literature Cited

Angelis K, Álvarez-Carretero S, Dos Reis M, Yang Z. 2017. An evaluation of different partitioning strategies for Bayesian estimation of species divergence times. Syst Biol. doi: 10.5061/dryad.d7839/1.

Angiosperm Phylogeny Group APG. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 181(1): 1–20.

Beaulieu JM, O'Meara B, Crane P, Donoghue MJ. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. Syst Biol. 64(5): 869–878.
Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. Am J Bot. 97(8): 1296–1303.
Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc Natl Acad Sci U S A. 92(25): 11331–11338.
Bouckaert R, et al. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp Biol. 10(4): e1003537.
Burnham KP, Anderson DR. 2003. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.
Chambers KL, Poinar G Jr, Buckley R. 2010. *Tropidogyne*, a new genus of Early Cretaceous Eudicots (Angiospermae) from Burmese amber. Novon 20(1): 23–29.
Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst Biol. 65(6): 997–1008.
Clarke JT, Warnock R, Donoghue PCJ. 2011. Establishing a time-scale for plant evolution. New Phytol. 192(1): 266–301.
dos Reis M, et al. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc R Soc Lond B Biol Sci. 279(1742): 3491–3500.
dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol. 28(7): 2161–2172.
dos Reis M, Yang Z. 2013. The unbearable uncertainty of Bayesian divergence time estimation. J Syst Evol. 51(1): 30–43.
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4(5): e88.
Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol Biol Evol. 32(11): 2986–2995.
Duchêne S, Foster CSP, Ho SYW. 2016. Estimating the number and assignment of clock models in analyses of multigene data sets. Bioinformatics 32(9): 1281–1285.
Duchêne S, Ho SYW. 2014. Using multiple relaxed-clock models to estimate evolutionary timescales from DNA sequence data. Mol Phylogenet Evol. 77: 65–70.
Duchêne S, Molak M, Ho SYW. 2014. ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. Bioinformatics 30(7): 1017–1019.
Eguchi S, Tamura MN. 2016. Evolutionary timescale of monocots determined by the fossilized birth-death model using a large number of fossil records. Evolution 70(5): 1136–1144.
Foster CSP. 2016. The evolutionary history of flowering plants. J Proc Roy Soc New South Wales 149: 65–82.
Foster CSP, et al. 2017. Evaluating the impact of genomic data and priors on bayesian estimates of the angiosperm evolutionary timescale. Syst Biol. 66(3): 338–351.
Gaut B, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. Annu Rev Ecol Evol Syst. 42(1): 245–266.
Herendeen PS, Friis EM, Pedersen KR, Crane PR. 2017. Palaeobotanical redux: revisiting the age of the angiosperms. Nat Plants 3: 17015.
Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Syst Biol. 44(1): 3–16.
Ho SYW. 2014. The changing face of the molecular evolutionary clock. Trends Ecol Evol. 29(9): 496–503.
Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol. 23(24): 5947–5965.
Hughes NF. 1994. The enigma of angiosperm origins. Cambridge: Cambridge University Press.
Iles WJD, Smith SY, Gandolfo MA, Graham SW. 2015. Monocot fossils suitable for molecular dating analyses. Bot J Linn Soc. 178(3):346–374
Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215): 1320–1331.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4): 772–780.
Kaufman L, Rousseeuw PJ. 2009. Finding groups in data: an introduction to cluster analysis. Hoboken, NJ, USA: John Wiley and Sons.
Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol. 18(3): 352–361.
Kumar S, Hedges SB. 2016. Advances in time estimation methods for molecular data. Mol Biol Evol. 33(4): 863–869.
Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 29(6): 1695–1701.
Lanfear R, et al. 2013. Taller plants have lower rates of molecular evolution. Nat Commun. 4: 1879.
Magallón S. 2010. Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. Syst Biol. 59(4): 384–399.
Magallón S, Castillo A. 2009. Angiosperm diversification through time. Am J Bot. 96(1): 349–365.
Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207(2): 437–453.
Martínez-Millán M. 2010. Fossil record and age of the Asteridae. Bot.Rev. 76(1): 83–135.
Massoni J, Couvreur TLP, Sauquet H. 2015a. Five major shifts of diversification through the long evolutionary history of Magnoliidae (angiosperms). BMC Evol Biol. 15: 49.
Massoni J, Doyle JA, Sauquet H. 2015b. Fossil calibration of Magnoliidae, an ancient lineage of angiosperms. Palaeontol Electron. 18.1.2FC: 1–25.
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5): 1188–1195.
Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci U S A. 107(10): 4623–4628.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11(5): 715–724.
Nagalingum NS, et al. 2011. Recent synchronous radiation of a living fossil. Science 334(6057): 796–799.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1): 268–274.
R Core Team. 2016. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol. 56(3): 453–466.
Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3): 539–542.
Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19(1): 101–109.
Sauquet H, et al. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). Syst Biol. 61(2): 289–313.
Sauquet H, et al. 2017. The ancestral flower of angiosperms and its early diversification. Nat Commun. 8: 16047. doi: 10.1038/ncomms16047

Smith SA, Beaulieu JM, Donoghue MJ. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc Natl Acad Sci U S A. 107(13): 5897–5902.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322(5898): 86–89.

Soltis DE, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 98(4): 704–730.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34: W609–W612.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 15(12): 1647–1657.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. J Roy Stat Soc B Stat Methodol. 63(2): 411–423.

Tong KJ, Lo N, Ho SYW. 2016. Reconstructing evolutionary timescales using phylogenomics. Zool Syst. 41: 343–351.

Wernersson R. 2006. Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res. 34: W385–W388.

Wertheim JO, Sanderson MJ, Worobey M, Bjork A. 2010. Relaxed molecular clocks, the bias–variance trade-off, and the quality of phylogenetic inference. Syst Biol. 59(1): 1–8.

Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 111(45): E4859–E4868.

Wolfe KH, Li W-H, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 84(24): 9054–9058.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8): 1586–1591.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol. 23: 212–226.

Zanne AE, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506(7486): 89–92.

Zeng L, et al. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytol. 214: 1338–1354.

Zeng L, et al. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun. 5: 4956.

Zhu T, Dos Reis M, Yang Z. 2015. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. Syst Biol. 64(2): 267–280.

Associate editor: Laura Katz