Syst. Biol. 0(0):1–17, 2014

DOI:10.1093/sysbio/syu068

Methods for the Quantitative Comparison of Molecular Estimates of Clade Age and the Fossil Record

JULIA A. CLARKE*,[1] AND CLINT A. BOYD[2]

[1]*Jackson School of Geosciences, The University of Texas at Austin, 1 University Station C1100, Austin, TX 78712;* [2]*Department of Geology and Geological Engineering, South Dakota School of Mines and Technology, 501 East Saint Joseph Street, Rapid City, SD 57701, USA*
**Correspondence to be sent to: Jackson School of Geosciences, The University of Texas at Austin, 1 University Station C1100, Austin, TX 78712, USA; E-mail: Julia_Clarke@jsg.utexas.edu.*

Received 13 January 2013; reviews returned 20 July 2014; accepted 24 July 2014
Associate Editor: Norm MacLeod

Abstract.—Approaches quantifying the relative congruence, or incongruence, of molecular divergence estimates and the fossil record have been limited. Previously proposed methods are largely node specific, assessing incongruence at particular nodes for which both fossil data and molecular divergence estimates are available. These existing metrics, and other methods that quantify incongruence across topologies including entirely extinct clades, have so far not taken into account uncertainty surrounding both the divergence estimates and the ages of fossils. They have also treated molecular divergence estimates younger than previously assessed fossil minimum estimates of clade age as if they were the same as cases in which they were older. However, these cases are not the same. Recovered divergence dates younger than compared oldest known occurrences require prior hypotheses regarding the phylogenetic position of the compared fossil record and standard assumptions about the relative timing of morphological and molecular change to be incorrect. Older molecular dates, by contrast, are consistent with an incomplete fossil record and do not require prior assessments of the fossil record to be unreliable in some way. Here, we compare previous approaches and introduce two new descriptive metrics. Both metrics explicitly incorporate information on uncertainty by utilizing the 95% confidence intervals on estimated divergence dates and data on stratigraphic uncertainty concerning the age of the compared fossils. Metric scores are maximized when these ranges are overlapping. MDI (minimum divergence incongruence) discriminates between situations where molecular estimates are younger or older than known fossils reporting both absolute fit values and a number score for incompatible nodes. DIG range (divergence implied gap range) allows quantification of the minimum increase in implied missing fossil record induced by enforcing a given set of molecular-based estimates. These metrics are used together to describe the relationship between time trees and a set of fossil data, which we recommend be phylogenetically vetted and referred on the basis of apomorphy. Differences from previously proposed metrics and the utility of MDI and DIG range are illustrated in three empirical case studies from angiosperms, ostracods, and birds. These case studies also illustrate the ways in which MDI and DIG range may be used to assess time trees resultant from analyses varying in calibration regime, divergence dating approach or molecular sequence data analyzed. [angiosperm; calibration; divergence dating; ostracod; penguin; stratigraphic consistency metrics; Time tree.]

The fossil record, via the paleontological literature, is being accessed at an unprecedented rate. Indeed, papers citing fossils for use as calibrations generally far exceeds citation by paleontologists. Methods utilizing molecular sequence divergence and fossil calibrations to estimate the timing of lineage splitting and clade origin continue to be forwarded and refined (e.g., Drummond et al. 2006; Rutschmann et al. 2007; Xia and Yang 2011). Criteria for calibration choice including degree of temporal constraint, phylogenetic evaluation, and other properties of available reported fossils have been proposed (e.g., Hug and Roger 2007; Parham et al. 2012). These studies specifically address assessment of fossil data prior to analysis. Other approaches assessed sets of fossil calibrations by utilizing their fit with the estimated branch lengths of ultrametric trees during analysis and time tree estimation (Near and Sanderson 2004; Near et al. 2005; Marshall 2008; Lukoschek et al. 2012) or documented the effect of calibrations of uncertain phylogenetic position on confidence intervals of the resulting divergence dates (Lee et al. 2009). Fossil records have often been interpreted as unreliable when the perceived mismatch with a given time tree is deemed sufficiently large (Brochu et al. 2004). However, there has been only very limited work on metrics to quantify postanalysis this relative fit, or misfit, between time trees and the available fossil record for a given clade.

Four studies (Smith et al. 2006; Clarke et al. 2007; Marjanovic and Laurin 2007; Tinn and Oakley 2008), have investigated the use of metrics to describe the fit of fossil data and results from analysis of molecular divergence estimates of clade age after calibrations have been selected. Smith et al. (2006) and Tinn and Oakley (2008) propose metrics to reflect the relative fit of sets of recovered molecular dates and fossil data for individual nodes with both sets of temporal information available. Alternatively, Clarke et al. (2007) and Marjanovic and Laurin (2007) utilized modifications of existing stratigraphic consistency metrics to quantify the implied missing fossil record when a broad sample of extinct taxa, including wholly extinct clades, were included along with extant taxa.

Here, we compare previously proposed methods and metrics for quantifying fit between fossil oldest known records (OKRs; Walsh 1998) and molecular estimates of clade age (Smith et al. 2006; Marjanovic and Laurin 2007; Clarke et al. 2007; Tinn and Oakley 2008) and discuss limitations of these approaches. Two new metrics are proposed that differ from previous metrics in the treatment of uncertainty and in discriminating between cases in which fossil-based OKRs are younger or older

than molecular divergence estimates. One of these new metrics is node specific and the second is used for calculating the total incongruence implied for a phylogeny that includes both extant taxa and wholly extinct clades. These metrics are used to compare among time trees obtained using different methodologies (e.g., penalized likelihood [Sanderson 2003] and Bayesian approaches [Thorne and Kishino 2002; Drummond et al. 2006]), distinct genes, partitioning schema, or sets of calibrations. The joint prior from a set of individual priors based on fossil dates may be at odds with the minimum clade ages the fossils were used to provide (Warnock et al. 2012). With the exception of simultaneous evaluation approaches (Ronquist et al. 2012), the primary morphological data placing the fossils are not reassessed. Multiple exemplars of often species-rich and wholly extinct clades are also not included in time tree estimation. The previously proposed metrics discussed herein are the only published approaches to quantitatively describe the effect of enforcing dates from time trees on, for example, our current understanding of the completeness of the fossil record and the fit of fossil-based priors with posterior estimates of clade age.

To explore the performance of new and previously proposed metrics for comparing fossil and molecular minimum estimates of cladogenesis, we first evaluate time trees for Ostracoda (data from Tinn and Oakley [2008]) and Sphenisciformes (data from Baker et al. [2006] and Clarke et al. [2010]). We chose these two case studies because both datasets were previously used to propose the methods and specific metrics reviewed here and because they represent extreme cases. In the Tinn and Oakley (2008) dataset, the majority of recovered divergence dates are younger than the OKRs, while in the Clarke et al. (2010) dataset; most divergence dates are substantially older than fossil-based OKRs. A third case study illustrates how the new metrics may be used together to describe and compare recovered time trees. This example utilizes *Nothofagus* time trees recovered by Sauquet et al. (2012) in their exploration of the effect of distinct calibration regimes and divergence dating methods on the recovered ages for a targeted set of nodes.

Limitations of Existing Metrics of Fit Between Divergence Dates and Fossil Ages

Two metrics were previously proposed to quantify the incongruence between fossil-based OKRs and molecular divergence estimates at the nodes in the tree for which both sets of data are available (Smith et al. 2006; Tinn and Oakley 2008). The first metric assigns each node with fossil and molecular divergence estimates of clade age a score between 0 and 2 based on how many standard deviations (SDs) a compared OKR is from the recovered divergence date (Smith et al. 2006). A value of 2 is assigned when the OKR is within one SD of the divergence estimate. A value of 1 is assigned when the fossil age is >1 but <2 SDs, and a value of 0 is assigned when the fossil age is >2 SDs away from the divergence date (Smith et al. 2006). The Smith et al. (2006) approach was the basis for the ensemble "SEA" score (Tinn and Oakley 2008) that summarizes incongruence present at all nodes within a tree topology for which both fossil and molecular divergence dates are available. In the computation of that metric ($\text{SEA} = \frac{\sum_1^n Z_n}{n}$), n is each node in a topology for which both types of age data are available and Z_n is the score (0, 1, or 2) at each node n.

It was recognized at the publication of SEA that this metric has some undesirable properties (Tinn and Oakley 2008). High scores of SEA typically result from better overall fit among the two sources of data for all nodes compared. However, in situations where there is greater uncertainty around the divergence dates (i.e., higher standard errors), the odds of obtaining higher (i.e., "better") scores are increased (Tinn and Oakley 2008). Thus, Tinn and Oakley (2008) introduced a second metric, a weighted sum of squares based approach (WSS) intended to be a complementary metric to SEA. Values of WSS can be calculated either for a single node or averaged across the entire tree. In the implementation of (Tinn and Oakley, 2008), only a single calibration is used at the base of the clade of interest. The equation for WSS is $1 - \left(\frac{\sum_1^n \frac{(F_n - M_n)^2}{(F_n)^2}}{n} \right)$ where F_n and M_n are the fossil-based and molecular-based point estimates for node n, respectively. Unlike SEA, WSS does not take into account the uncertainty associated with the divergence dates. When fossil-based OKRs and divergence dates closely agree, values of WSS will approach 1. Unfortunately a negative WSS score is produced when the difference between the assessed fossil-based OKR and divergence date is greater than the absolute age of the fossil-based OKR. Tinn and Oakley (2008) also noted this systematic bias in WSS scores: it produces inflated values (i.e., closer to 1) when working with absolutely older fossils because the result is weighted by the square of the fossil age.

Other key limitations of both SEA and WSS methods were not previously noted. A first, and primary, limitation is that both methods fail to discriminate between situations in which divergence dates for nodes are older than the fossil-based OKR for a clade vs. situations where divergence dates are younger than these OKRs (Fig. 1). In one case (Fig. 1a,d), the divergence date is older than the OKR, consistent with an understanding of the nature of the fossil record as incomplete, fossils as minimum estimates of cladogenesis, and expectations that sequence divergence generally predates significant morphological divergence. In the second case, (Fig. 1c,f), the fossil-based OKR is identified as an *overestimate* of lineage first appearance. Divergence dates that are younger than the fossil-based OKRs could only be correct if the fossil record is wrong, and nearly ubiquitous assumptions about the incomplete nature of the record are also unrealistic. Figures 1a–c and 1d–f depict the behavior

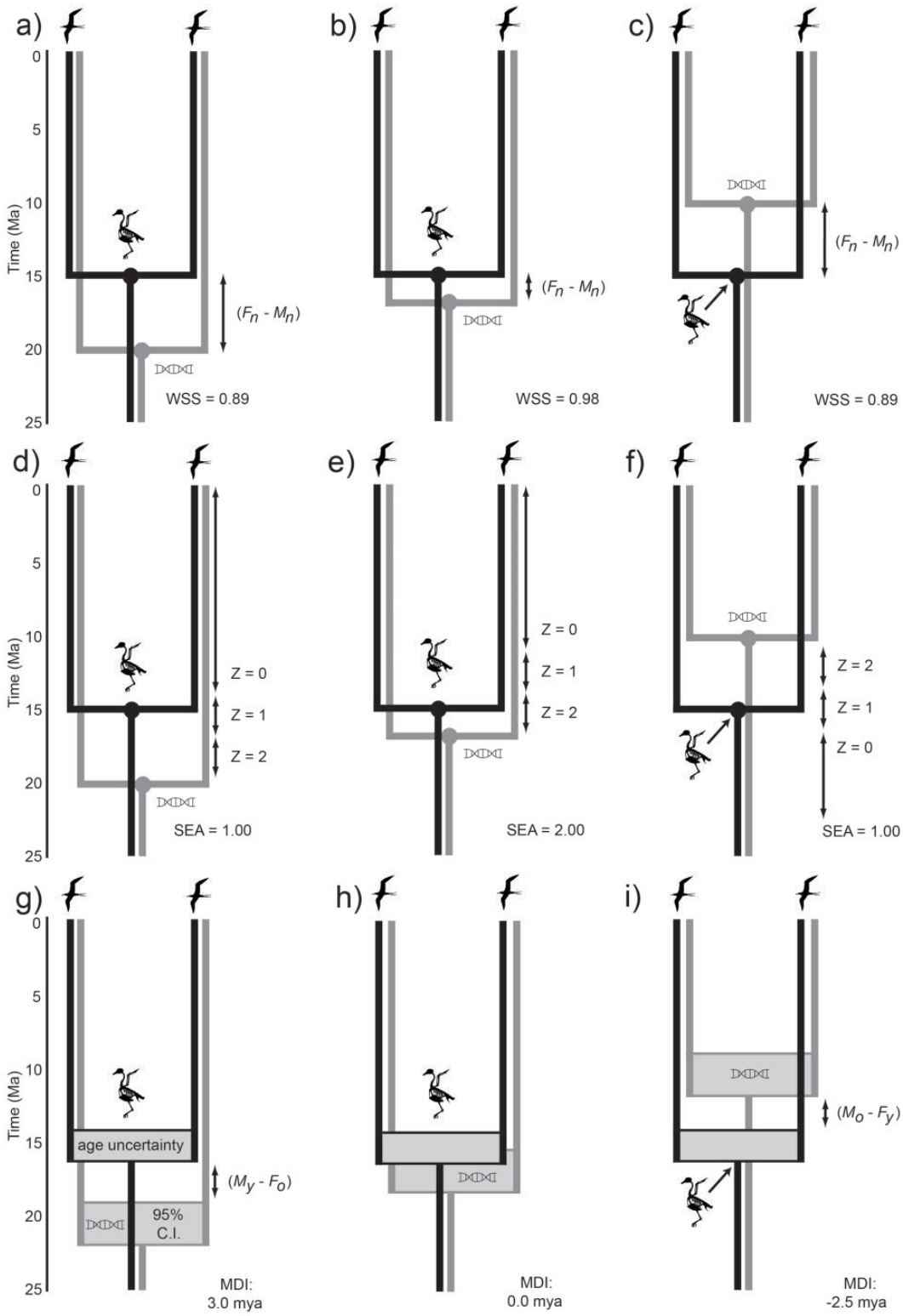

FIGURE 1. Illustration of how WSS (a–c), SEA (d–f), and MDI (g–i) are calculated and how each metric performs when the divergence date is older than (a, d, and g), closely approximates (b, e, and h), or is younger than (c, f, and i) the fossil-based estimate of cladogenesis. The divergence date is shown in grey, while the fossil-based estimate is shown in black. Grey boxes (g–i) indicate the uncertainty surrounding the age of the fossils and the 95% confidence interval for the molecular dates. Abbreviations: Fn, age of the fossil-based datum for node n; Fo, oldest possible age for the fossil; Fy, youngest possible age for the fossil; MDI, minimum divergence incongruence; Mn, divergence date for node n; Mo, oldest possible age for the divergence date; My, youngest possible age for the divergence date; SEA, Smith et al. (2006) metric; WSS, weighted sum of squares; Z, score for SEA metric.

TABLE 1. Comparison of methods for calculating incongruence among time trees and a given set of molecular divergence dates

	Marjanovic and Laurin (2007)	Clarke et al. (2007)	This article
Age data reporting	List of references	Table of age ranges and references for fossil and molecular data	Table of age ranges and references for fossil and molecular data
Tree topology/taxonomic sampling	Modified between analyses	Held constant	Held constant
Stratigraphic uncertainty	Single age used	Ranges to represent uncertainty	Ranges to represent uncertainty
Molecular uncertainty	Single divergence date used	Single divergence date used	Ranges (95% confidence interval) to represent uncertainty
Systematic uncertainty (Treatment of polytomies)	Taxa rearranged to maximize stratigraphic fit	ComPoly approach[a]	ComPoly approach[a]
Metric Used	AIG (=MIG)	MIG range	DIG range, MDI

Notes: AIG = actual implied gap; DIG = divergence implied gap; MDI = minimum divergence incongruence; MIG = minimum implied gap.
[a]Utilizes maximum and minimum fit rearrangements to compute a range (Boyd et al. 2011).

of both WSS and SEA, respectively, in situations where divergence dates are older than (Fig. 1a,d), closely agree with (Fig. 1b,e), and are younger than (Fig. 1c,f) the fossil-based OKR. Because the calculation of SEA for a single node is based only on how many SDs the OKR is from the assessed divergence date, it cannot differentiate between situations where the divergence date is older than or younger than the fossil-based OKR (Fig. 1d vs. f). Similarly, because WSS squares the difference between each OKR and divergence date, identical values are produced when the scale of disagreement is equal (Fig. 1a vs. c).

A second problematic characteristic of existing SEA and WSS methods concerns uncertainty associated with both sets of temporal data (i.e., fossil and molecular). Neither SEA nor WSS incorporate uncertainty surrounding the age of fossil-based OKRs into their calculations. Both metrics use a fixed age for the OKR rather than an age range. Incorporating uncertainty in the absolute age assigned to the deposit containing the fossil would involve setting a minimum and maximum possible range of ages for each fossil (i.e., an age range). SEA does incorporate uncertainty in the age of the divergence date into its calculations directly; it asks where the fossil age is relative to 1 or 2 SDs of the divergence date compared. WSS does not consider uncertainty associated with the divergence dates; instead, this metric compares between fixed dates for both the fossil-based OKRs and divergence dates. Accurately representing the uncertainty surrounding temporal data has been shown to substantially affect related metrics utilized to compare the fit of competing phylogenetic hypotheses to the fossil record (e.g., Pol and Norell 2006; Wills et al. 2008) and needs to be incorporated into any metric designed to compare fossil-based and molecular-based estimates of cladogenesis.

Two prior studies attempted to quantify the effect that enforcing sets of divergence dates has on the estimated pattern of cladogenesis and inferred missing fossil record across recovered topologies, including all phylogenetically evaluated extinct lineages (Clarke et al. 2007; Marjanovic and Laurin 2007). This approach lies in contrast with the SEA and WSS approaches which limit comparisons to specific nodes for which both molecular-based and fossil-based dates are available (i.e., clades with extant descendants). A summary of differences between these two approaches is given in Table 1. For example, Clarke et al. (2007) assessed the effect on the estimated missing fossil record for penguins if molecular-based dates for several dated nodes were enforced (e.g., penguin crown and penguin total group; Baker et al. 2006). These authors calculated the total implied missing record before and after enforcement of these dates utilizing a stratigraphic congruence metric, MIG range, that sums all implied missing fossil records (i.e., ghost lineages: Norell 1992) given a tree topology and a set of temporal data (Pol and Norell 2006). Unlike previous approaches utilizing this metric, which was developed to compare differing phylogenetic hypotheses for a single set of fossil-based dates, the topology was held constant while the ages were varied via the insertion of 'anchor taxa' (see below; Fig. 2).

Marjanovic and Laurin (2007) also defined what they considered a new metric aimed at quantifying the incongruence between fossil and molecular-based temporal data across a tree including an array of extinct taxa. They called this metric the Actual Implied Gap (AIG). They calculated AIG by "counting the branch lengths (in [myr]) that lie between the first fossil of a clade and its estimated origin as ghost lineages" and then summing "the total length of all ghost lineages" (Marjanovic and Laurin 2007: p. 376). By this definition, their AIG metric is identical to MIG (i.e., "the total ghost range implied by a given set of stratigraphic ranges on a given tree": Wills (1999): p. 559), although Wills (1999) did not consider the ghost lineages implied by molecular divergence.

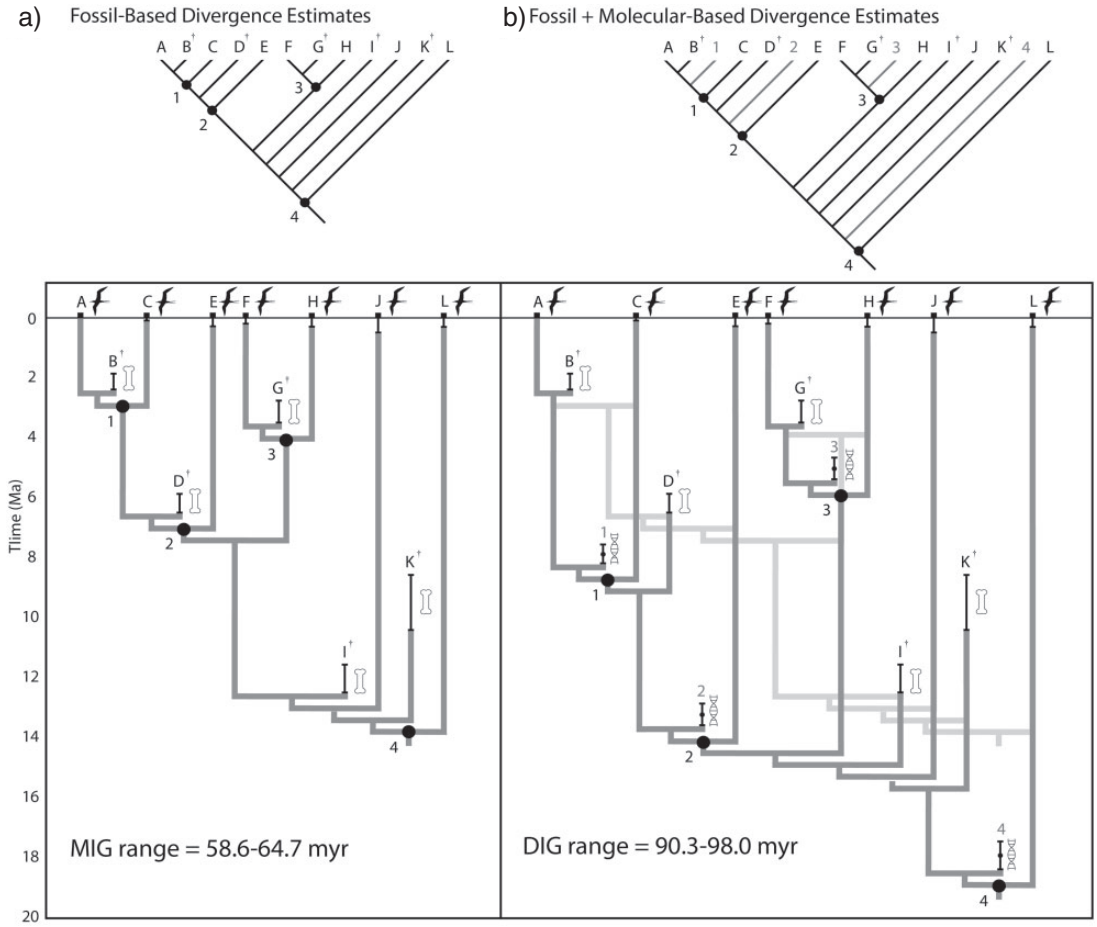

FIGURE 2. Analysis of a hypothetical example illustrating how divergence dates are incorporated into DIG range calculations: a) hypothetical phylogeny containing twelve terminal taxa (A–L), b) phylogeny with insertion of anchor taxa (1–4) to constrain the ages of the four nodes (black circles) to the divergence dates. The topology indicating the calibrated nodes and the placement of anchor taxa (grey in b) is given at top; corresponding chronograms are provided below. In the chronograms, temporal data are represented either by black squares (for extant taxa with no fossil history) or by black whisker bars that represent either stratigraphic uncertainty (fossils) or the 95% confidence interval (divergence dates). For (a), values of MIG range are reported (total implied ghost lineages, fossil-based dates only; factoring in stratigraphic uncertainty), while in (b) values of DIG range are reported (taking into account all fossil-based dates and enforcing the set of divergence dates using anchor taxa; factoring in both stratigraphic uncertainty and the 95% confidence intervals). Notes: the chronogram from (a) is underlain in light grey in (b) to highlight differences; visualization also necessitated slightly extending some branches in the chronograms relative to the actual temporal data analyzed in the calculation of the MIG and DIG range values reported. Abbreviations: †, extinct lineage; DIG range, divergence implied gap range; MIG range, minimum implied gap range.

While the explicit definition given for AIG is the same as that for MIG, the manner in which AIG calculations were made raises additional methodological concerns. Marjanovic and Laurin (2007) utilized the Stratigraphic Tools module (Josse et al. 2006) for the program Mesquite (Maddison and Maddison 2005), supplying their tree topologies and temporal data. When constructing their tree, they required all internal branch lengths to be at least 3 myr (considered a plausible minimum internal branch length based on Laurin [2004]), an approach that ensured the tree could be easily assessed visually (Marjanovic and Laurin 2007: p. 373). AIG values were then calculated by hand based on a visual examination of tree using the default scale built into Mesquite, which had a built-in minimum error of ±2 myr for each ghost lineage duration (Marjanovic and Laurin 2007: p. 373). Thus, error and internode lengths were similar. Marjanovic and Laurin (2007) also varied topology and sets of sampled taxa when comparing trees, obscuring the relative impact of the change in tree topology and enforcement of the molecular-based dates had on resultant estimates of implied missing fossil record. Additionally, implied missing fossil record was only calculated for those nodes with both molecular-based dates and fossils in common between trees, and ghost lineages for branches collapsed, or unresolved, in a compared tree were not summed. Thus, the influence of each of these changes on the resulting AIG values for the compared topologies was not clear.

As with the node specific metrics SEA and WSS (Tinn and Oakley 2008), the studies of Clarke et al. (2007) and Marjanovic and Laurin (2007) insufficiently

incorporated knowledge of uncertainty in the ages of fossil-based and molecular-based dates. Clarke et al. (2007) used the procedure proposed by Pol and Norell (2006) to incorporate the uncertainty in the fossil-based temporal data, which allows the user to define the full range of stratigraphic uncertainty and then iteratively sample from within that range when calculating MIG range. However, they did not use this method for the molecular-based dates compared; instead, they treated each molecular-based date as a fixed age. The study of Marjanovic and Laurin (2007) also treated both minimum clade ages from fossils and the divergence dates as fixed ages, though their study was completed and submitted for review prior to the publication of these range methods in Pol and Norell (2006).

MDI: A Metric for Assessing the Fit of Divergence Dates and Fossil Ages for Specific Nodes

We propose a metric, minimum divergence incongruence (MDI), to be computed for nodes of interest for which both OKRs and divergence dates are available. This metric improves characterization of the fit between divergence dates and OKRs at these nodes by taking into account uncertainly in the age of both sources of temporal data and by differentiating between situations where known fossils for a clade are younger or older than their respective divergence dates. MDI takes into account the uncertainty surrounding the age of each fossil using the range method proposed by Pol and Norell (2006) and utilizes the 95% confidence interval associated with the corresponding divergence date (Fig. 1g–i). It should be noted that uncertainty surrounding the OKR of a taxon can be based on a variety of criteria, including the age range of the containing stratigraphic unit or via the calculation of confidence intervals for fossil occurrences (e.g., Marshall 1990), and the preferred method will vary between taxa.

The methodology of calculating MDI varies depending upon the relative ages of the fossil-based and molecular-based dates. When the oldest possible age for the fossil-based temporal datum (F_o) is younger than the youngest possible age for the molecular-based temporal datum (M_y), MDI is calculated as: $M_y - F_o$ (Fig. 1g) and positive values are produced. When the youngest possible age for the fossil-based temporal datum (F_y) is older than the oldest possible age for the molecular-based temporal datum (M_o), MDI is calculated as: $M_o - F_y$ (Fig. 1i) and negative values are produced. When the uncertainty ranges for the fossil-based OKRs and 95% confidence intervals of divergence dates overlap, a value of zero is assigned to the node to indicate the close fit between these two sets of data.

The ways in which SEA and MDI incorporate uncertainty around the age of molecular-based estimates of cladogenesis differ significantly. While SEA asked whether the OKR lies within 1 or 2 SDs (Tinn and Oakley 2008), MDI utilizes the reported 95% confidence intervals (or posterior densities). If variance is normally distributed around the divergence date, then the 95% confidence interval is equivalent to 1.96 SDs, making these two approaches relatively similar. However, variance is not expected to be normally distributed in these situations. For example, Baker et al. (2006: table 1) report 95% confidence intervals that indicate variance for the estimated divergence dates is log-normally distributed. Therefore, use of SD of an untransformed log-normal distribution in the calculation of a comparative score may less accurately represent the data, making the use of the reported 95% confidence intervals a preferred approach. It should be noted that ideally, the precise variance would be compared, but this is rarely reported in practice.

As with all other node specific metrics (SEA and WSS), MDI is reported for individual nodes (Tinn and Oakley 2008). However, in contrast to these other metrics, each node is assigned a negative or positive value that indicates whether the fossil-based OKR is younger (positive) or older (negative) than the molecular-based date. Unlike SEA and WSS, ensemble MDI values are reported by summing only negative scores across nodes and reporting the number of nodes with such scores. Thus, ensemble MDI scores quantify the degree to which compared fossils are recovered as older than estimated divergence dates. Utilizing negative MDI in this way summarizes data that would be incompatible with these fossils actually comprising minimum estimates of clade origin. Summing both negative and positive MDI values would have no discernible meaning and would result in the recovery of "better" scores for trees with more divergence dates that are younger than the compared fossil-based OKRs (see "Discussion" section).

DIG range: A Metric for the Missing Fossil Record Implied by a Given Set of Divergence Dates

A second proposed metric, the divergence implied gap (DIG), is defined as: the sum total of all ghost ranges implied given a set of OKRs, a tree topology, and a time tree taking into account associated uncertainty (95% confidence intervals). DIG reports the total length of ghost lineages implied for all clades represented in a phylogeny. It can be assessed for topologies that include clades with no living descendants. The total increase in ghost lineages that results from enforcing a set of divergence dates can be ascertained by subtracting MIG (Benton 1994; Wills 1999) or MIG range: Pol and Norell 2006) values from DIG (or DIG range) values calculated for the same set of terminal taxa on the same tree topology. To clearly differentiate DIG from MIG, we clarify the definition of the latter metric (MIG) as follows: the sum total of all ghost ranges implied solely by the OKRs of a set of extant and extinct terminal taxa on a given tree topology. As discussed below, MDI and DIG range are intended to be used together, but they differ fundamentally in what they assess and how they are calculated.

Operationally, calculating DIG or DIG range values involves incorporating divergence dates by inserting additional proxy terminal taxa with zero branch lengths, referred to as "anchor taxa," into the original tree topology (Fig. 2b). For example, in Figure 2b the anchor taxon representing the origin of clade 2 is placed above terminal taxon E, constraining the age of clade 2 to the older divergence date. Because relative missing fossil record is being assessed by comparing DIG and MIG range (rather than a modified MSM* or GER approach; Pol et al. 2004), altering the size and shape of the tree via the insertion of anchor taxa does not invalidate comparisons of the resulting values. MIG range and DIG range only compare raw total implied missing fossil record values in millions of years; thus, anchor taxa enforce the molecular-based ages, but do not otherwise affect the values calculated. Anchor taxa should not be inserted for molecular-based estimates that are younger than the fossil-based estimates at a given node (i.e., those nodes with negative MDI values) because these dates necessarily cannot increase the implied missing fossil record for the clade under study. As a result, time trees with a large number of incompatible nodes (i.e., nodes with negative MDI values) would be more likely to produce low DIG values, since the impacts of fewer molecular-based ages are being assessed. To counter this issue, we only propose calculating DIG for time trees in which all nodes produce MDI values of zero or higher, making DIG a complimentary measure to MDI. However, in situations where all of the time trees being assessed have at least one node with a negative MDI value (proposed fossil minima older than estimated divergence date; e.g., the Tinn and Oakley [2008] study), strictly enforcing this requirement is not possible if we wish to use these metrics to select a preferred time tree (see "Discussion" and "Case Studies" section for illustration of how these metrics work in concert and potential biases).

When calculating both DIG and MIG, uncertainty in phylogenetic position, fossil ages, and molecular-based dates is addressed in two primary ways. First, both fossil and molecular divergence dates are set as ranges instead of fixed points. For fossils, this range describes stratigraphic uncertainty and is set equal to the minimum and maximum possible ages of the fossil containing stratigraphic unit, or may be set equal to a set of confidence intervals calculated for that fossil age (e.g., via methods described by Marshall 1990). For the divergence dates, this range is equal to the 95% confidence intervals associated with these dates. MIG range and DIG range values are then calculated as proposed by Pol and Norell (2006) for MIG range. A series of replicates are run and during each replicate a single age is randomly selected from the age ranges defined for each fossil-based and molecular-based date. The highest and lowest values obtained during these replicates are combined to yield the final DIG range or MIG range values. These values incorporate the full range of known uncertainty for all temporal data included in the analysis, assuming a large enough number of replicates are conducted (Pol and Norell 2006).

DIG and MIG range scores should ideally be calculated from the full set of trees produced by the original phylogenetic analysis, combining the highest and lowest recovered MIG and DIG values into the final range values to describe the degree of phylogenetic resolution present in the original dataset. The effect that uncertainty in phylogenetic position (i.e., presence of polytomies in tree topologies) can have on stratigraphic consistency metrics like DIG range was recently explored (Boyd et al. 2011). That study found that when calculating values for the original set of trees is not possible, the ComPoly approach (Boyd et al. 2011) should be employed, in which taxa in polytomies in the strict consensus tree are rearranged in two ways: stratigraphic order and reverse-stratigraphic order. The highest and lowest values obtained are combined to form the final DIG and/or MIG range scores. While this is the general method used by Clarke et al. (2007), Marjanovic and Laurin (2007) resolved polytomies only to maximize stratigraphic congruence (Table 1).

Empirical Case Studies

Methods

To compare the descriptive utility of previously proposed metrics and those proposed herein, we calculated the incongruence between fossil-based OKRs and divergence dates reported for Ostracoda (data from Tinn and Oakley 2008) and Sphenisciformes (fossil-based data from Clarke et al. [2010]). We chose these datasets as both were used in the papers that established the methods evaluated (SEA and WSS: Tinn and Oakley [2008]; modified MIG range applied to divergence dates, Clarke et al. [2007]). They also exemplify extreme cases. As discussed below, in Tinn and Oakley (2008) most divergence dates are younger than the OKRs. By contrast, the divergence dates from Baker et al. (2006) are all much older than the OKRs (Clarke et al. 2007, 2010). Finally, we also evaluated time trees recovered by Sauquet et al. (2012). That paper considered the effect of distinct calibration regimes and divergence dating methods on estimated divergence dates within *Nothofagus*.

All three examples illustrate the application of MDI and DIG range to describe and compare time trees. Multiple sets of divergence dates were available for comparison in each study, either owing to the use of multiple methods for estimating divergence dates (i.e., clock, linear/log penalized likelihood in r8s [Sanderson 2003], Mean Path Length in PATHd8 [Britton et al. 2006] and Bayesian approaches in MultiDivTime [Thorne and Kishino 2002] in Tinn and Oakley [2008]), or resulting from analysis of different sets of sequence data with the same method (RAG-1, mtDNA, combined dataset of Baker et al. [2006]). In Sauquet et al. (2012) both calibration regimes and methods (i.e., ML-PL utilizing r8s [Sanderson 2003] and Bayesian analyses in BEAST

[citealp16Drummond2007) were varied between the 30 reported analyses.

Values of the node specific metrics MDI, SEA, and WSS were calculated for all sets of dates reported in the Spheniscifomes and Ostracoda datasets (Tables 2 and 3). Ensemble SEA and WSS values were also summarized across the entire tree for each of these analyses and compared to the values of DIG range obtained for each tree (Table 4). For the sphenisciform data, values of MIG range (maximum and minimum age taking into account uncertainty in phylogenetic resolution and the fossil-based ages) were calculated for comparison to the resulting DIG range values (implied missing fossil record with the same minimum and maximum fossil-based ages and address of polytomies, but enforcing divergence estimates using the 95%

TABLE 2. Data and resulting SEA, WSS, and MDI scores for the twenty individual nodes assessed by Tinn and Oakley (2008) for the time trees resulting from five different divergence dating methods implemented in R8s, PATHd8, and MultiDivtTime

	Clock (LF)									
	—First Fossil—							———Scores———		
Node	Min	Reported	Max	Mean	SD	95% Confidence Intervals[a]		SEA	WSS	MDI
1	476.9	480.0	490	330.42	25.53	280.38	380.46	0.00	0.90	−96.44
2	442.2	443.7	447.1	498.60	9.27	480.43	516.77	0.00	0.98	33.33
6	293.8	299.0	299.8	26.52	3.45	19.76	33.28	0.00	0.17	−260.52
8	420.4	426.2	428.6	66.17	7.71	51.06	81.28	0.00	0.29	−339.12
9	425.9[b]	433.7[b]	437.9[b]	187.10	14.47	158.74	215.46	0.00	0.69	−200.54
11	181.5	189.6	191.6	151.07	10.09	131.29	170.85	0.00	0.96	−10.65
12	164.2[b]	171.6[b]	174.6[b]	136.15	10.51	115.55	156.75	0.00	0.96	−7.45
13	164.2[b]	171.6[b]	174.6[b]	88.82	6.50	76.08	101.56	0.00	0.77	−62.64
14	164.2	171.6	174.6	68.5	7.36	54.07	82.93	0.00	0.64	−81.27
15	425.9	433.7	437.9	125.07	8.78	107.86	142.28	0.00	0.49	−283.62
16	248.8[b]	249.7[b]	254.5[b]	87.2	5.74	75.95	98.45	0.00	0.57	−150.35
17	248.8	249.7	254.5	83.63	5.53	72.79	94.47	0.00	0.56	−154.33
18	248.8[b]	249.7[b]	254.5[b]	60.76	3.95	53.02	68.50	0.00	0.43	−180.30
20	248.8[b]	249.7[b]	254.5[b]	54.40	3.99	46.58	62.22	0.00	0.39	−186.58
21	248.8	249.7	254.5	50.28	3.43	43.56	57.00	0.00	0.36	−191.80
23	131.9	136.4	143.2	31.32	3.75	23.97	38.67	0.00	0.41	−93.23
26	131.9[b]	136.4[b]	143.2[b]	47.38	3.24	41.03	53.73	0.00	0.57	−78.17
27	124	130.0	135.9	31.87	3.51	24.99	38.75	0.00	0.43	−85.25
29	60.9	65.5	71.2	51.23	4.95	41.53	60.93	0.00	0.95	0.00
30	425.9[b]	433.7[b]	437.9[b]	149.12	10.01	129.50	168.74	0.00	0.57	−257.16

	PL (linear)							**PL (log)**						
					Scores							Scores		
Node	Mean	Std	95% Confidence Intervals[a]		SEA	WSS	MDI	Mean	Std	95% Confidence Intervals[a]		SEA	WSS	MDI
1	518.26	10.55	497.58	538.94	1.00	0.99	7.58	501.46	19.96	462.34	540.58	1.00	**1.00**	**0.00**
2	483.82	11.21	461.85	505.79	0.00	0.99	14.75	421.51	15.34	391.44	451.58	1.00	**1.00**	**0.00**
6	186.63	28.26	131.24	242.02	0.00	0.86	−51.78	63.45	17.60	28.95	97.95	0.00	0.38	−195.85
8	326.22	28.17	271.01	381.43	0.00	0.94	−38.97	190.58	47.95	95.60	284.56	0.00	0.69	−135.84
9	411.28	14.81	382.25	440.31	**2.00**	**1.00**	**0.00**	276.43	22.72	231.90	320.96	0.00	0.88	−95.04
11	349.18	14.62	320.52	377.84	0.00	0.29	129.42	204.67	14.19	176.86	232.48	1.00	0.99	**0.00**
12	321.93	15.82	290.92	352.94	0.00	0.23	116.32	177.96	17.57	143.52	212.40	**2.00**	**1.00**	**0.00**
13	216.97	14.28	188.98	244.96	0.00	0.93	14.38	87.01	10.43	66.57	107.45	0.00	0.76	−56.75
14	188.94	15.42	158.72	219.16	**2.00**	0.99	**0.00**	65.67	12.94	40.31	91.03	0.00	0.62	−73.17
15	297.78	17.85	262.77	332.75	0.00	0.90	−93.15	144.29	17.07	110.83	177.75	0.00	0.55	−248.15
16	201.32	16.66	168.67	233.97	0.00	0.96	−14.83	63.11	10.47	42.59	83.63	0.00	0.44	−165.17
17	189.74	16.19	158.01	221.47	0.00	0.94	−27.33	57.14	10.85	35.87	78.41	0.00	0.40	−170.39
18	128.33	15.27	98.40	158.26	0.00	0.76	−90.54	28.27	5.83	16.84	39.70	0.00	0.21	−209.10
20	112.28	15.42	82.06	142.50	0.00	0.69	−106.30	23.73	5.30	13.34	34.12	0.00	0.18	−214.68
21	99.75	15.21	69.94	129.56	0.00	0.64	−119.24	18.68	3.49	11.84	25.52	0.00	0.14	−223.28
23	52.97	11.78	29.88	76.06	0.00	0.63	−55.84	8.44	2.63	3.29	13.59	0.00	0.12	−118.31
26	90.29	13.99	62.87	117.71	0.00	0.89	−14.19	16.09	3.13	9.96	22.22	0.00	0.22	−109.68
27	52.04	11.33	29.83	74.25	0.00	0.64	−49.75	8.93	2.34	4.34	13.52	0.00	0.13	−110.48
29	110.51	16.01	79.13	141.89	0.00	0.53	7.93	28.04	9.20	10.01	46.07	0.00	0.67	−14.83
30	350.65	18.27	314.84	386.46	0.00	0.96	−39.44	205.05	16.60	172.51	237.59	0.00	0.72	−131.80

(*Continued*)

TABLE 2. Continued

	MPL							Bayesian (MultiDivTime)						
					Scores							Scores		
Node	Mean	Std	95% Confidence Intervals[a]		SEA	WSS	MDI	Mean	Std	95% Confidence Intervals[a]		SEA	WSS	MDI
1	510.91	17.33	476.94	544.88	1.00	**1.00**	**0.00**	434.46	57.51	321.74	547.18	**2.00**	**1.00**	**0.00**
2	418.52	31.86	356.07	480.97	**2.00**	**1.00**	**0.00**	428.69	66.32	298.70	558.68	**2.00**	**1.00**	**0.00**
6	47.29	36.43	−24.11	118.69	0.00	0.29	−175.11	60.37	21.61	18.01	102.73	0.00	0.29	−191.07
8	89.76	48.73	−5.75	185.27	0.00	0.38	−235.13	136.53	37.46	63.11	209.95	0.00	0.38	−210.45
9	367.13	20.41	327.13	407.13	0.00	0.98	−8.87	348.10	56.98	236.42	459.78	1.00	0.98	**0.00**
11	255.31	21.69	212.80	297.82	0.00	0.88	21.70	303.82	55.68	194.69	412.95	0.00	0.88	3.59
12	227.6	24.80	178.99	276.21	0.00	0.89	4.39	250.74	50.75	151.27	350.21	1.00	0.89	**0.00**
13	131.07	11.91	107.73	154.41	0.00	0.94	−9.79	158.68	38.34	83.53	233.83	**2.00**	0.94	**0.00**
14	96.50	15.66	65.81	127.19	0.00	0.81	−37.01	128.23	32.39	64.75	191.71	1.00	0.81	**0.00**
15	305.05	26.74	252.64	357.46	0.00	0.91	−68.44	260.59	49.70	163.18	358.00	0.00	0.84	−67.90
16	194.35	13.20	168.48	220.22	0.00	0.95	−28.58	188.57	39.74	110.68	266.46	1.00	0.94	**0.00**
17	191.66	14.04	164.14	219.18	0.00	0.94	−29.62	176.74	37.79	102.67	250.81	1.00	0.91	**0.00**
18	126.49	16.07	94.99	157.99	0.00	0.75	−90.81	139.25	31.61	77.29	201.21	0.00	0.80	−47.59
20	107.77	9.52	89.11	126.43	0.00	0.67	−122.37	123.92	28.69	67.69	180.15	0.00	0.74	−68.65
21	105.28	22.91	60.38	150.18	0.00	0.66	−98.62	111.53	26.09	60.39	162.67	0.00	0.69	−86.13
23	67.80	14.75	38.89	96.71	0.00	0.75	−35.19	66.28	20.64	25.83	106.73	0.00	0.74	−25.17
26	97.44	10.20	77.45	117.43	0.00	0.92	−14.47	101.38	24.17	54.01	148.75	1.00	0.93	**0.00**
27	66.61	0.76	65.12	68.10	0.00	0.76	−55.90	69.29	20.56	28.99	109.59	0.00	0.78	−14.41
29	94.4	12.43	70.04	118.76	0.00	0.81	**0.00**	112.71	31.24	51.48	173.94	1.00	0.48	**0.00**
30	345.24	19.97	306.10	384.38	0.00	0.96	−41.52	297.35	53.25	192.98	401.72	0.00	0.90	−24.18

Notes: Values of SEA, WSS, and MDI in this table represent the discordance present between fossil-based and molecular-based estimates of clade origin at each node, not for the entire tree. All calculations used the oldest known fossils for each clade, requiring some ages assigned to nodes by Tinn and Oakley (2008) to be revised (modified ages indicated by [b]). Negative values of MDI indicate molecular divergence estimates are younger than the known fossil record. Best possible values for each metric across nodes are in boldface. LF = Langley-Fitch molecular clock implemented in r8s (Sanderson 2003); MDI = Minimum Divergence Incongruence (this study); MPL = Mean Path Length implemented in PATHd8 (Britton et al. 2006); MultiDivTime = see Thorne and Kishino (2002); PL = penalized likelihood in r8s (Sanderson 2003); SEA = method of Smith et al. (2006); Std = SD; WSS = weighted sum of squares (Tinn and Oakley 2008). [a]95% confidence intervals estimated using the reported mean and SD.[b]Modified fossil age datum.

confidence intervals) (Table 4). DIG range and MIG range values were calculated using the program Assistance with Stratigraphic Consistency Calculations [ASCC: Boyd et al. 2011 v.4.0.0; www.stratfit.org (last accessed December 09, 2014)]. SEA, WSS, and MDI were calculated by hand, as in all previous studies utilizing these metrics.

For reanalysis, several minimum clade ages were changed from Tinn and Oakley (2008) owing to differences in the metric proposed here and interest in implied missing fossil record explicitly. When calculating values of SEA and WSS, Tinn and Oakley (2008) did not always use the oldest known fossil for the clade being dated (Tinn and Oakley 2008: table 1). Their approach prevents a situation where a single fossil OKR influences the minimum divergence age for numerous nodes on the tree. However, it does not prioritize consideration of all of the fossil evidence available to constrain minimum clade age (e.g., the oldest known part of a given clade). For this study, all values were recalculated using the maximum OKR for a clade, and the modified node ages used for reanalysis are noted in Table 2. Uncertainty in the age of the OKR was not reported by Tinn and Oakley (2008). To be able to compare the performance of SEA, WSS, and MDI for the Tinn and Oakley (2008) dataset, uncertainty in the age of each fossil was approximated by setting the age range equal to the temporal range of the geologic stage of the OKR. While ideally these age ranges should be further constrained by data on uncertainty in the assessment of the age of the containing deposit, use of this proxy allowed us to compare the metrics without taking on a detailed reassessment of the ostracod fossil record.

In the analysis of the sphenisciform dataset, fossil-based OKRs and divergence dates were both available for seven nodes in the tree. Values of SEA, WSS, and MDI were calculated for these seven nodes (Table 3), while values of DIG range included the entire tree (Table 4). Online Appendix 1 lists the age data utilized in the MDI and DIG range calculations. The phylogeny utilized for extinct and extant taxa is from Clarke et al. (2010). The combined analysis from that study included the data (RAG-1, mtDNA data) generated by Baker et al. (2006). Only nodes recovered by both Clarke et al. (2010) and Baker et al. (2006) and dated in the latter study were evaluated (i.e., 16 out of the 20 nodes dated).

Sauquet et al. (2012) evaluated the effect of different calibration regimes on resultant time trees. Calibration regimes included sets of phylogenetically vetted or apomorphy-referred fossil records deemed "safe" by the authors, combinations of these records with "risky" fossil calibrations, as well as the use of vicariance events

TABLE 3. Data and resulting SEA, WSS, and MDI scores for the 15 nodes in common between the Clarke et al. (2010) and Baker et al. (2006) studies on Sphenisciformes (penguins) to compare among time trees from the latter's analysis of the single gene (Rag 1), mitochondrial, and combined datasets

				Combined Rag-1+mtDNA						
	—First Fossil—							———Scores———		
Node	Min	Mean	Max	Mean	Std[a]	95% Confidence Intervals[a]		SEA	WSS	MDI
1	60.50	61.05	61.60	77.00	3.37	69.90	83.10	0.00	0.93	8.30
2	60.50	61.05	61.60	70.60	3.80	62.40	77.30	0.00	0.98	0.80
3	11.00	12.00	13.00	40.50	3.42	34.20	47.60	0.00	−4.64	21.20
4	0.00	0.00	0.00	13.50	2.12	9.90	18.20	0.00	–	9.90
5	11.00	12.00	13.00	37.70	3.34	31.60	44.70	0.00	−3.59	18.60
6	0.00	0.00	0.00	19.20	2.17	15.40	23.90	0.00	–	15.40
7	0.00	0.00	0.00	14.10	1.91	10.80	18.30	0.00	–	10.80
8	11.00	12.00	13.00	27.80	3.04	22.50	34.40	0.00	−0.73	9.50
9	11.00	12.00	13.00	25.10	2.83	20.10	31.20	0.00	−0.19	7.10
11	0.00	0.00	0.00	6.10	1.05	4.30	8.40	0.00	–	4.30
12	0.00	0.00	0.00	4.00	0.48	2.90	4.80	0.00	–	2.90
13	0.00	0.00	0.00	3.50	0.56	2.30	4.50	0.00	–	2.30
14	9.70	10.00	10.30	15.30	1.94	11.90	19.50	0.00	0.72	1.60
17	0.00	0.00	0.00	1.40	0.43	0.70	2.40	0.00	–	0.70
19	0.00	0.00	0.00	1.80	0.48	1.00	2.90	0.00	–	1.00

	Rag1							**mtDNA**						
					Scores							Scores		
Node	Mean	Std[a]	95% Confidence Intervals		SEA	WSS	MDI	Mean	Std[a]	95% Confidence Intervals		SEA	WSS	MDI
1	77.10	5.10	65.90	85.90	0.00	0.93	4.30	77.10	3.27	70.30	83.10	0.00	0.93	8.70
2	70.70	5.20	60.00	80.40	1.00	0.98	0.00	68.80	3.47	61.80	75.40	0.00	0.98	0.20
3	40.70	7.19	22.90	51.10	0.00	−4.72	9.90	41.80	3.19	35.70	48.20	0.00	−5.17	22.70
4	14.40	3.34	8.90	22.00	0.00	–	8.90	14.40	2.07	10.80	18.90	0.00	–	10.80
5	37.80	4.85	29.30	48.30	0.00	−3.62	16.30	38.50	3.09	32.80	44.90	0.00	−3.88	19.80
6	20.70	3.95	15.00	30.50	0.00	–	15.00	21.40	2.24	17.50	26.30	0.00	–	17.50
7	15.20	2.96	10.30	21.90	0.00	–	10.30	15.90	1.94	12.50	20.10	0.00	–	12.50
8	27.80	4.26	20.60	37.30	0.00	−0.73	7.60	29.60	2.73	24.60	35.30	0.00	−1.15	11.60
9	25.30	4.26	17.50	34.20	0.00	−0.23	4.50	27.20	2.63	22.50	32.80	0.00	−0.60	9.50
11	5.90	1.48	3.50	9.30	0.00	–	3.50	6.30	0.99	4.60	8.50	0.00	–	4.60
12	4.40	1.38	2.10	7.50	0.00	–	2.10	4.10	0.46	3.10	4.90	0.00	–	3.10
13	3.60	0.77	1.90	4.90	0.00	–	1.90	3.50	0.56	2.40	4.60	0.00	–	2.40
14	16.80	3.11	11.70	19.80	0.00	0.54	1.40	16.00	1.79	12.80	19.80	0.00	0.64	2.50
17	1.50	0.64	0.50	3.00	0.00	–	0.50	1.60	0.48	0.70	2.60	0.00	–	0.70
19	1.80	0.61	0.70	3.10	0.00	–	0.70	1.90	0.43	1.10	2.80	0.00	–	1.10

Notes: Fossil age data is from Clarke et al. (2007, 2010) and papers cited in online Appendix 1, http://dx.doi.org/10.5061/dryad.0vk92. Time tree node ages are from Baker et al. (2006: Table 1). MDI = minimum divergence incongruence (this study); SEA = method of Smith et al. (2006); Std = SD; WSS = weighted sum of squares (Tinn and Oakley 2008). [a]SD estimated using the reported means and 95% confidence interval from Baker et al. (2006) for the purpose of calculating WSS (though see text). The symbol (–) indicates that WSS could not be computed because this method requires a non-zero fossil age and no fossil records for these extant species were available. Abbreviations: see Table 2.

and single secondary calibrations. Posterior probabilities on clade age were assessed for 30 analyses, 15 calibration approaches evaluated in r8s (ML-PL) and BEAST. We compared the set of "safe" fossil OKRs utilized in that study (reported for 14 nodes [A, B, C, D, E, F, G, I,J, K, M, N, Q, and U]: see Table 2 in Sauquet et al. [2012]) to each recovered time tree and calculated ensemble MDI. DIG range was only compared for those nodes with 0 incompatible (negative) scores in MDI comparison, again using only the phylogenetically vetted and apomorphy-based "safe" fossils OKRs. Both 95% posterior densities and uncertainty for the fossil dates were reported by Sauquet et al. (2012), allowing the use of those data in our calculations.

Results

Data from the penalized likelihood (linear) analysis of Tinn and Oakley (2008) are shown in Figure 3. Nodes for

TABLE 4. Ensemble SEA, WSS, and MIG or DIG range values (whole tree) calculated for the ostracod and penguin datasets illustrating the use of these metrics to compare results from distinct datasets (e.g., single gene, mtDNA + Rag1) and analytical approaches

Analysis	Dating Method	SEA	WSS	MIG or DIG Range (myr)		Incompatible Nodes	MDI (whole tree)
TO	Fossil-only (MIG)	–	–	8651.37	8858.08	–	–
	Clock (LF)	0.00	0.60	–	–.	18	−2719.42
	PL (linear)	0.13	0.79	–	–	12	−701.36
	PL (log)	0.13	0.56	–	–	16	−2272.52
	MPL (Pathd8)	0.08	0.81	–	–	15	−1051.43
	Bayesian (MDT)	0.33	0.80	–	–	9	−735.55
C/B	Fossil-only (MIG)	–	–	250.26	425.65	–	–
	Combined	0.00	−0.93	543.81	863.63	0	–[a]
	Rag1	0.00	−0.98	499.15	927.89	0	–[a]
	mtDNA	0.00	−1.18	575.00	872.05	0	–[a]

Notes: In the case of the Tinn and Oakley (2008) dataset, as many as 18 out of 20 nodes have divergence estimates younger than the fossil-based minima. The Bayesian (MDT) analysis was preferred by SEA and that analysis also exhibited the fewest incompatible nodes. For the penguin dataset, while all estimated node ages were significantly older than their fossil-based minima, the Rag1 dataset implied the least amount of missing fossil record. Incompatible nodes are those with molecular dates are younger than the fossil dates. C/B = Baker et al. (2006) time trees with fossil-based data from Clarke et al. (2010); DIG = divergence inferred gap (this study); MIG = minimum implied gap (Pol and Norell 2006); TO = Tinn and Oakley (2008) dataset; SEA = method of Smith et al. (2006); WSS = weighted sum of squares (Tinn and Oakley 2008). The symbol (−) indicates the relevant metric could not be computed. [a]Ensemble MDI scores were not computed for the penguin dataset as that metric only sums the values produced in situations where the divergence dates are younger than the fossil-based ages, and all divergence dates in Baker et al. (2006) are significantly older than the fossil-based minimum ages.

which both fossil-based and molecular-based estimates of cladogenesis were available are labeled in Figure 3 and listed in Table 2, with the numbering matching that used by Tinn and Oakley (2008: figure 1). In Tinn and Oakley (2008), ensemble SEA scores favored the MultiDivTime results (0.80) with the next highest value for Pathd8 (0.30), while the three other analyses had SEA values <0.25. Reanalysis of these data with the updated clade ages yields distinct results. SEA still favors the MultiDivTime results but with an SEA value of only 0.33. The second most favored results are the two penalized likelihood analyses in r8s with values of 0.13, and SEA values for all other analyses were 0.08 and 0. In the original Tinn and Oakley (2008: table 3) findings, calculation of the WSS metric favored the PATHd8 time tree (0.79) with the MultiDivTime tree a close second (0.75). Reanalysis yields similar results for this dataset (WSS values: PATHd8 : 0.81; and MultiDivTime 0.80).

As assessed with MDI, the penalized likelihood (linear) analysis in r8s and the Bayesian MultiDivTime analysis produced the ostracod time trees most congruent with the fossil record. The MDI values for these two analyses were substantially lower than all other analyses, indicating a "negative missing record" of −701.36 and −735.55 myr, respectively (Table 2). However, in the Bayesian MultiDivTime analysis only 9 (of 20) nodes analyzed underestimate the timing of cladogenesis, and for 10 nodes the uncertainty surrounding the fossil-based and molecular-based temporal data overlap. For the penalized likelihood (linear) analysis in r8s, 12 (of 20) nodes underestimate the timing of cladogenesis with only 2 nodes displaying overlapping temporal data (Table 2). The other three time trees had between 15 and 18 younger-than-fossil node ages and a total MDI ranging from −1051.43 to −2719.42 myr. The worst fit was obtained from the "clock" time tree (18 nodes; −2719.42 myr).

Moving to the sphenisciform analysis, none of the divergence dates from Baker et al. (2006) underestimate the timing of cladogenesis implied solely by the fossil record and tree topology, as indicated by the positive MDI values obtained for each node (Table 3). Because none of the divergence dates were younger than the fossil-based dates for the nodes, ensemble MDI values were not computed because only negative scores are summed when describing fit across the entire tree. In this regard, that dataset represents an opposite case from the Tinn and Oakley (2008) dataset and here MDI at individual nodes, SEA, WSS, and DIG range values are compared. All nodes analyzed for the sphenisciform tree produced SEA values of 0 regardless of the set of molecular-based temporal data used (Table 3) because most of the divergence dates are substantially older than those implied by the fossil record (Figure 4). However, congruence between the two sets of temporal data is reasonable at some of the individual nodes, with some nodes (e.g., node 2) producing low values for MDI and high values for WSS (Table 3). WSS values calculated for the total time trees are negative because, on average, the mean divergence dates are more than twice as old as the mean fossil-based dates, indicating a poor fit to the fossil-based data. Divergence dates resulting from analysis of the combined RAG-1 + mtDNA dataset have the least negative WSS values (−0.93). The RAG-1-only WSS score is similar (−0.98) with the mtDNA-only dataset producing the poorest fit by this measure (−1.18: Table 4).

Comparison of the MIG range and DIG range values for the sphenisciform dataset indicates that enforcing the molecular-based estimates of cladogenesis approximately doubles the amount of implied missing fossil record for the clade, regardless of the molecular

FIGURE 3. Chronograms comparing (dashed) the ostracod time tree from the penalized likelihood results (lin, r8s; Tinn and Oakley 2008: figure 1) and (solid) estimates of cladogenesis based on a minimum fit to the fossil record utilized by those authors with adjusted minimum clade ages (see "Methods" section). As assessed with MDI, the penalized likelihood (lin) analysis in r8s and Bayesian MultiDivTime analysis produced the time trees most congruent with the fossil record; MDI values were substantially lower for these two analyses than all other analyses. MDI for the penalized likelihood (linear) analysis were least negative, but 12 (of 20) nodes yield molecular dates younger than the fossil minimum ages with only 2 nodes displaying overlapping temporal data. In comparison, the MultiDivTime time tree contains only 9 nodes that are younger than the fossil minimum ages (Tables 2 and 4). The previously proposed metric WSS preferred the Pathd8 tree, which had a highly negative MDI value and 15 molecular estimates younger than fossil minima. Numbers next to the 20 compared nodes for which both fossil and molecular-based estimates were available reflect the referenced node numbers (1–30) in Tinn and Oakley (2008) and are cited in Table 2. Grey boxes represent 1.96 SDs to approximate 95% confidence intervals for each molecular date because Tinn and Oakley (2008) did not report 95% confidence intervals.

dataset used (Table 4), which agrees with the results obtained by Clarke et al. (2007: p. 11549). However, utilizing DIG range to compare the relative fit of dates inferred from the molecular datasets provides insight beyond this previously reported result. Enforcing the dates from the mtDNA only dataset implies the greatest missing fossil record, followed by the combined RAG-1 +mtDNA dataset (Table 4). The RAG-1 dataset exhibits the closest fit to the fossil-based dates, resulting in a DIG range of 499.15–927.89 myr (Table 4).

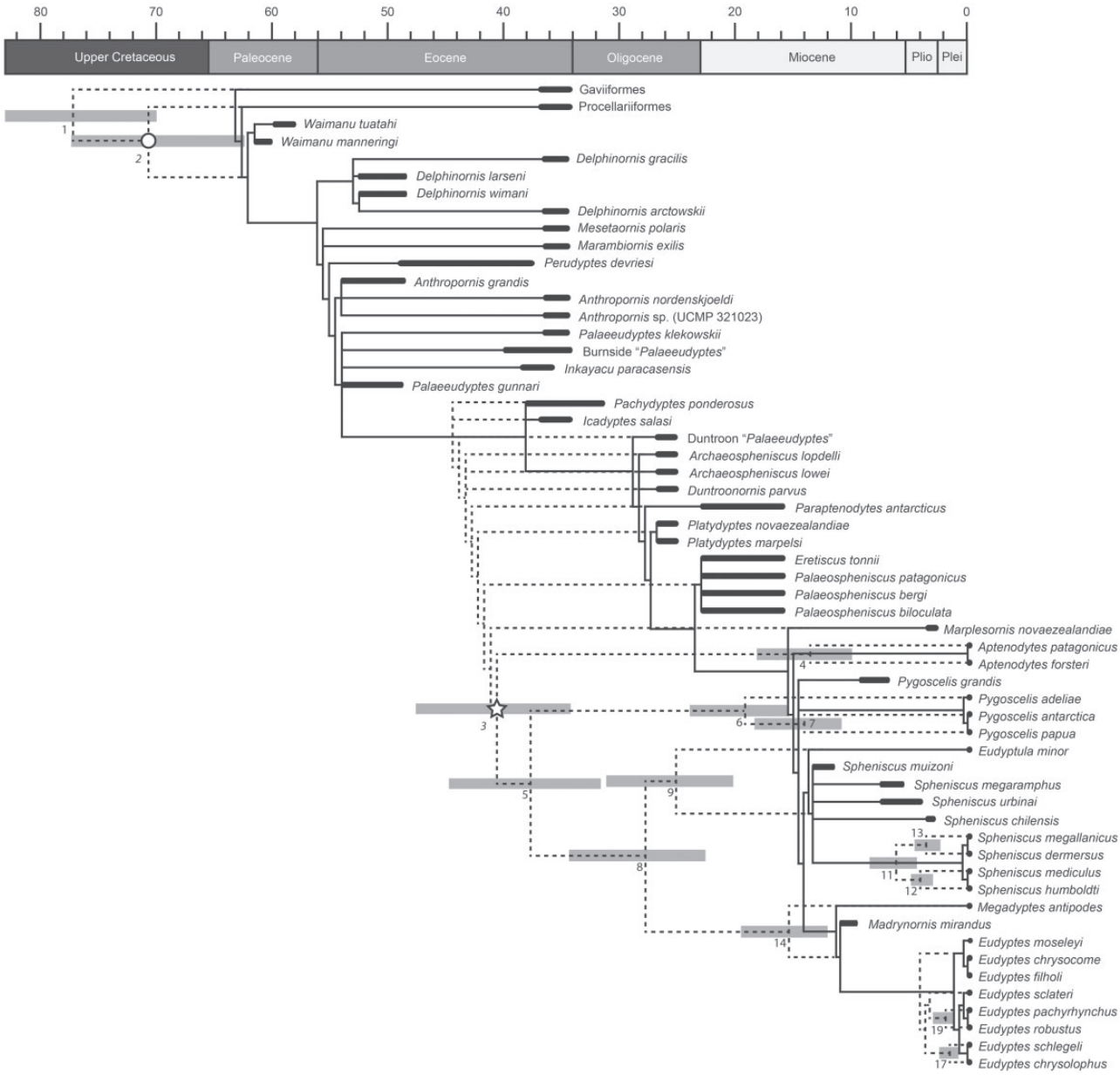

FIGURE 4. Chronograms comparing (dashed) a penguin time tree when enforcing dates from Bayesian analysis of the combined RAG-1+mtDNA dataset of Baker et al. (2006: table 1) and (solid) estimates of cladogenesis based on a minimum fit to the fossil record (see online Appendix 1 for fossil ages). The divergence of penguins from Procellariformes is indicated with a circle (node 2) and the penguin crown clade is indicated by a star (node 3). Grey boxes represent the 95% confidence intervals for each divergence date. The RAG-1-only time tree yielded the lowest estimate of DIG range followed by the combined RAG-1+mtDNA time tree (shown). The combined data time tree was utilized by Clarke et al. (2007) in their proposal of a precursor metric to DIG range (Table 1). Node numbers are from Baker et al. (2006: table 1).

Sauquet et al. (2012) chose not to make recommendations concerning specific calibration strategies or dating methods, instead focusing on the need to evaluate sets of calibrations and approaches as another way of bounding uncertainty on recovered time trees. Significant differences in the posterior probabilities of clade ages were recovered under distinct calibration regimes and with distinct methods (Sauquet et al. 2012). Comparing MDI values across the 30 analyses shows that six analyses did not yield any divergence dates younger than the any of the 14 "safe" OKRs (Table 5; BEAST analyses with calibration sets 1, 3, 4, and 6; ML-PL with calibration sets 1 and 4). Unsurprisingly, these six analyses included one calibration regime (1) that used the total set of 14 fossil OKRs considered "safe". Two other calibration regimes used sets of these "safe" calibrations but with the addition of calibrations deemed "risky" (4, 6) and one used a subset of "safe" calibrations only for the outgroup (3). The rest of the 24 analyses

TABLE 5. Ensemble MDI and DIG range values for *Nothofagus* time trees from Sauquet et al. (2012) illustrating the use of these metrics to compare results from distinct calibration regimes and analytical approaches

Calibration scenario	Analysis method	Marginal likelihoods	# of Incompatible Nodes	MIG range (in Ma)	DIG range (in Ma)	MDI (in Ma)
0	BEAST	−20375.94	9	–	–	−172.4
1	BEAST	−20375.59	0	1228.2-1297.1	1773.5-2253.8	–
2	BEAST	−20375.24	5	–	–	−91.4
3	BEAST	−20375.67	0	1228.2-1297.1	1746.6-2212.6	–
4	BEAST	−20376.04	0	1228.2-1297.1	2078.2-2685.5	–
5	BEAST	−20375.37	3	–	–	−46.7
6	BEAST	−20375.80	0	1228.2-1297.1	1762.8-2229.2	–
7	BEAST	−20374.94	2	–	–	−34.3
8	BEAST	−20376.08	13	–	–	−292.6
8a	BEAST	−20376.01	13	–	–	−284.3
8b	BEAST	−20375.97	9	–	–	−141.4
8c	BEAST	−20375.83	8	–	–	−131.6
8d	BEAST	−20376.17	13	–	–	−309.8
8e	BEAST	−20375.66	6	–	–	−112.6
8f	BEAST	−20375.37	8	–	–	−128.4
0	ML-PL	N/A	14	–	–	−342.7
1	ML-PL	N/A	0	1228.2-1297.1	1642.3-1801.9	–
2	ML-PL	N/A	8	–	–	−217.5
3	ML-PL	N/A	3	–	–	−35.1
4	ML-PL	N/A	0	1228.2-1297.1	1921.3-2052.3	–
5	ML-PL	N/A	7	–	–	−188.3
6	ML-PL	N/A	3	–	–	−33.3
7	ML-PL	N/A	7	–	–	−158.1
8	ML-PL	N/A	14	–	–	−369.8
8a	ML-PL	N/A	14	–	–	−357.7
8b	ML-PL	N/A	11	–	–	−265.7
8c	ML-PL	N/A	11	–	–	−269.3
8d	ML-PL	N/A	9	–	–	−188.9
8e	ML-PL	N/A	10	–	–	−202.3
8f	ML-PL	N/A	13	–	–	−277.1

Notes: Six of the analyses in Sauquet et al. (2012) yielded 0 incompatible nodes; thus, ensemble MDI was not calculated for those analyses; DIG range values are compared for these analyses (see text). Marginal likelihoods from the BEAST analyses were generously provided by the authors. Marginal likelihoods are not reported by the program r8s, so these values were unavailable for the ML-PL analyses. The symbol (–) indicates no incompatible nodes were recovered; therefore, ensemble MDI is not calculated in those situations.

yielded at least two clade ages younger than these 14 fossils.

DIG range values quantifying implied ghost lineages are computed for analyses with no nodes younger than fossil records. The implied increase in missing fossil record for all six of the time trees when compared to the minimum fit tree (MIG) was relatively low (e.g., compare the sphenisciform results) comprising a small percent of the DIG range. The minimum estimates of implied missing fossil record for the six time trees compared ranged from 1642.3 to 2078.2 million years of missing record across the clade compared to a MIG range of 1228.2–1297.1. Calibration set 1 in r8s (ML-PL), which used all 14 "safe" fossils as calibrations, yielded divergence dates minimizing the implied missing fossil record for these compared nodes. Calibration sets 1, 3, and 6 assessed in BEAST yielded only slightly higher minimum values, though the range of DIG values are broader for the BEAST analyses than the ML-PL analyses (Table 5).

In the BEAST time trees, as expected, some minimum molecular age estimates approach the fossil calibration minimum age, but they were never identical. However, the reported r8s (ML-PL) results sometimes included a single molecular divergence date (i.e., not a range of possible ages) that was equal to the minimum age of the fossil calibration. The latter situation resulted in a narrower DIG range for the ML-PL analyses than the BEAST analyses (Table 5). R8s (ML-PL) analyses also often yielded narrower confidence intervals for the molecular ages than BEAST. This characteristic also resulted in less inclusive DIG range values for ML-PL analyses than for BEAST analyses for the same scenario (Table 5).

Discussion

While the case studies discussed herein serve to illustrate differences between the new and previously proposed metrics, they also highlight particular models or datasets that better fit the compared fossil records. Due to the nature of the datasets chosen, the first example illustrates use of MDI, the second, DIG range, and the third how these metrics may be used together. Penalized likelihood (linear) analysis in r8s and the Bayesian MultiDivTime analysis yielded ostracod time trees most congruent with the fossil record as assessed

by MDI (Table 2). Of these two best-fit time trees, the MultiDivTime analysis had the fewest estimated divergence dates younger than compared fossils and the most nodes with overlapping confidence intervals. If the methods proposed here were utilized to choose among time trees, the MultiDivTime results are to be preferred. By contrast, PATHd8 results were preferred by the previously proposed metrics (SEA, WSS; Tinn and Oakley 2008) (Table 4).

In the penguin case study, MDI values were all >0, and comparison of DIG range values for time trees from analysis of distinct molecular datasets highlights that the RAG-1 dataset marginally fit the known fossil record better. DIG highlights the difference among datasets in the implied incompleteness of a fossil record considered to be one of the best within Aves (Fig. 4). Enforcement of divergence dates changed the perceived shape of early stem penguin diversification, indicating apparent explosive radiation of all stem group penguins in the early Paleogene (Clarke et al. 2007). While there were only a small number of fossils relevant to calibrating a penguin time tree comprised of only extant species, there was a rich fossil record with more than thirty well-preserved fossils whose relationships were resolved in phylogenetic analyses. Enforcing even a small number of divergence dates modified the hypothesized minimum age of cladogenesis for numerous species-rich extinct clades in the tree. DIG values draw attention to major differences in the estimated pattern of early penguin diversification with the enforcement of small number of estimated divergence dates for the crown.

Analysis of the Sauquet et al. (2012) dataset examined differences in fit among calibration regimes and distinct methods. The phylogenetically vetted fossil OKRs deemed "safe" by these authors were compared to resultant time trees. Only 6 of the 30 analyses from that study required that none of these fossil OKRs be an incorrect minimum estimate of clade origin. Seven analyses utilizing a single secondary calibration and one vicariance-based approach with six calibrated nodes recovered no fewer than eight node ages significantly younger than the "safe" fossils, and several analyses required nearly all of the 14 fossils to be incorrectly assessed minimum estimates of clade age.

We agree with Sauquet et al. (2012) in their conclusions that the basis for use of a fossil as a clade age minimum should be specified and that the effect of calibration regime on node age estimates should be fully explored. To this we add that consideration of the number of underestimates of clade age may be another useful descriptor of time trees. While we were only able to compare estimated likelihoods for the BEAST models, they are very similar ($\leq$ 2 Bayes Factors difference; Table 5) but, as noted, vary markedly in how many prior hypotheses about fossil placement must be wrong for estimated divergence dates to be correct.

Unsurprisingly, MDI scores highlight that fit with the fossil record is more commonly, but not exclusively, yielded by analyses including these same "safe" fossils as calibrations. These models yielded a better fit regardless of approach (ML-PL or Bayesian) used. To prefer models that included all of the compared "safe" fossils based on MDI score would be circular. However, they serve as a base in this case to compare the rest of the results. Of the four analyses where from 4 to 10 of these 14 safe calibrations were deployed, variably with other "risky" calibrations, only two yielded time trees requiring no divergence estimate to be younger than the compared OKRs. Which analysis was recovered with no negative MDI values varied by estimation method (ML-PL calibration set 3 and BEAST calibration set 6) but both used only outgroup calibrations. Interestingly, analyses that only used ingroup calibrations (scenarios 2 and 5) consistently recover divergences in the outgroup younger than "safe" outgroup fossils.

In this case, while MDI was a a useful descriptor of the fit of distinct models to the set of vetted fossils, DIG was less useful; as predicted, the models with all of these fossils included as calibrations showed a better fit, but differences in DIG range were limited (e.g., differences among all six time trees approximated differences from MIG). As treated in the Results, marginal differences in DIG range values may not be as informative, especially when comparing results from different dating methods.

Conclusions and Perspectives

There has been extensive discussion of fossil calibration choice, divergence dating methods, and cross-validation approaches, but few attempts to quantify the fit of estimated time trees to the known fossil record after calibration. Some might question whether this is appropriate or desirable. If fossils are carefully chosen and vetted for analysis then why should they or other fossil records be compared after time trees are generated? Or, why should the misfit between minimum ages used as priors in analysis be compared to posterior estimates of node age (e.g., via DIG range)? We believe that proposed fossil oldest known occurrences constitute prior hypotheses concerning the clade of interest and knowing how many of these hypotheses must be positively false for a time tree to be true is worth quantifying. Comparisons among time trees of the implied extent of missing fossil record in a clade may additionally productively inform estimates of the relative quality of this record or the degree of incongruence in hypotheses concerning the timing of radiation.

We recommend using the metrics MDI and DIG range to quantitatively compare congruence between the fossil record and multiple sets of divergence dates involving distinct calibration sets (e.g., Douzery et al. 2003), different methods for implementing fossil calibration points (e.g., Yang and Rannala 2006), different molecular clock models (e.g., Douzery et al. 2004; Tinn and Oakley 2008), or for comparing different sets of molecular-based temporal data generated using different genes to date the same clade (e.g., Baker et al. 2006). Ultimately, interpretation of MDI and DIG range scores will depend

on the set of assumptions employed by individual researchers.

All methods so far proposed to model fossil preservation and recovery indicate that the earliest part of a clade is least likely to be recovered (e.g., Marshall 1997, 1998; Weiss et al. 2003). The probability of fossilization early in a lineage's history is expected to be lower as the lineage either would not be widespread and individual-rich (e.g., Marshall 1998), or it would not be morphologically distinct (Cooper and Fortey 1998). The proposed systematic effect of hard part sampling in the rock record on the phylogenetic placement of fossil calibrations also supports the use of these fossils as minima (Sansom and Wills 2013). Indeed, proposed taphonomic models also consistently focus on assessing the extent of an older missing or unsampled fossil record (e.g., Behrensmeyer et al. 2000, see also discussion in Near et al. 2005). If this characterization of the fossil record is assumed to be correct, then time trees with the best MDI (i.e., fewest incompatible nodes with fossil OKRs predating estimated clade origin) should be preferred.

DIG values quantify how much of the fossil record of a clade is not yet known. We would expect better explanations of the data to minimize these values. This metric differs importantly from MDI, and must be seen as secondary to it, as poorer fit as assessed in DIG range does not require any of the published hypotheses concerning the fossil record to be wrong. DIG range values are also sensitive to the absolute age of the clade under consideration in a way that the number of MDI-assessed incongruent nodes is not. Furthermore, as we discuss, comparison of DIG range among time trees generated by different divergence dating methods should take into account potential systematic differences in the breadth of associated confidence intervals (see "Discussion" section).

The metrics proposed herein may be used in conjunction with similar previously described metrics (e.g., SEA, WSS), but have additional benefits. In particular, they are the first metrics that take into account both stratigraphic uncertainty in both fossil age and confidence bounds on the molecular dates as well as distinguishing between cases in which divergence dates are younger than or older than the known fossil record.

All of these metrics differ from other approaches that have assessed the fit among a set of calibrations and ultrametric trees in the process of time tree estimation (e.g., Near et al. 2005). The ways in which they differ illuminates the intention behind their development. First, they are applied descriptively after a set of calibration regimes are chosen based on criteria that assess the fossil data themselves (e.g., phylogenetically vetted, apomorphy-based referral) and are not intended to inform decisions about calibration choice. Incompatible nodes, as assessed in MDI comparison, are not interpreted as implying that the fossil records are unreliable or need to be excluded from analysis. As shown here, the number of incompatible nodes can vary markedly among analyses with the same calibration set but differing in divergence dating approach (Table 5). Aspects of the model, the shape of prior associated with a given fossil-based minimum age and other aspects of analysis equally may be involved in explaining the observed misfit. While simultaneous estimation of a time tree of life including all fossil and extant taxa is ideal, we hope in its absence that these metrics might be used to quantify our improvement in generating hypotheses that can explain both sequence-based branch lengths and the often separately assessed data from the fossil record.

Supplementary Material

Supplementary material can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.0vk92.

Funding

This work was supported by the National Science Foundation (grant numbers DEB 0949897, DEB 1355292) and the Jackson School of Geosciences at The University of Texas at Austin.

Acknowledgments

We thank Hervé Sauquet for critical discussion of the metrics proposed and suggestion of an additional target dataset. We additionally thank H. Sauquet and S. Ho for providing additional data on their analyses. We thank M. Benton, T. Cleland, A. Debee, D. Eddy, M. Householder, S. Nesbitt, T. Oakley, C. Brochu, D. Pol, N.A. Smith, W. Thompson, D. Wills, and one anonymous reviewer for comments on previous versions of this manuscript, and M. Norell for discussion on the metrics.

References

Baker A.J., Pereira S.L., Haddrath O.P., Edge K.A. 2006. Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. Proc. R. Soc. Lond. B Biol. Sci. 217:1117.

Behrensmeyer A.K., Kidwell S.M., Gastaldo R.A. 2000. Taphonomy and paleobiology. Paleobiology. 26:103–147.

Benton M. J. 1994. Paleontological data and identifying mass extinctions. Trends Ecol. Evol. 9:181–185.

Boyd C.A., Cleland T.P., Marrero N.L., Clarke J.A. 2011. Exploring the effects of phylogenetic uncertainty and consensus trees on stratigraphic consistency scores: a new program and a standardized method proposed. Cladistics. 27:52–60.

Britton T., Anderson C.L., Jaquet D., Lundquist S., Bremer K. 2006. PATHd8 — a new method for estimating divergence times in large phylogenetic trees without a molecular clock. Available from: http://www.math.su.se/PATHd8 (last accessed December 9, 2014).

Brochu C.A., Sumrall C.D., Theodor J.M. 2004. When clocks (and communities) collide: estimating divergence time from molecules and the fossil record. J. Paleontol. 78:1–6.

Clarke J.A., Ksepka D.T., Stucchi M., Urbina M., Giannini N., Bertelli S., Narvaez Y., Boyd C.A. 2007. Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. Proc. Natl. Acad. Sci. USA. 104:11545–11550.

Clarke J.A., Ksepka D.T., Salas-Gismondi R., Altamirano A.J., Shawkey M.D., D'Alba L., Vinther J., DeVries T.J., Baby P. 2010. Fossil evidence for evolution of the shape and color of penguin feathers. Science. 330:954–957.
Cooper A., Fortey R. 1998. Evolutionary explosions and the phylogenetic fuse. Trends Ecol. Evol. 3:151–156.
Donoghue C.J., Smith M.P. 2003. Telling the evolutionary time: molecular clocks and the fossil record. New York: Taylor & Francis.
Douzery E.J.P., Delsuc F., Stanhope M.J., Huchon D. 2003. Local molecular clocks in three nuclear genes: divergence times for rodents and other mammals and incompatibility among fossil calibrations. J. Morphol. Evol. 57:S201–S231.
Douzery E.J.P., Snell E.A., Bapteste E., Delsuc F., Philippe H. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils. Proc. Natl. Acad. Sci. USA. 101:15386–15391.
Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.
Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.
Hug L.A., Roger A.J. 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. Mol. Biol. Evol. 24:1889–1897.
Laurin M. 2004. The evolution of body size, Cope's rule and the origin of amniotes. Syst. Biol. 53:594–622.
Lee M.S.Y., Oliver P.M., Hutchinson M.N. 2009. Phylogenetic uncertainty and molecular clock calibrations: a case study of legless lizards (Pygopodidae, Gekkota). Mol. Phylogenet. Evol. 50:661–666.
Lukoschek V., Keogh J.S., Avise J.C. 2012. Evaluating fossil calibrations for dating phylogenies in light of rates of molecular evolution: a comparison of three approaches. Syst. Biol. 61:22–43.
Maddison W.P., Maddison D.R. 2005. Mesquite: A modular system for evolutionary analysis. Version 1.06. Available from: http://mesquiteproject.org.
Marjanovic D., Laurin M. 2007. Fossils, molecules, divergence times, and the origin of lissamphibians. Syst. Biol. 56:369–388.
Marshall C.R. 1990. Confidence intervals on stratigraphic ranges. Paleobiology. 16:1–10.
Marshall C.R. 1997. Confidence intervals on stratigraphic ranges with nonrandom distributions of fossil horizons. Paleobiology. 23:165–173.
Marshall C.R. 1998. Determining stratigraphic ranges. In: Donovan S.K., Paul C.R.C., editors. The adequacy of the fossil record. Chichester (UK): Wiley. p. 23–53.
Marshall C.R. 2008. A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. Am. Nat. 171:726–742.
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 Biol. Sci. 359:1477–1483.
Near T.J., Meylan P.A., Shaffer H.B. 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. Am. Nat. 165:137–146.
Norell M.A. 1992. Taxic origin and temporal diversity: the effect of phylogeny. In: Novacek M.J., Wheeler Q.D., editors. Extinction and phylogeny. New York (NY): Columbia University Press. p. 88–118.
Parham J.F., Phillip C., Donoghue J., Bell C.J., Calway T.D., Head J.J., Holyroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61(2):346–359.
Pol D., Norell M.A. 2006. Uncertainty in the age of fossils and the stratigraphic fit to phylogenies. Syst. Biol. 55:512–521.
Pol D., Norell M.A., Siddall M.E. 2004. Measures of stratigraphic fit to phylogeny and their sensitivity to tree size, shape and scale. Cladistics. 20:64–75.
Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn, A.P. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61: 973–999.
Rutschmann F., Eriksson T., Salim K.A., Conti E. 2007. Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points. Syst. Biol. 56:591–608.
Sauquet H., Ho S., Gandolfo M., Jordan G., Wilf P., Cantrill D., Bayly M., Bromham L., Brown G., Carpenter R., Raymond J. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). Syst. Bio. 61:289–313.
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.
Sansom R.S., Wills M.A. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Scientific Rep. 3. doi:10.1038/srep02545.
Smith A.B., Pisani D., Mackenzie-Dodds J.A., Stockley B., Webster B.L., Littlewood T.J. 2006. Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). Mol. Biol. Evol. 23:1832–1851.
Thorne J.L., Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689–702.
Tinn O., Oakley T.H. 2008. Erratic rates of molecular evolution and incongruence of fossil and molecular divergence time estimates in Ostracoda (Crustacea). Mol. Phylogenet. Evol. 48:157–167.
Walsh S.L. 1998. Fossil datum and paleobiological event terms, paleontostratigraphy, chronostratigraphy, and the definition of land mammal "age" boundaries. J. Vert. Paleontol. 18:150–179.
Weiss R.E., Basu S., Marshall C.R. 2003. A framework for analyzing fossil record data. In: Buck C.E., Millard A.R., editors. Tools for constructing chronologies: crossing disciplinary boundaries. London (UK): Springer. p. 15–232.
Warnock R.C., Yang Z., Donoghue P.C. 2012. Exploring uncertainty in the calibration of the molecular clock. Biol. Lett. 8:156–159.
Wills M.A. 1999. Congruence between stratigraphy and phylogeny: randomization tests and the gap excess ratio. Syst. Biol. 48:559–580.
Wills M.A., Barrett P.M., Heathcote J.F. (2008) The modified gap excess ratio (GER*) and the stratigraphic congruence of dinosaur phylogenies. Syst. Biol. 57:891–904.
Xia X., Yang Q. 2011. A distance-based least-square method for dating speciation events. Mol. Phylogenet. Evol. 59:342–353.
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.