The Trouble with Topology: Phylogenies without Fossils Provide a Revisionist Perspective of Evolutionary History in Topological Analyses of Diversity

JAMES E. TARVER1,2,* AND PHILIP C. J. DONOGHUE1,*

1School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen’s Road, Bristol, BS8 1RJ, UK; and
2Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA;

*Correspondence to be sent to: School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen’s Road, Bristol, BS8 1RJ, UK;
E-mail: james.tarver@bristol.ac.uk; phil.donoghue@bristol.ac.uk.

Received 10 February 2010; reviews returned 7 May 2010; accepted 7 January 2011
Associate Editor: Norman MacLeod

The decidedly asymmetrical architecture of the Tree of Life betrays the fact that some evolutionary lineages have diversified to a greater or lesser extent than have others. It has been a goal of evolutionary ecology to identify shifts in diversification rate (speciation rate minus extinction rate) and their causal bases. Historically, the question of relative diversity was addressed in terms of phylogenies of fossils (Simpson 1944) with focus shifting to taxonomies in the 1970s (Raup et al. 1973; Gould et al. 1977). The dramatic rise in the availability of molecular sequence data in the 1990s led to renewed interest in phylogenies and diversity from evolutionary biologists and attempts to explain patterns of diversity returned to phylogenies. For the first time, however, historical events inferred to have shaped current diversity were identified without recourse to palaentological data (Harvey et al. 1994).

Two principal approaches have been developed—"distance" and "topological" methods. Distance (aka "temporal") methods represent a natural development of taxonomic methods that were initially aphylogenetic. They exploit branch length data or the temporal spacing of branching events within phylogenetic trees and while distance methods remain popular for analyzing rates of evolution, they have lost out in favor of topological methods for the analysis of the history of clade diversity.

Topological (aka "tree shape") methods were developed in parallel with distance methods but are distinguished in that they exploit only tree balance as a record of evolutionary history (Slowinski and Guyer 1989). They have been adopted widely because they eschew temporal and distance data and so they can be applied readily to phylogenetic trees, such as supertrees, where distance data are lacking (Moore et al. 2004). The purpose of these methods is to determine 1) if lineages within a given phylogenetic tree diversified under different rates and 2) which particular lineages within a given phylogenetic tree are more diverse than would be expected under a given model, such as a Yule model (Yule 1924). These methods were not, however, conceived to identify the causal bases of tree imbalance but merely to identify whether these phenomena might exist in the first place (Raup 1985; Slowinski and Guyer 1989). In this endeavour, topological methods perform their task admirably. However, in practice, the overwhelming majority of studies (Table 1) that have employed topology-based methods have sought to identify causality “to know if shifts in diversification rate are correlated with changes in some other variable (e.g., the origin of morphological or behavioral novelties, ecological associations, biogeographic events)” (Moore et al. 2004; p. 524). Intrinsic causal factors are identified through coincidence with diversification rate shifts in tree topology and extrinsic causes are identified through their temporal correspondence to the age of the node on which a diversification rate shift is identified. Increasingly sophisticated methods are being developed to test the coincidence between the putative causal factors and the diversification rate shifts to which they are attributed (Moore and Donoghue 2007, 2009). However, the modus operandi is invariably inductive—first identifying diversification rate shifts and then seeking their causal bases rather than deductive—testing hypotheses of causal association between innovations and their impact upon diversity (but see Moore and Donoghue 2009).

Ultimately, hypotheses of causality rest on the intuitively reasonable assumption that the position of diversification rate shifts within the topology of a phylogenetic tree reflects their relative timing. Thus, diversification rate shifts identified near the root or the tips of a tree are considered to have occurred early or late, respectively, within the evolutionary history of the clade. In this contribution, we demonstrate that this assumption is not a natural expectation of phylogenetic trees of standing diversity. The topologies of census trees are in constant flux as taxa are added by speciation and pruned by extinction. Nonrandom imbalances in the diversity of sister clades need not be...
Table 1. All studies which use either the shift statistic (SymmeTREE) or the relative Cladogenesis statistic (End-Epi) to seek a causal explanation for perceived shifts in diversification rates. Additional papers have cited these two programs, however, they are either reviews or use other features implemented within said programs. Likewise, other methods (Rabosky et al. 2007; Alfaro et al. 2009) have been developed which are subject to the same limitations.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Clade</th>
<th>Software used</th>
<th>Causal explanation sought</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steeman et al. 2009</td>
<td>Cetaceans</td>
<td>SymmeTREE</td>
<td>Palaeoceanographic changes</td>
</tr>
<tr>
<td>Kazancioglu et al. 2009</td>
<td>Parrotfishes</td>
<td>SymmeTREE</td>
<td>Sexual selection</td>
</tr>
<tr>
<td>Li et al. 2009</td>
<td>Schizoflorax fishes</td>
<td>End-Epi</td>
<td>Geographic elevation</td>
</tr>
<tr>
<td>Egan and Crandall 2008</td>
<td>Flowering plants (Psoraleaeae)</td>
<td>SymmeTREE</td>
<td>Climate change</td>
</tr>
<tr>
<td>Tolley et al. 2008</td>
<td>South African chameleons</td>
<td>SymmeTREE</td>
<td>Climate change</td>
</tr>
<tr>
<td>Merckx et al. 2008</td>
<td>Flowering plants (Burmanniaceae)</td>
<td>SymmeTREE</td>
<td>Climate change</td>
</tr>
<tr>
<td>Williams and Duda 2008</td>
<td>Gastropods</td>
<td>SymmeTREE and End-Epi</td>
<td>Tectonic activity</td>
</tr>
<tr>
<td>Hardman and Hardman 2008</td>
<td>Bullhead catfishes</td>
<td>SymmeTREE and End-Epi</td>
<td>Climate change</td>
</tr>
<tr>
<td>Ruta et al. 2007</td>
<td>Tetanospondyls</td>
<td>SymmeTREE</td>
<td>Recovery after mass extinctions</td>
</tr>
<tr>
<td>Moore and Donohue 2007</td>
<td>Flowering plants (Adoxaceae and Valerianaeae)</td>
<td>SymmeTREE and End-Epi</td>
<td>Biogeographic and morphological changes</td>
</tr>
<tr>
<td>Verdu et al. 2007</td>
<td>Multiple plant clades</td>
<td>SymmeTREE</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Forest et al. 2007</td>
<td>Milkwoks</td>
<td>End-Epi</td>
<td>Morphological novelty</td>
</tr>
<tr>
<td>Wiens 2007</td>
<td>Tropical salamanders</td>
<td>SymmeTREE</td>
<td>Geographic elevation</td>
</tr>
<tr>
<td>McLeish et al. 2007</td>
<td>Actinia thrips</td>
<td>End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Wiens et al. 2006</td>
<td>North American salamanders</td>
<td>SymmeTREE and End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>McKenna and Farrell 2006</td>
<td>Leaf beetles</td>
<td>End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Davison and Chiba 2006</td>
<td>Land snails</td>
<td>End-Epi</td>
<td>Climate change</td>
</tr>
<tr>
<td>Kozak et al. 2006</td>
<td>North American salamanders</td>
<td>SymmeTREE</td>
<td>Coevolution with flowering plants</td>
</tr>
<tr>
<td>Jones et al. 2005</td>
<td>Bats</td>
<td>End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Kozak et al. 2005</td>
<td>Dusky salamanders</td>
<td>End-Epi</td>
<td>Sea level fluctuations</td>
</tr>
<tr>
<td>Near and Benard 2004</td>
<td>Logperch darters</td>
<td>End-Epi</td>
<td>Climate change</td>
</tr>
<tr>
<td>Jennings et al. 2003</td>
<td>Australian lizards</td>
<td>End-Epi</td>
<td>Behavior</td>
</tr>
<tr>
<td>Ruber et al. 2003</td>
<td>American gobies</td>
<td>End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Vences et al. 2002</td>
<td>Malagasy treefrogs</td>
<td>End-Epi</td>
<td>Adaptive radiation</td>
</tr>
<tr>
<td>Bininda-Emonds et al. 1999</td>
<td>Carnivores</td>
<td>End-Epi</td>
<td>Chromosomal variation</td>
</tr>
<tr>
<td>Fumaragi et al. 1999</td>
<td>Shrews</td>
<td>End-Epi</td>
<td>Life history traits</td>
</tr>
<tr>
<td>Cook and Lessa 1998</td>
<td>Rodents</td>
<td>End-Epi</td>
<td>Climate change</td>
</tr>
<tr>
<td>Price et al. 1998</td>
<td>Warblers</td>
<td>End-Epi</td>
<td>Climate change</td>
</tr>
</tbody>
</table>
achieved by singular events or episodes of diversification, as topological methods presuppose. Nonrandom imbalances can also be achieved through temporally and causally unrelated episodes of random diversification (speciation and/or extinction), leading to spurious hypotheses of diversification rate shift. The contribution of extinction to standing diversity is especially critical, telescoping past flux in speciation and extinction into a single internal branch of a tree—an effect exploited in lineage through time analyses (Harvey et al. 1994). In this way, the summed effects of random processes can produce nonrandom tree topologies.

If the veracity and timings of diversification rate shifts are to be correctly identified, we argue that it is necessary to distinguish the relative timing of the contributory episodes of speciation and extinction. In achieving this objective, the architecture of trees must be considered in terms of extinct and not merely extant taxa. In effect, it is necessary to shift focus from the investigation of diversification rate, to speciation rate, which better fits the pure birth Yule model that underpins the majority of topology-based methods. Only once this has been established will it be possible to speculate on, and test among, putative causal bases underpinning the flux of tree shape over the evolutionary history of clades.

**MATERIALS AND METHODOLOGY**

Using empirical data sets, we demonstrate the effect of extinction upon tree topology, perceived shifts in diversification rate, and the relative timing of the action of extinction versus its perceived impact on a tree. We then show how the integration of extinct among extant members of a clade makes it possible to distinguish between the contributory effects of speciation and extinction to diversity, to distinguish genuine from artifactual diversification rate shifts, and establish the timing of their causal drivers. For our demonstration, we used two empirical data sets, a supertree of extant carnivore relationships (Bininda-Emonds et al. 2007) to examine the impact of extinction, and a phylogenetic tree of extant and extinct crocodilian relationships to examine the impact of speciation (Gatesy et al. 2004). Our focus is on the identification of diversification rate shifts because, in terms of the relative timing of the events they predict, they are more readily testable than other topology-based methods that, for instance, focus exclusively on tree symmetry. Of the two approaches most commonly used to identify diversification rate shifts, End-Epi (Rambaut et al. 1997) is no longer distributed and runs on an obsolete Macintosh operating system and SymmeTREE (Chan and Moore 2005). SymmeTREE has been adopted broadly, and we use it as an exemplar of the performance of topology-based methods.

Although topological methods seek to identify shifts in diversification rate, they do not measure rates in units of absolute time. Rather, topological methods assess whether local subtrees are more unbalanced than would be expected by chance if rates of speciation were equal. When analyses are run in SymmeTREE, two shift statistics $\Delta^1$ and $\Delta^2$ are produced for each node within the phylogenetic tree. A $P$ value is also calculated for each node to express the probability of the observed difference in sister-group diversity. Significant ($P < 0.05$) and substantial ($P < 0.1$) values indicating asymmetries in sister clade diversity that reject the equal rates Markov (ERM) null model indicating that the more speciose clade has undergone nonrandom diversification. If two clades are considered, the shift statistics look at the probability that a shift occurred on the internal branch that subtends the more speciose clade. The $\Delta^1$ shift statistic is simply the difference in the log-likelihood ratios between two models, one with a homogeneous rate, and the other with a heterogeneous rate at both the basal and the internal nodes of the clade ($a, (b,c)$). The $\Delta^2$ shift statistic selectively and systematically excludes ingroup subclades in estimating the net rate of diversification in order to pinpoint nodes where the rate may have changed radically (see Chan and Moore 2002, 2005; Moore et al. 2004). The specific settings used in our analyses were 1,000,000 random resolutions of the entire tree and 10,000 for nonbifurcating nodes using the taxon-size sensitive ERM algorithm, and the $\Delta^2$ shift statistic was used to identify diversification rate shifts.

**TOPOLOGIES OF DIVERSIFICATION RATE SHIFTS DO NOT REFLECT TIMING**

SymmeTREE analysis of the Carnivora phylogenetic tree identified three significant and one substantial shift in the rate of diversification, all at levels approximately midway along the temporal axis (Fig. 1). Molecular clock dates for these lineage divergences suggest that these shifts in the rate of diversification occurred in the interval 13–50 Ma (Bininda-Emonds et al. 2007). To demonstrate how the topological position of diversification phenomena flux with extinction, we simulated the impact that extinction would have on the perspective of evolutionary history for extant carnivorans. Taxa that are considered “critically endangered” or “endangered” in the latest available International Union for Conservation of Nature (IUCN) and the United States Endangered Species Act (US ESA) lists (Released 1st

---

**FIGURE 1.** Carnivoran phylogenetic tree taken from Bininda-Emonds et al. (2007). The SymmeTREE analysis was run on both this tree and a pruned version, with all taxa at risk from extinction removed (dashed lines). The results of both analyses are shown with letters above each node corresponding to the presence or absence of a diversification rate shift in the respective analyses. Nodes represented by filled arrows indicate significant shifts in diversification, unfilled arrows indicate substantial shifts. (A) corresponds to shifts present in both phylogenies, (E) corresponds to shifts created by the future extinction of taxa, whereas (L) corresponds to the loss of a shift by extinction. Both represent artifacts of the data. Black triangles indicate the collapse of a larger clade for ease of representation. This figure is available in black and white in print and in colour at Systematic Biology online.
September 2007) were pruned from the phylogenetic tree, representing a pessimistic census date perhaps 100 years hence. This is an entirely arbitrary but objective filter to simply demonstrate the effect of extinction upon the perception of historical diversity and, thus, perceived shifts in the rate of diversification.

The removal of 31 species (12%) leads to substantially different perception of diversification over what is essentially the same episode of evolutionary history. Although some shifts are identified in both topologies, others occur in only one of the topologies (Fig. 1). For example, a shift leading to the families Procyonidae and Mustelidae is lost in the pruned tree (Fig. 1L), and a new shift appears in the lineage leading to the clade encompassing Procyonidae, Mustelidae, and Pinnipedia (Fig. 1E). This greatly affects the timing of the putative events that may have led to either diversification rate shift, with the dating of this phenomenon pushed back from 38 to 50 Ma. Likewise, a new diversification rate shift is identified as a result of taxon pruning at the base of the Civet and Genet clade (Fig. 1E). The split between this clade and its sister taxon occurred at approximately 24 Ma during a period of global warming following a global decrease in temperatures after the Eocene–Oligocene extinction event (Cowie 2007). Critically, however, the causal factor of this putative diversification rate shift is not rooted in a narrative of global warming in the Oligocene, but prospective extinction resulting from poaching and habitat destruction by humans in the Holocene.

Evidently, the positions of putative diversification rate shifts in phylogenetic trees do not reflect their relative or absolute timing. Diversification rate shifts located deep within a phylogenetic tree can be driven by random or deterministic causal factors operating at any time between the present and the node on which the rate shift is identified. Further, diversification rate shifts need not represent “events,” as they are usually portrayed but instead may result from the net effects of multiple rounds of stochastic and/or deterministic speciation and/or extinction. This occurs because, as taxa are removed from consideration when they become extinct, the resolution of perceived changes in diversification rate is diminished. In some instances, this will mask real shifts in diversification rate, whereas in other instances, it will lead to the identification of spurious diversification rate shifts. As a result, the identification of diversification rate shifts may be little more than sampling artefact. The allusion to any deterministically causal macroevolutionary process is entirely inferential. No topology-based analyses of diversification rate have entertained this variable and so, for this reason, their results must be considered questionable.

**Constraining the Timing of Diversification Rate Shifts Requires the Inclusion of Extinct Taxa**

Irrespective of the performance of topology-based methods in identifying diversification rate shifts, the phenomenon that we wish to explain, the unequal distribution of species diversity is real, and it remains our objective to find an approach that allows us to identify the cause of imbalances in clade structure. Because the identification of causal factors underpinning diversification rate shifts relies upon little more than the weak inductive logic of temporal correlation, it is imperative that there is constraint on the timing of both the diversification rate shifts and their putative causes (Moore and Donoghue 2009). The timing of action of the causal factors (deterministic or stochastic) that gave rise to shifts in taxic richness can only be determined by distinguishing the contributory effects of speciation and extinction. This requires that topological analyses of extant diversity also consider the effect of cumulative random extinction as a causal explanation in addition to deterministic factors. However, rejecting such a hypothesis requires the placement of extinct taxa within such topologies to be considered. Ultimately, the goal of topology-based methods cannot be achieved by analyzing phylogenetic trees that encompass extant taxa alone—“complete” phylogenies are required that encompass not just extant members of the clade in question but also their extinct relatives.

Concern over taxon sampling in topological analyses was raised by Mooers (1995) who showed that incomplete trees are more imbalanced than complete trees. This is significant because an increase in imbalance may lead to the erroneous identification of diversification rate shifts (Heath et al. 2008). It has been argued that the imbalance between complete and incomplete phylogenies reflects the nonrandom manner by which systematists sample taxa; if a random sample were taken, the results for the incomplete and complete phylogenies would be the same (Guyer and Slowinski 1991; Kirkpatrick and Slatkin 1993; Mooers 1995; Purvis and Agapow 2002). However, more recent work has shown that the random removal of taxa from simulated complete phylogenies increases the observed imbalance and, thus, incomplete sampling can have a strong effect on the perception of macroevolutionary events inferred from tree topologies (Heath et al. 2008). This provides further support for the inclusion of extinct taxa among their extant relatives in topological analyses of macroevolutionary processes such as diversification.

**Extinct Taxa and Topology-Based Methods**

From their conception, it has been argued that topology-based methods should not be applied to extinct taxa:

When extinct lineages are used, relative differences in extinction can no longer be tested; instead, only differential speciation can be considered. However, if significant differences are found, they could be the results of certain lineages’ having slightly more extinct taxa than others and, therefore,
fewer chances through time for speciation.
(Slowinski and Guyer 1989; p. 910)

It has also been suggested that extinct taxa should not be included within topology-based methods of analyzing diversity because all such methods, including SymmeTREE (Chan and Moore 2005), End-Epi (Rambaut et al. 1997), and apTreeshape (Bortolussi 2006) contrast the degree to which the relative diversity of sister clades deviates from a pattern of random diversification, provided by a Yule Model (Kirkpatrick and Slatkin 1993; Fusco and Cronk 1995; Mooers and Heard 1997; McKenzie and Steel 2000; Chan and Moore 2002; Heard and Mooers 2002; Blum and Francois 2005). A Yule model is a pure birth model (Yule 1924; Aldous 2001; Nee 2006) that follows a random branching process in which each tip (taxon) has an equal probability of branching at any moment in time. Clearly, speciation is not equally probable in extinct and extant lineages and, hence, extinct taxa have either been actively pruned from phylogenetic trees before conducting analyses (e.g., Jones et al. 2005) or else omitted in the first instance (e.g., Bininda-Emonds et al. 2007).

Existing software used in analyzing diversification rate requires users to exclude extinct taxa because of the pure birth algorithms that were coded into them, but this is not a general requirement for software implementations of topology-based methods. Previously, analyses of tree topology have used statistics that have incorporated variable rates of speciation (Heard 1996), speciation followed by mass extinction events (Yule 1924; Heard and Mooers 2002), and variable rates of both speciation and extinction (Raup 1985; Harcourt-Brown et al. 2001; Harcourt-Brown 2002). To realize the aim of identifying the causal bases of diversification rate shifts it is necessary to distinguish the contributory effects of speciation versus extinction. This can only be achieved readily by analyzing holistic phylogenies that include extinct taxa.

**Time Slicing does not Resolve the Problems with Inclusion of Extinct Taxa**

A solution to the problem of including extinct taxa in topology-based analyses of diversity was proposed and applied by Ruta et al. (Ruta et al. 2007; Lloyd et al. 2008), but their approach fails to overcome the fundamental problem with topology-based approaches, viz. an inability to distinguish between the relative effects of speciation and extinction. Their method also creates several new problems concerning the interpretation of results.

Ruta et al. (2007) argued that topology-based methods may be extended to include extinct taxa without violating the Yule model by considering tree topology in cohorts of contemporaries. Once the phylogenetic tree has been mapped onto stratigraphy, it is broken down into separate time slices on which topology-based analyses are conducted sequentially from the oldest to the youngest. When an extinct taxon falls outside the temporal extent of this analytical window, it is removed. Lineages that are inferred to have existed within the time interval because subsequent intervals include descendent clades are shown as ghost lineages (Fig. 2). This approach overcomes the violation of the particular implementation of the Yule model in SymmeTREE (Chan and Moore 2005) because within each time slice every taxon is extant. It also overcomes the limitation of topology-based methods in that it provides a basis for including extinct taxa alongside extant lineages. However, because this method considers tree topology within successive temporal intervals, it remains a topological approach and subject to the same errors associated with the analysis of extant taxa alone, with workers unable to distinguish between the relative effects of both speciation and extinction.

To demonstrate this phenomenon, we utilized the crocodilian phylogenetic tree of Gatesy et al. (2004) mapped onto geological time (Fig. 3) using data from

**FIGURE 2.** The time slice approach as advocated by Ruta et al. (2007). The complete phylogenetic tree is shown on the left, whereas the 3 respective time slices are shown on the right. Each slice shows a different portion of the evolutionary history of the clade. Slice one shows the origination of the clade with the gray line representing ghost ranges of clades that have yet to arise and are not currently known in the fossil record. The black lines indicate taxa with a known fossil record. In slice two, the three taxa known from slice one have become extinct and so no longer included within the analysis, whereas some of the ghost ranges have led to taxa now known in the fossil record others are still present. In the third and final slice, all taxa are represented as present and no ghost ranges are shown but several taxa from slice two have been removed because of extinction. This figure is available in black and white in print and in colour at Systematic Biology online.
Figure 3. The crocodilian phylogenetic tree as presented in Gatesy et al. (2004) mapped to stratigraphy and dated according to Brochu (1997, 1999, 2000, 2003). This figure is available in black and white in print and in colour at Systematic Biology online.
In so doing, we acknowledge that this phylogenetic tree is contentious and that the temporal ranges of these lineages will undoubtedly be revised in light of further sampling of the fossil record and molecular clock analyses. However, none of the competing phylogenies encompass so many extant and extinct taxa and the stratigraphic range data are constant in our comparative analyses of diversification rates. Therefore, with these caveats, we use these resources to make our conceptual point; the results of our analyses should not be read as providing insight into crocodilian evolutionary history.

The method advocated by Brochu (1997, 1999, 2000, 2003). In so doing, we acknowledge that this phylogenetic tree is contentious and that the temporal ranges of these lineages will undoubtedly be revised in light of further sampling of the fossil record and molecular clock analyses. However, none of the competing phylogenies encompass so many extant and extinct taxa and the stratigraphic range data are constant in our comparative analyses of diversification rates. Therefore, with these caveats, we use these resources to make our conceptual point; the results of our analyses should not be read as providing insight into crocodilian evolutionary history.

The method advocated by Ruta et al. (2007) was applied to 6 time slices (Fig. 4). No shifts were observed within the Pliocene, Miocene, and Oligocene. There was one shift in both the Eocene and the Cretaceous, whereas 6 were observed during the Paleocene. How to interpret the results? Should only those shifts present in all time slices be considered? Those shifts that are present in the majority? Or are all shifts identified in all analyses equally valid regardless of their transience? Another option would be to consider the time slices themselves: 6 shifts are observed in the Paleocene slice of which 5 are unique to this slice. If this were the correct approach, then the high number of diversification rate shifts in the Paleocene (Fig. 4) would imply a deterministic factor, possibly associated with the recovery following the Cretaceous–Palaeogene mass extinction. However, when these shifts are considered in light of their stratigraphic occurrence, it is clear to see that they all occurred at least 10 myr before the Cretaceous–Palaeogene mass extinction.

The underlying problem with the method advocated by Ruta et al. (2007) is that, like topological analyses of extant taxa, it relies upon incomplete phylogenies. As per topological analyses of exclusively extant taxa, the solution is to discriminate the relative effects of speciation and extinction on diversity by analyzing speciation alone and entertaining extinction as causal factor in explanation of lineages that are more diverse than would be expected under the Yule Model.

A Solution: Analysis of Speciation Rate

Although topology-based methods perform their task well, they cannot discriminate between stochastic and deterministic differences in diversity or the relative timing of shifts in diversity. However, it remains possible to achieve the principal aims underpinning topology-based analyses of diversity. This can be done by discriminating between changes in diversity that are genuinely the outcome of contemporaneous extinction and speciation from changes in diversity that are merely the outcome of telescoping multiple episodes of stochastic extinction.

These aims can be achieved by analyzing speciation alone and attempting to distinguish between imbalances in diversity that deviate from the spectrum of outcomes expected on the basis of the Yule model. This is the approach advocated for analysis of trees encompassing extinct taxa when topology-based methods of diversification were first proposed (Slowinski and Guyer 1989: p. 910). It is a logical expectation that, unlike analyses of diversification, shifts in the speciation rate are contemporaneous with the relative age of the branch on which the rate shift is identified. It is necessary to consider extinction as an explanatory factor, among other possible intrinsic and extrinsic factors. As we have shown, however, this was always an implicit requirement of effective interpretations of diversification rate shifts. This approach also has the benefit of allowing—arguably requiring—the inclusion of extinct taxa, facilitating discrimination between apparent and actual historical diversities (cf. Harvey et al. 1994).

Rather than applying a time slicing approach, we used a nested-growth method wherein topology-based analytical methods are applied to increasingly inclusive nested sets of taxa, from the oldest to the youngest. In this instance, the phylogenetic tree is grown through time so that all speciation events are recorded, and the SymmeTREE analysis is repeated after the addition of each successive time interval (i.e., the first analysis is on time period T1, the second is on T1 + T2 and the third on T1 + T2 + T3). This makes it possible to discriminate the pattern of speciation alone to which the temporally varying pattern of diversification may be compared. In these analyses, taxa move into, but not out of, the analytical reading frame of an analysis of tree topology (Fig. 5). Thus, tree topology changes only as a result of speciation not extinction. Anomalous shifts in diversification rates that arise as a result of the inclusion of extinct taxa are identifiable because they appear at the base of extant clades long after their fossil relatives have gone extinct (cf. Slowinski and Guyer 1989).

In the crocodilian example, Shift 4 can be identified as the product of extinction rather than speciation (Fig. 5). Although the shift is inferred to have occurred in the Late Cretaceous, it is not apparent until the reading frame is extended to the Miocene. In this case, the diversification rate shift was caused by two factors; the later speciation within the Crocodyliinae during the Miocene and the extinction of numerous (now stem) taxa during the Eocene. These two events occurred approximately 50 and 20 myr after the shift is inferred to have occurred showing that the causative factors were not associated with the origination of the more speciose clade.

This is in contrast to Shift 1 that is present in every reading frame from the Cretaceous to the present. This shift is identified while both clades are contemporaneous, with the less speciose clade surviving more than 35 myr after the shift is inferred to have occurred showing that this lineage had sufficient longevity for further speciation. These results indicate that the shift at the base of crown Crocodylia reflects a genuine event during the end Cretaceous rather than an artifact caused by later extinction or speciation.
Figure 4. The crocodylian phylogenetic tree divided into the 6 time slices shown in Fig. 3 following the method of Ruta et al. (2007). Black lines indicate taxa present within the time slice, whereas gray lines show taxa known from their ghost ranges. Nodes represented by filled arrows indicate significant shifts in diversification, unfilled arrows indicate substantial shifts. This figure is available in black and white in print and in colour at Systematic Biology online.
FIGURE 5. The Gatesy et al. (2004) crocodilian phylogenetic tree analyzed using the nested-growth method. The phylogenetic tree is grown through the 6 time periods, nodes represented by filled arrows indicate significant shifts in diversification, and unfilled arrows indicate substantial shifts. As the tree is grown through time, the relative effects of both speciation and extinction becomes clear and it is possible to distinguish between the two. When we consider the diversification rate shift labeled 1, we see that it is present in every time period. It appeared in the Cretaceous, while both lineages were extant, the extinction of Pristichampsus vorax during the Eocene would not have had any affect on this diversification rate shift as it was already present when both lineages were contemporaneous. This is in contrast to the shift labeled 4, this shift appeared in the Miocene, although it is inferred to have occurred in the Cretaceous and is an artifact caused by both the speciation of the Crocodylinae in the Miocene and the extinction of many stem taxa during the Eocene. This figure is available in black and white in print and in colour at Systematic Biology online.
**Caveat Ex Tempore**

Our approach to using topology-based methods to identify clades that are more speciose than would be anticipated under a pure birth Yule model has the benefit of effectively identifying the relative timing of shifts in speciation. If only extant taxa are considered, it is not possible to discriminate whether extinction (stochastic or deterministic) is a causal factor in explaining differences in diversity, nor indeed, the relative timing of extinction or any other causal factor. However, with the inclusion of extinct taxa and subsequent conversion to timetrees (whether by using the temporal data with which fossil taxa are intrinsically imbued or else relying on molecular clock estimates), it is possible to discriminate between stochastic and nondeterministic extinction as a causal factor underpinning diversity, as well as the relative, even absolute, timing of nonstochastic shifts in diversity.

Despite our enthusiasm for the integration of extinct and extant taxa in topology-based analyses of diversity, the view that such a holistic approach in some way provides for “complete” phylogenies must be tempered with the realization that it will never be possible to obtain phylogenies that encompass all members of a clade. This is because only a small proportion of species preserve a fossil record that survives the rigors of tectonism and is recovered and identified as relevant (Paul 2009). Further, the temporal extent of those extinct species that are known to science will always fall short of their true longevity because of secular biases in the preservation of environments in the rock record, peculiarities in the way that fossils are sampled, and the impact of provincialism in their biogeographic distribution, among other factors (e.g., Donoghue and Benton 2007). These limitations can be constrained somewhat using methods such as confidence intervals (Marshall 1997), and graphic correlation (Mann and Lane 1995) on stratigraphic range data, as well as molecular clock analyses (Smith et al. 2006). Indeed, these limitations may be overcome altogether if preservation is controlled for alongside estimates of speciation and extinction (Connolly and Miller 2001; Foote 2001; Alroy et al. 2008), but no methods yet exist that integrate all three variables in a tree-based approach. Regardless, the temporal context of those species that are preserved, collected, and studied, should provide sufficient constraint to discriminate between stochastic extinction and deterministic causal factors in explanation of nonstochastic patterns of clade diversity. The temporal correlation of putative causal factors may then be tested probabilistically (Moore and Donoghue 2009).

At the conception of topology-based methods, Slowinski and Guyer (1989) shied away from analyses that encompass extinct taxa, perhaps implicitly because of the problems we have just outlined but explicitly because of the problems perceived with identifying the phylogenetic position of fossil taxa. Empirical evidence shows that extinct taxa are just as influential as their extant relatives in affecting the precision of phylogenetic analyses (Gauthier et al. 1988; Donoghue et al. 1989; Cobb et al. 2007) and simulations show in attempting to achieve an integrated phylogenetic tree of extinct and extant taxa, molecular phylogenetic data greatly improves the accuracy with which the phylogenetic position of extinct taxa is identified (Wiens et al. 2009).

Finally, we return to the nub of the issue, that is, why are certain clades considerably more speciose than their immediate relatives? Answering this question has been one of the principal goals of biology and, if it is to be achieved, it will be necessary to overcome both the limitations of the fossil record and the negative prejudices that exist toward it. This echoes the call for the integration of fossil taxa into timetrees through time (Harvey et al. 1994; Pybus and Harvey 2000) based analyses of historical diversity (Crisp and Cook 2009; Rabosky 2009, 2010a,b; Liow et al. 2010; Quental and Marshall 2010). Ultimately, the fossil record may not be complete but, then, neither is a molecular phylogeny that, by necessity, encompasses only extant taxa. The only means of approximating a complete phylogeny and, therefore, a complete perspective on historical diversity is by restoring extinct taxa to their rightful place in phylogenies alongside their living relatives (Wiens et al. 2010).

**Conclusions**

Topology-based methods provide a means of describing tree shape and determining whether or not the shape of a particular phylogenetic tree deviates from the kind of shape that would be produced by a homogenous process of random branching. Inevitably, where differences in the diversity of sister lineages deviate significantly from random chance, causal drivers are sought. However, phylogenetic tree topology is the sum of multiple rounds of speciation and extinction, resulting in tree shapes that are significantly nonrandom. Furthermore, the relative timing of causal drivers underpinning tree asymmetries need not relate to the position of the root of the asymmetry within the topology of the tree. To determine the material basis of tree asymmetries, it is necessary to discriminate between those that occur as a result of contemporaneous changes in diversification rates versus tree asymmetries that occur through extinction-driven loss of phylogenetic history. This cannot be achieved using topological data from extant species alone. Fossil taxa incorporated into phylogenies of their extant relatives provide a means of temporally constraining and discriminating between the relative contributions of speciation and extinction in effecting diversification.

**Funding**

Funding for this work was provided by the Natural Environment Research Council.
ACKNOWLEDGMENTS

We would like to thank M. Benton, M. Ruta, and N. Stroeymeyt for comments on an earlier draft. We also
thank the reviewers, C. Brochu, J. Muller, D. Polly, and P. Wagner, our associate editor, N. MacLeod, and jour-
nal’s editors, R. DeBry and J. Sullivan, for thoughtful
discussions that greatly improved the paper.

REFERENCES

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky
plus high turnover explain species diversity in jawed vertebrates.
Alroy J., Aberhan M., Bottjer D.J., Foote M., Fursich F.T., Harries P.J.,
Hendy A.J.W., Holland S.M., Ivancy L.C., Kiessling W., Kosnik M.A.,
Patzkowski M.E., Peters S.E., Villier L., Wagner P.J., Bonuo N.,
Borkow P.S., Brennies B., Clapham M.E., Fall L.M., Ferguson C.A.,
Hanson V.L., Krug A.Z., Layou K.M., Leckey E.H., Nurnberg S.,
Powers C.M., Sessa J.A., Simpson C., Tomasovych A., Visaggi C.C.
2008. Phanerozoic trends in the global diversity of marine inverte-
Bininda-Emonds O.R.P., Cardillo M., Jones K.E., MacPhee R.D.E., Beck
A. 2007. The delayed rise of present-day mammals. Nature. 446:
507–512.
large trees by combining phylogenetic information: a complete
phylogeny of the extant Carnivora (Mammalia). Biol. Rev. 74:
143–175.
Blum M.G.B., Francois O. 2005. On statistical tests of phylogenetic tree
Bortolussi N., Durand E., Blum M., Francois O. 2006. apTreeshape:
statistical analysis of phylogenetic tree shape. Bioinformatics. 22:
363–364.
Brochu C.A. 1997. Morphology, fossils, divergence timing, and the
Brochu C.A. 1999. Taxon sampling and reverse successive weighting.
Brochu C.A. 2000. Phylogenetic relationships and divergence timing of
Crocodylus based on morphology and the fossil record. Copeia.
3:657–673.
Brochu C.A. 2003. Phylogenetic approaches toward crocodylian his-
Chan K.M.A., Moore B.R. 2005. SYMMETREE: whole-tree analysis of
turnover rates from fossil databases: capture-mark-recapture meth-
living taxa in parsimony analyses of morphology. Syst. Biol. 56:
753–766.
South American tuco-tucos (genus Ctenomys, Rodentia: Octo-
Cambridge: Cambridge University Press.
Crisp M.D., Cook L.G. 2009. Explosive radiation or cryptic mass ex-
tinction? Interpreting signatures in molecular phylogenies. Evolu-
tion. 63:2257–2267.
Davison A., Chiba S. 2006. Labile ecotypes accompany rapid cladoge-
nesis in an adaptive radiation of Mandarin (Bradybaenidae) land
Donoghue P.C.J., Benton M.J. 2007. Rocks and clocks: calibrating the
tree of life using fossils and molecules. Trends Ecol. Evol. 22:
424–431.
Donoghue, M.J., Doyle J., Gauthier J., Kluge A., Rowe T. 1989. The
importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol.
Syst. 20:431–460.
Egan A.N., Crandall K.A. 2008. Divergence and diversification in
North American Ponerine (Formicidae) due to climate change. BMC
Foote M. 2001. Inferring temporal patterns of preservation, origina-
tion, and extinction from taxonomic survivorship analysis. Paleobi-
ology. 27:602–630.
Foote M. 2006. Labs and cladogenesis: the importance of fossils in phy-
The importance of fossil clades to the unification of abiotic and
13307–13312.
Guyer C., Slowinski J.B. 1991. Comparisons of observed phylogenetic
topologies with null expectations among three monophyletic lin-
Harcourt-Brown K.G. 2002. Tree balance, time slices, and evolution-
ary turnover in Cretaceous planktonic foraminifera. Syst. Biol. 51:
908–916.
cence of paleontological trees. Paleobiology. 27:188–204.
Hardman M., Hardman L.M. 2008. The relative importance of body
dimension and paleoclimatic change as explanatory variables influencing
lineage diversification rate: An evolutionary analysis of bulbous
Harvey P.H., May R.M., Nee S. 1994. Phylogenies without fossils. Evo-
lution. 48:523–529.
Heard S.B. 1996. Patterns in phylogenetic tree balance with variable
analysis of ecomorphological divergence, community structure,
Kirkpatrick M., Slatkin M. 1993. Searching for evolutionary patterns in
analysis of ecomorphological divergence, community structure,
and diversification rates in dusky salamanders (Plethodontidae: Plethodon). Syst.
Biol. 54:879–897.
Kozak K.H., Zwickl D.J., Kim J., Hills D.M. 2008. Taxon sampling af-
fers influences of macroevolutionary processes from phylogenetic
family pygopodidae with implications for the diversification of
Jones K.E., Bininda-Emonds O.R.P., Gittleman J.L. 2005. Bats,
clocks, and rocks: diversification patterns in chiroptera. Evolution.
Kazanciglu E., Near T.J., Hanel R., Wainwright P.C. 2009. Influence of
sexual selection and feeding functional morphology on diver-
276:3439–3446.
Kirkpatrick M., Slatkin M. 1993. Searching for evolutionary patterns in
analysis of ecomorphological divergence, community structure,
Kozak K.H., Wessick D.W., Larson A. 2006. Rapid lineage accumu-
lization in a non-adaptive radiation: phylogenetic analysis of
diversification rates in eastern North American woodland salam-
273:539–546.
Model identifiability is a key component of any proof of statistical consistency. Identifiability means that it is possible to infer all of the model’s parameters given an infinite amount of data from the model. For phylogenetic inference under the F81+Γ model, for example, the parameters are the unrooted phylogenetic tree with branch lengths, the particular F81 rate matrix (Felsenstein 1981), and a shape parameter for the gamma (Γ) distribution describing the rate heterogeneity. The F81 rate matrix is particularly simple with the rate of mutating to state x depending only on the long-run frequency of state x. This model is not identifiable using only pairwise species comparisons, that is, the joint pairwise DNA state distributions (Steel 2009).

For any F81 matrix, any two distinct gamma distributions, and any set of four or more species, Steel (2009) showed that there are distinct topologies for each gamma distribution that give the same joint pairwise DNA sequence distributions for those species. Any statistical estimator of these model parameters using only pairwise comparisons, such as distance-based methods, will be trying to estimate two separate points in parameter space, violating the definition of consistency.

For the purposes of this work, we define generic identifiability to mean that the set of parameters for which a model is not identifiable has a smaller dimension than the whole parameter space. Wu and Susko (2010) proved generic identifiability for the general time reversible (GTR) + Γ model from pairwise comparisons. The model parameters are the phylogeny with branch lengths, the 4 × 4 instantaneous rate matrix describing GTR DNA evolution along the phylogeny, and a shape parameter for the gamma distribution. Specifically, they proved that for all but the F81 family of matrices and for all phylogenies with at least two distinct interspecies distances, the rate matrix, the shape of the gamma distribution, and the phylogeny and its branch lengths are identifiable from pairwise comparisons. Because a GTR Markov matrix is parameterized using three nonzero eigenvalues but the F81 subfamily of matrices is parameterized using only one nonzero eigenvalue, the F81 model is a lower dimensional subset of the whole GTR class. Similarly, given more than two taxa, phylogenies with only one value for all of their interspecies distances make up a lower dimensional subset of the whole of tree space.

For phylogenetic inference under GTR + Γ + I, we add an additional parameter, which is the proportion of invariable (I) sites (Gu et al. 1995). Rogers (2001) argued that this popular model was generically identifiable from pairwise comparisons and that all other aspects of Wald’s proof that maximum likelihood estimators are consistent held for phylogenetic inference. His argument of identifiability, however, contained a flaw (Allman et al. 2008).

Allman et al. (2008) gave a correct proof, using three-taxon comparisons, of identifiability for the model without invariable sites for generic GTR Markov matrices. The exceptional cases when using joint three-taxon distributions only involve conditions on the GTR rate matrices, thus eliminating the exception against phylogenies with only one interspecies distance. For 4 × 4 DNA matrices, the authors go on to handle the exceptional cases, so the specific results of Allman et al. (2008) show that the model with an F81 rate matrix and gamma-distributed rate heterogeneity is always identifiable using joint three-taxon distributions. Thus, although pairwise distributions can be the same for two distinct models with F81 rate matrices, three-taxon distributions cannot be. However, the sufficiency of pairwise comparisons is of interest because distance-based methods traditionally use only pairwise comparisons and proofs of their consistency rely on identifiability based only on pairwise information.

In this paper, we complete Rogers’ proof of generic identifiability for the model with invariable sites using pairwise comparisons for all but members of the F81 family of rate matrices on any phylogeny with more than two distinct interspecies distances. Our proof also works on phylogenies with only two distinct pairwise interspecies distances if the rate matrix has three distinct nonzero eigenvalues, a condition which rules out not only the F81 family of matrices but also families such as K2P (Kimura 1980) with only two distinct nonzero eigenvalues. However, we show in the next section that deciding whether we are in the generic case or in an exceptional case as well as the nature of the exceptional case (whether it comes from not having enough distinct pairwise interspecies distances or from not having enough distinct nonzero eigenvalues) can be determined solely from the pairwise comparisons. All of these exceptional cases may still be identifiable.
using information from the joint distributions of three or more taxa in which case maximum likelihood techniques would be consistent as would appropriately generalized distance-based methods (Ranwez and Gascuel 2002; Contois and Levy 2005).

Software packages such as PAUP (Swofford 1991) and PhyML (Guindon and Gascuel 2003) use a discrete gamma distribution introduced by Yang (1995) for computational efficiency and do not actually use the theoretical continuous version discussed in this paper. PHYLIP (Felsenstein 2005) uses either a discrete gamma distribution or a Bayesian method for incorporating the gamma distribution for rate heterogeneity efficiently into its software code. Identifiability of models using the discrete gamma distribution has not been analyzed except in the case where the number of discrete classes is fewer than the number of states in the transition matrix, so fewer than four for DNA (Allman and Rhodes 2006).

IDENTIFIABILITY OF THE GRT +Γ+ I MODEL

We state the formal result and briefly give Rogers’ proof up to the missing component. We sketch the idea behind the proof of the missing component in this section and give the rigorous details in the Appendix.

Theorem 1. Let T be an unrooted phylogenetic tree with three or more distinct interspecies distances. Let Q be any GTR instantaneous rate matrix with two or more distinct nonzero eigenvalues. Let R be a mixture of a gamma distribution with mean 1 and invariable sites. Then, sequence data produced under this model identifies this model using pairwise comparisons. That is, with infinite sequence data at the species tips, we can recover all of the model parameters in T, Q, and R used to generate the species data using only pairwise species comparisons. Additionally, if the tree has only two distinct interspecies distances but the rate matrix has three distinct nonzero eigenvalues, then the model is also identifiable.

We work only with four-state DNA rate matrices although everything herein would generalize to more states, such as for models of protein evolution. Rogers’ proof is nicely constructive in the sense that it not only says that the model can be identified but also explains how to identify the components.

What Was Previously Proved

Most of the components are easily identifiable.

1. The stationary nucleotide frequencies, \( \pi = (\pi_A, \pi_C, \pi_G, \pi_T) \), can be identified simply by determining the relative proportion of A’s, C’s, G’s, and T’s in the species sequences. In general, the stationary nucleotide frequency is not enough to determine the rate matrix Q. However, if additionally we know that the rate matrix has only one nonzero eigenvalue, then we know that the rate matrix is in the F81 family and thus has the form

\[
Q = \begin{pmatrix}
A & C & G & T \\
A & -\pi_C & \pi_G & \pi_T \\
C & \pi_A & -\pi_G & \pi_T \\
G & \pi_A & \pi_C & -\pi_T \\
T & \pi_A & \pi_C & \pi_G \\
\end{pmatrix}
\]

where the dashes are understood to be what is required to make the row sum to zero.

2. The rate matrix Q is determined by its eigenvalues and its eigenvectors. The observed DNA state transition matrix for a pair of species separated by distance \( \tau \) is \( P(\tau) = \exp(\tau \mathbf{Q}) \), where \( \tau \) is the random rate from the mixture distribution of the gamma distribution plus invariable sites over which the expectation is taken. The observed matrix \( P(\tau) \) has the same eigenvectors as the rate matrix Q. The eigenvalues of the two matrices are also related. If the eigenvalues for Q are \( (-\lambda_1, -\lambda_2, -\lambda_3, 0) \) (where \( 0 < \lambda_2 \leq \lambda_3 \leq \lambda_4 \)), then the eigenvalues for \( P(\tau) \) will be \( \{E[e^{-\tau \lambda_1}], E[e^{-\tau \lambda_2}], E[e^{-\tau \lambda_3}], 1\} \). These latter values are observable from the transition matrix, \( \mathbf{P}(\tau) \). Let \( \mu(x) = E[e^{-\tau x}] = \pi + (1-\pi)(1 + x/\alpha)^{-\alpha} \) be the moment generating function evaluated at \(-x\) for the rate distribution which is a mixture of invariable sites (with proportion \( \tau \)), not to be confused with the stationary distribution frequencies \( \pi_A, \pi_C, \pi_G, \pi_T \), which have subscripts plus a mean 1 gamma distribution (with proportion \( 1-\pi \)) and shape parameter \( \alpha \). If \( \alpha \) and \( \pi \) were known, then the functional form of \( \mu \) is known and one can solve for the nonzero eigenvalues. That is, for the \( i \)th pair of species tips with distance \( \tau_i \) between them and for eigenvalue \( \lambda_i \), \( m_{ij} = E[e^{-\tau_i \lambda_i}] = \mu(\tau_i \lambda_i) \) is observed and \( \tau_i \lambda_i = \mu^{-1}(m_{ij}) \).

3. Again, if the functional form of \( \mu \) were known, then all the interspecies distances can be determined by inverting. Knowing all pairwise distances exactly determines the phylogeny (Buneman 1971).

There is a small redundancy that all of this is known up to an arbitrary scale factor because the distances and rates always appear multiplied together. We can set the distance between an arbitrary pair of species to be 1 to resolve this difficulty as Rogers suggested. Other branch lengths would then be relative to the distance between this pair. Alternatively, we can set a scale for the DNA evolution rate matrix, Q. Phylogenetic software typically sets the scale in the model with invariant sites so that \( \pi_AQ_{AA} + \pi_CQ_{CC} + \pi_GQ_{GG} + \pi_TQ_{TT} = \frac{1}{\tau} \) with the interpretation that the branch length measures the expected number of substitutions on the branch per site.

Note that the observed moment,

\[
m_{ij} = E[e^{-\tau_i \lambda_i}] = \mu(\tau_i \lambda_i) = \pi + (1-\pi)(1 + \tau_i \lambda_i/\alpha)^{-\alpha}
\]

is a strictly decreasing function of both the distances \( \tau_i \) and the nonzero eigenvalues \( \lambda_i \). By fixing a pair of
species with pairwise distance $\tau_i$, we can determine how many distinct nonzero eigenvalues there are. If there is only one distinct nonzero eigenvalue, the matrix comes from the F81 family and one can determine the exact F81 matrix, as noted above. The observed moments also, for a fixed nonzero eigenvalue, determine how many distinct interspecies distances are in the phylogenetic tree. Thus, we can tell if we are in the exceptional case where there is only one or two distinct interspecies distances.

**What Was Incorrectly Proved**

The difficulty in the argument is determining the functional form for $\mu$, that is, in determining $\pi$ and $\alpha$. At the point in Rogers’ argument where he is trying to show that the form of the moment generating function is determined by the pairwise joint sequence distributions, the tree and the rate matrix describing DNA evolution generally have not yet been determined. Thus, the question is whether there can be two different moment generating functions are identical was flawed (Rogers 2001, Allman et al. 2008).

We note the symmetry here played by the distances and the nonzero eigenvalues. If there are just two distinct nonzero eigenvalues ($\lambda_2 \neq \lambda_3$) but many values for the distances separating pairs of tips, then we have $\nu^{-1}(\mu(\tau_{i2}))/\nu^{-1}(\mu(\tau_{i3})) = l_{i2}/l_{i3}$ for at least three distinct interspecies distances $\tau_i$ on a generic phylogeny.

Our main result, Proposition 1, proved in the Appendix, states that $\nu^{-1}(\mu(\tau_{i2}))/\nu^{-1}(\mu(\tau_{i3}))$ equals a constant for at most two distinct values of $\tau$ unless the two moment generating functions $\nu$ and $\mu$ are identical. A mathematically identical argument works for two distinct interspecies distances with a rate matrix with three distinct nonzero eigenvalues by showing that Rogers’ original ratio, $\nu^{-1}(\mu(\tau_{i1}))/\nu^{-1}(\mu(\tau_{i2}))$, can equal a given constant at most twice in the eigenvalue variable $\lambda$. We omit the statement and proof of this symmetrical result from the Appendix.

**The Ideas Behind Our Proof**

We make use of Rogers’ original and useful trick of considering $\nu^{-1}(\mu(z))$, which changes the nonlinear multivariable problem of determining how many solutions are possible to the equations

$$\mu(\tau_{i1}) = \pi + (1 - \pi)(1 + \tau_{i1}/\alpha)^{-\alpha} = m_{ij}$$

in terms of the scale of the gamma distribution $\alpha$ and the proportion of invariant sites $\pi$ into a single variable calculus problem of determining how many times the function $\nu^{-1}(\mu(\tau_{i2}))/\nu^{-1}(\mu(\tau_{i3}))$ of the single variable $\tau$ crosses a horizontal line. A continuous, smooth function crosses a horizontal line at most once if it is monotone (increasing or decreasing); it crosses at most twice if its derivative is zero at most once; and it crosses at most three times if its derivative is zero at most twice, which happens if its second derivative is zero at most once. Those observations and some algebraic manipulations form the basis of the proof of this missing component in Rogers’ argument. The rigorous proof is provided in the Appendix.

This is the one place where Rogers’ proof is not constructive in that the identification of $\alpha$ and $\pi$ is not explicit at this point. In practice, for $m_{ij} = \mu(\tau_{i1}) = \pi + (1 - \pi)(1 + \tau_{i1}/\alpha)^{-\alpha}$, we would use three distinct interspecies distances $\tau_1, \tau_2$, and $\tau_3$, two distinct nonzero eigenvalues $\lambda_2$ and $\lambda_3$, and solve the equations

$$\mu^{-1}(m_{12})/\mu^{-1}(m_{13}) = \mu^{-1}(m_{22})/\mu^{-1}(m_{23})$$

for $\alpha$ and $\pi$ using numerical software.

These equations can be solved exactly in this context because we have an infinite amount of data and thus have the observed moments $m_{ij}$ exactly. With estimates of these moments from a finite amount of pairwise data, a least squares approach derived from these theoretical equations for estimating the common quotient, $\alpha$, and $\pi$ could be developed. This approach is unnecessary for maximum likelihood methods that have natural ways for estimating these parameters but might be useful for distance-based methods that do not currently provide estimates for $\alpha$ and $\pi$.

**ACKNOWLEDGMENTS**

We thank John Rhodes, Elizabeth Allman, the three attentive reviewers—Cecile Ané, Mike Steel, and one who remained anonymous—and Associate Editor Olivier Gascuel for numerous corrections and helpful comments, which substantially improved the manuscript.

**REFERENCES**


APPENDIX

Proof of Rogers’ Claim

PROPOSITION 1 Assume

\[
\mu(x) = \pi + (1 - \pi)(1 + x/\alpha)^{-\alpha}
\]

\[
\nu(y) = p + (1 - p)(1 + y/a)^{-a}
\]

are the parameterizations of the two moment generating functions (evaluated at \(-x\)) for the rate distribution which is a mixture of a gamma distribution with mean 1 and a point mass at zero for invariable sites. Also assume that \(\pi \geq p\). (If this is not true, reverse the roles of \(\nu\) and \(\mu\).) Furthermore, assume that the rate matrix \(Q\) in the model for DNA evolution associated with \(\mu\) has at least two distinct nonzero eigenvalues \((-\lambda_2\) and \(-\lambda_3\)) so that \(\lambda_2 < \lambda_3\). Let the corresponding eigenvalues associated with \(\nu\) be \(-l_2\) and \(-l_3\), which need not be distinct. Then the equation

\[
\frac{\nu^{-1}(\mu(\lambda_2x))}{\nu^{-1}(\mu(\lambda_3x))} = l_2/l_3 = A
\]

has at most two solutions in the variable \(x > 0\) whenever \(\mu \neq \nu\).

Proof. Note that

\[
\nu^{-1}(\mu(x)) = a \left\{ \frac{\pi - p + (1 - \pi)(1 + x/\alpha)^{-\alpha}}{1 - p} \right\}^{-1/a} - 1
\]

and that showing (1) has at most two solutions for \(x > 0\) is equivalent to showing that the function

\[
f(x) = [(\pi - p) + (1 - \pi)(1 + x/\alpha)^{-\alpha}]^{-1/a}
\]

satisfies

\[
f(x) - Af(Bx) = (1 - A)[1 - p]^{-1/a}
\]

for at most two \(x > 0\). In this equation, \(A = l_2/l_3 \leq 1\) because the eigenvalues are ordered and \(B = \lambda_3/\lambda_2 > 1\). The inequality for \(B\) is strict by the hypothesis that the rate matrix of the model associated with \(\mu\) has at least two distinct nonzero eigenvalues. We further note that the equality holds at \(x = 0\), and we will bound the number of times that the derivative of \(f(x) - Af(Bx)\) is zero. This is equivalent to bounding the number of times \(f'(x)/f''(Bx) = AB\).

We will use the following equivalent formulas for the derivative:

\[
f'(x) = \frac{1 - \pi}{a} \left\{ [(\pi - p)(1 + x/\alpha)^{-\alpha}]^{-1/a} + (1 - \pi)(1 + x/\alpha)^{-\alpha/a} \right\}
\]

\[
f''(Bx) = \frac{1}{a} \frac{1 + Bx/\alpha}{1 + x/\alpha} \left\{ [(\pi - p)(1 + Bx/\alpha)^{\alpha} + (1 - \pi)]^{1/\alpha} \frac{(\pi - p)(1 + x/\alpha)^{-\alpha} + (1 - \pi)}{(\pi - p)(1 + x/\alpha)^{\alpha} + (1 - \pi)} \right\}.
\]

Because both parts on the right hand side are increasing, \(f'(x)/f''(Bx)\) crosses \(AB\) at most once. Thus the equation \(f(x) - Af(Bx) = (1 - A)[1 - p]^{-1/a}\) has at most one additional solution for \(x > 0\). That the second term is increasing in \(x\) is not completely obvious. With some algebraic manipulation its derivative can be written as:

\[
\frac{d}{dx} \left\{ (\pi - p)(1 + \frac{Bx}{\alpha})^{\alpha} + (1 - \pi) \right\} = \frac{(\pi - p)(B - 1)(1 + \frac{Bx}{\alpha})^{\alpha - 1}(1 + \frac{x}{\alpha})^{\alpha - 1} + \frac{(\pi - p)(B - 1)}{(1 + \frac{x}{\alpha})^{\alpha - 1}}}{((\pi - p)(1 + \frac{x}{\alpha})^{\alpha} + (1 - \pi))^2}
\]

In the case \(\alpha \geq 1\) the term \(B(1 + \frac{Bx}{\alpha})^{\alpha - 1}\) is increasing with a minimum of \(B\) at \(x = 0\). In the case \(\alpha < 1\), it is decreasing with a minimum of \(B^\alpha\) as \(x \to \infty\). In either
case, it is greater than 1. So the derivative is positive and the function is increasing as claimed.

If \( \alpha > a \), we will use the third expression for the derivative. For simplicity of notation and to facilitate our calculations, we will change the parameterization. Bounding the number of times \( f'(x)/f''(Bx) = AB \) is equivalent to bounding the number of times the function

\[
g(x) = \frac{g(1 + Bx)^r + (1 + Bx) - s}{q(1 + x)^r + (1 + x) - s} = c
\]

where \( c \) is a constant, \( g = (\pi - p)/(1 - \pi) < 1, B > 1, 0 < r = a(\alpha + 1)/(a + 1), 0 < s = (\alpha - a)/(a + 1), \) and \( s + r = \alpha \). To do this, we will show that \( g'(x) = 0 \) at most once. Note that \( g'(x) = 0 \) if and only if

\[
0 = B[q(1 + Bx)^r - s(1 + Bx)^{-s-1}]|g(1 + x)^r - (1 + x)^{-s}]
- [q(1 + Bx)^r + (1 + Bx)^{-s}][q(1 + x)^r - (1 + x)^{-s}]
- q^2r(B - 1)(1 + Bx)^{r+1}
- + q^2r(B - 1)(1 + Bx)^{r+1}(B - 1)
\]

Placing the terms with a multiple of \( s \) on the other side and multiplying through by \((1 + x)^{-r+1}(1 + Bx)^{r+1}\), we see that the last equation is zero if and only if

\[
\frac{s(B - 1)}{(1 + x)^{r+1}} + q^2r(1 + x)
\left( B - \left( \frac{1 + Bx}{1 + x} \right)^{r+1} \right) = 0
\]

\[
q^2r(B - 1)(1 + Bx)^{r+1}
+ q^2r(1 + Bx)
\left( B - \left( \frac{1 + Bx}{1 + x} \right)^{r+1} - 1 \right) = 0
\]

We claim that the function \( (4) \) on the left hand side of this equation is decreasing and the function \( (5) \) on the right hand side of this equation is increasing, so that the intersection occurs at most once. Thus \( g'(x) = 0 \) at most once and \( f'(x)/f''(Bx) = AB \) at most twice for \( x > 0 \). Thus, \( f(x) - Af(Bx) = (1 - A)(1 - p)^{-1/a} \) for at most two \( x > 0 \) because that equation holds for \( x = 0 \).

To prove the claims, let

\[
h_1(x) = (1 + x)
\left( B - \left( \frac{1 + Bx}{1 + x} \right)^{r+1} \right)
\]

so that

\[
h'_1(x)
= B \left( \frac{1 + Bx}{1 + x} \right)^{r+1}
\left( \frac{1}{1 + x} \right)(1 + Bx + (r + s + 1)(B - 1)).
\]

Note that \( h'_1(0) = -(B - 1)(r + s) < 0 \) and \( h'_1(x) \) is decreasing because both \( \frac{1 + Bx}{1 + x} \) and \( \frac{1 + (r+s+1)(B-1)+Bx}{1 + x} \) are increasing functions. Thus, \( h_1(x) \) and, consequently, function \( (4) \) are decreasing functions.

If \( \alpha = r + s \geq 1 \), then the right hand side function \( (5) \) is clearly increasing. If \( \alpha = r + s < 1 \), then let

\[
h_2(x) = (1 + Bx)
\left( B - \left( \frac{1 + Bx}{1 + x} \right)^{r+s-1} - 1 \right)
\]

so that

\[
h'_2(x) = B \left( -1 + \left( \frac{1 + x}{1 + Bx} \right) \left[ B - \left( \frac{1 - (r + s)(B - 1)}{1 + x} \right) \right] \right)
\]

Because \( \frac{1 + x}{1 + Bx} \) is between 0 and 1 and because \( 1 - (r + s) > 0 \), it is also between 0 and 1, \( \left( \frac{1 + x}{1 + Bx} \right) \geq \frac{1 + x}{1 + Bx} \). Thus,

\[
h'_2(x) \geq B \left( -1 + \left( \frac{1 + x}{1 + Bx} \right) \left[ B - \left( 1 - (r + s)(B - 1) \right) \right] \right)
\]

\[
= B \left( -1 + \frac{1 + Bx + (r + s)(B - 1)}{1 + Bx} \right) \geq 0.
\]

Thus, \( h_2(x) \) and, consequently, \( (5) \) are increasing functions.

The F81 Exceptional Case

We will show that the F81 family of rate matrices are the only ones with only one nonzero eigenvalue because this was not pointed out in Wu and Susko (2010). We can set the one nonzero eigenvalue to any negative number and the argument will be the same, so assume that the nonzero eigenvalue is \( -1 \). Let \( V \) be the matrix of eigenvectors with the last column being the eigenvector associated with the eigenvalue zero so that it is the transpose of \((1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2)\). Then,

\[
Q = V \left( \begin{array}{cccc}
-1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} \right)
\]

\[
V^{-1} = -I + \left( \begin{array}{cccc}
0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 1/2 \\
\end{array} \right) V^{-1}.
\]

Thus, \( Q \) has the identical entries in a given column except for the diagonal entry, making the rate matrix a member of the F81 family. The Jukes—Cantor matrix is a special member of the F81 family and can be distinguished by considering the eigenvectors because, for Jukes—Cantor, the eigenspace of the nonzero eigenvalue is orthogonal to \((1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2)\). Other families can overlap with the F81 family and, for the specific cases in which they do so, they fall into the exceptional (non-generic) case.

Clearly, it is not possible to determine the branch lengths, shape parameter, and proportion of invariable sites for an F81 matrix using one three-taxon tree as done in Allman et al. (2008) for the model without invariable sites because there are at most four pieces of information (the changes between pairs of species and for the
entire three-taxon tree) and five parameters to solve for. Thus, at least four taxa are necessary to identify the rate parameters and branch lengths under the F81 rate matrix. For a four-taxon tree, we will have four three-taxon subtrees and thus four equations similar to Equation (8) (more specifically, the equation just above equation (8)) in Allman et al. (2008) with two unknown parameters, \( \alpha \) and \( \pi \) to determine.

We have no proof to date that there cannot be multiple solutions to these four equations for all parameter choices, although we hypothesize that there will not be, at least for generic branch lengths. Using Rogers’ trick of assuming that two distinct models give the same data, we do have a simple proof that two models that lie in one part of bivariate parameter space can be distinguished. Let \( \mu \) be associated with tree \( T' = ((AB)(CD)) \) and rate distribution mixture parameters \( \alpha \) and \( \pi \). Let \( \nu \) be associated with an unknown tree \( T \) and parameters \( a \) and \( p \). Assume \( p \leq \pi \) and \( \alpha \leq a \). The first restriction, \( p \leq \pi \) simply places an ordering on the models. The line \( \alpha = a \) divides the ordered bivariate parameter space into two parts. If the models fall on one side of this line, the side where \( \alpha \leq a \), then we show below that they can be distinguished.

Let \( t_i \) be the interior branch length in \( T' \) and \( t_A, t_B, t_{CD}, t_{ABD} \) and \( t_{ABCD} \) be the branch lengths to the respective tips in \( T' \). Then applying Rogers’ trick to the four equations corresponding to Equation (8) or, more specifically, the equation appearing just above equation (8), in Allman et al. (2008), we have, for any F81 matrix with eigenvalue set to \(-1 \) (and for the moment generating functions evaluated at the negative value):

\[
\begin{align*}
\nu^{-1}(\mu(t_A + t_B)) + \nu^{-1}(\mu(t_A + t_{CD} + t_{ABD})) + \nu^{-1}(\mu(t_B + t_{CD} + t_{ABD})) &= 2
\nu^{-1}(\mu(t_A + t_B + t_{CD} + t_{ABD})) \\
\nu^{-1}(\mu(t_B + t_A)) + \nu^{-1}(\mu(t_B + t_B + t_{CD} + t_{ABD})) + \nu^{-1}(\mu(t_B + t_{CD} + t_{ABD})) &= 2
\nu^{-1}(\mu(t_A + t_B + t_{CD} + t_{ABD})) \\
\nu^{-1}(\mu(t_{CD} + t_{ABD})) + \nu^{-1}(\mu(t_A + t_{CD} + t_{ABD})) + \nu^{-1}(\mu(t_A + t_{CD} + t_{ABD})) &= 2
\nu^{-1}(\mu(t_A + t_{CD} + t_{ABD} + t_x)) \\
\nu^{-1}(\mu(t_{ABD} + t_{CD})) + \nu^{-1}(\mu(t_A + t_{CD} + t_x)) + \nu^{-1}(\mu(t_A + t_{CD} + t_x)) &= 2\nu^{-1}(\mu(t_A + t_{CD} + t_x))
\end{align*}
\]

As before, writing these equations in terms of the function \( f(x) = [(\pi - p) + (1 - \pi)(1 + x/\alpha)^{-\alpha}]^{-1/\alpha} \) and setting \( s = t_{CD} + t_x \) for \( s \), we see that the first two equations have the form:

\[
2f(t_A + t_B + s) - f(t_A + s) - f(t_B + s) = f(t_A + t_B) - f(0).
\]

When \( \left(\frac{1 - \pi}{\pi - p}\right) \left(\frac{\alpha - a}{\alpha(\alpha + 1)}\right) \leq 1 \), \( f \) is concave. In this case, one can show that there is no solution \( s \) unless \( a = \alpha \) and \( \pi = p \). When \( \left(\frac{1 - \pi}{\pi - p}\right) \left(\frac{\alpha - a}{\alpha(\alpha + 1)}\right) > 1 \), \( f \) has one inflection point and there appears to be at most one solution \( s \) implying that the parameters are identifiable for generic trees. However, we have been at a loss to prove this last point formally.

Discretized Gamma Distribution

Phylogenetic software often uses a discrete gamma distribution where the gamma distribution is broken into \( k \) intervals of equal probability and the mean of each interval is used to represent the interval. This changes the functional form of the moment generating function. Let \(-\lambda \) be a nonzero eigenvalue for the instantaneous rate matrix \( Q \) and \( \tau \) be the distance between a pair of species. The moment generating function evaluated at \(-\lambda \tau \) is then the function

\[
\mu(\lambda \tau) = \pi + \frac{1 - \pi}{k} \sum_{j=1}^{k} e^{-r_j(\alpha) \lambda \tau},
\]

where \( r_j(\alpha) \) is the mean of the gamma distribution, with shape parameter \( \alpha \) and mean 1, over the \( j \)th equally probable interval (out of \( k \) intervals in total). It is an open mathematical problem to determine whether or not the model with this discrete rate distribution is generically identifiable.
Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data

JOHN J. WIENS* AND MATTHEW C. MORRILL

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA;
*Correspondence to be sent to: Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA;
E-mail: wiensj@life.bio.sunysb.edu.

Received 22 May 2010; reviews returned 1 September 2010; accepted 21 February 2011

Associate Editor: Karl Kjer

This paper will attempt to resolve some controversies about the effects of missing data on phylogenetic analysis. Whether missing data are generally problematic is a critical issue in modern phylogenetics, especially as wildly different amounts of molecular data become available for different taxa, ranging from entire genomes, to single genes, to none (e.g., fossils). Our perception of the impact of missing data (or lack thereof) may strongly influence which taxa and characters we include in a phylogenetic analysis (Wiens 2006) and may lead to a diversity of serious errors. For example, if we think that missing data are problematic when they are not, then we may exclude taxa and characters that would otherwise benefit our analyses, given the abundant evidence that increasing numbers of both taxa and characters can potentially improve the accuracy of phylogenetic analyses (e.g., Huelsenbeck 1995; Rannala et al. 1998; Poe 2003), where accuracy is generally defined as the similarity between the estimated tree and the correct, known phylogeny. In contrast, if missing data cells are themselves intrinsically problematic (e.g., Huelsenbeck 1991), including taxa or characters with many missing data cells may lead to inaccurate phylogenetic estimates.

Several studies have explored how missing data may impact phylogenetic analyses, using both empirical and simulated data. Many simulation and empirical studies now suggest that it is often possible to include taxa that have large amounts of missing data without ill effects (e.g., Wiens 2003b; Driskell et al. 2004; Philippe et al. 2004; Wiens et al. 2005; Wiens and Moen 2008; Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010). However, a recent simulation study (Lemmon et al. 2009) suggested instead that missing (“ambiguous”) data are generally problematic for phylogenetic analysis and implied that these previous simulation and empirical studies are therefore incorrect. They justified their study based on the grounds that previous studies were supposedly in conflict about the impacts of missing data (p. 131).

In this paper, we will show that the paper by Lemmon et al. (2009; LEA hereafter) is problematic for several reasons. First, despite their statement that previous studies are in conflict, most simulation and empirical results on missing data can be easily explained within an existing theoretical framework (Wiens 2003b). Furthermore, many contradictory studies suggesting that missing data are not generally problematic for Bayesian and likelihood analyses (given some assumptions) were not addressed by LEA. Second, the sweeping negative conclusions of LEA are not necessarily supported by their results. LEA find missing data to be problematic primarily when using sets of invariant or saturated characters and/or when obvious rate heterogeneity is ignored. Their results do not support the idea that missing data generally lead to incorrect inferences about topology when informative data are analyzed with appropriate methods. We conduct new simulations under more realistic conditions, and these results show no evidence that missing data generally lead to inaccurate Bayesian estimates of phylogeny. In fact, we show that the practice of excluding characters simply because they contain missing data cells may itself reduce accuracy. We reanalyze the “manipulated” empirical example from LEA and find that, without these artificial “manipulations” of the data, their conclusions are not supported. We also analyze eight empirical data sets, each containing many taxa with extensive missing data. We show that these incomplete taxa are consistently placed into the expected higher taxa, often with very strong support. Overall, our results confirm previous simulation and empirical studies showing that taxa with extensive missing data can be accurately placed in phylogenetic analyses and that adding characters with missing data can be beneficial (at least under some conditions). We conclude by pointing out important areas for future research on the topic of missing data and phylogenetic analysis.

A General Framework for Interpreting Simulation and Empirical Results

There are two main ways that missing data might be added to a phylogenetic analysis, either through the addition of incomplete characters or incomplete taxa. For example, imagine having data from two genes for a given genus of organisms, in which the first gene has been sequenced for all 10 species and the second gene has been sequenced for only 5 species. Given this situation, one might (a) analyze only the first gene for all 10 species and then decide whether or not to (b) add...
the second gene (adding incomplete characters that are missing data for 5 of the species). Or, one might (c) analyze only the 5 species having data for both genes, and then decide whether or not to (d) add the 5 species lacking data for the second gene (adding incomplete taxa that are lacking data for the first gene). Note that (b) and (d) are effectively identical. Most of the literature on missing data has focused on whether to include taxa lacking data for some characters (c versus d). LEA did not actually address this question, but they imply that their results overturn earlier studies that did (e.g., p. 141). Below, we briefly review previous studies on incomplete taxa (treating fossil studies collectively rather than individually), and address incomplete characters (a vs. b) in our simulations and under “Areas for Future Research.”

Rather than being in conflict, we argue that most of the diverse empirical and simulation studies on missing data are largely consistent when viewed in light of the hypothesis that highly incomplete taxa can potentially be accurately placed if enough informative characters are sampled overall (Wiens 2003b). Thus, the apparent impacts of extensive missing data in these studies fall along a continuum (from negative to inconsequential) based on the overall number of characters in the analysis.

The issue of missing data first became prominent in association with parsimony analyses of morphological data for fossil taxa (e.g., Donoghue et al. 1989; Platnick et al. 1991). These studies have found incomplete taxa to be problematic in some cases (e.g., generating many equally parsimonious trees and poorly resolved consensus trees; Novacek 1992; Wilkinson 1995; Anderson 2001) but not others (e.g., Kearney 2002; Cobbett et al. 2007). However, these studies included relatively few characters (up to a few hundred, but often <100). The simulations of Huelsenbeck (1991) included only 100 characters and found highly incomplete taxa (75% missing data) were often problematic. Wiens and Reeder (1995) found that including highly incomplete taxa (75%) reduced accuracy somewhat in parsimony analyses of known viral phylogenies, but their two data sets (sequence and restriction site) each included less than 100 parsimony informative characters. Draggoo and Honeycutt (1997) showed that their parsimony analyses were largely insensitive to missing data with ~1500 characters (three mitochondrial genes for carnivorn mammals), with no effect on topology when one or two of the three genes were replaced with missing data cells, but that some resolution was lost when some taxa had ~87% missing data. The simulations of Wiens (2003b; see also 2003a) showed that highly incomplete taxa (e.g., 90% missing data) can be accurately placed given enough characters in parsimony, likelihood, and neighbor-joining analyses, but the exact level of character sampling needed depends on the phylogenetic method, distribution of missing data among characters, and branch lengths (e.g., accurate placement is more difficult with neighbor joining and parsimony, when missing data are randomly distributed among taxa, and when overall branch lengths are long and/or characters evolve rapidly).

Dunn et al. (2003) performed a limited set of simulations based on their data for 2293 rapidly evolving mitochondrial DNA characters for mylobatiform fish and found that the impact of incomplete taxa varied depending on the method, from negative (parsimony) to none (likelihood), relative to simulations in which all taxa were complete. Philippe et al. (2004) included 30,399 characters from 129 protein sequences among eukaryotes and found that highly incomplete taxa were placed with strong support in their empirical likelihood analyses (i.e., the four most incomplete taxa had 56%, 60%, 61%, and 76% missing data, and the likelihood bootstrap values placing them with their sister taxa are respectively 100%, 92%, 98%, and 95%). They also found high accuracy in their simulations based on those data (e.g., 100% accuracy for all nodes when 50% of the data were missing, and 89% mean accuracy across nodes when 90% were missing). Driskell et al. (2004) examined DNA data sets with very large numbers of characters (469,497 for metazoans and 96,698 for green plants) and extensive missing data (92% and 84%, respectively) and found that highly incomplete taxa were placed in clades expected from previous taxonomy with strong support based on parsimony bootstrapping. Wiens et al. (2005) included 3519 (mostly DNA) characters for treefrogs and showed that highly incomplete taxa were placed in the expected clades with very strong support by parsimony and Bayesian analyses (e.g., 10 species with >90% missing data each were all placed in the expected clades, with monophyly of these clades each supported with a Bayesian posterior probability (PP) of 1.00). Other recent empirical studies (described below) have also shown that highly incomplete taxa are placed in the expected clades with strong support, and most of these studies included >4000 characters each (e.g., Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010).

Wiens (2005) used simulations to show that adding highly incomplete taxa (i.e., 90% missing data) could be as effective as complete taxa in rescuing likelihood and Bayesian analyses from long-branch attraction, even when the models utilized in these analyses were misspecified (i.e., among-site rate heterogeneity and transition-termination bias were simulated but not included in the estimation models), given a sample of 1000 characters. The simulations of amino acid data by Hartmann and Vision (2008) showed reduced accuracy with extensive missing data for parsimony, likelihood, and neighbor-joining analyses, but only included 500 characters. Wiens and Moen (2008; their fig. 2) used simulations to show that highly incomplete taxa could be accurately placed in Bayesian analyses given enough characters (e.g., 2000), even when rate heterogeneity and substitution bias were simulated but not included in the Bayesian model.

In summary, all of these simulation and empirical studies seem to fit into this common framework, with highly incomplete taxa being potentially problematic...
when the overall number of characters is small and generally unproblematic when the number is large. This common framework seems to apply to all phylogenetic methods, not simply likelihood and Bayesian analysis.

The results of many of these studies contradict the conclusions of LEA but were not mentioned by them, including the ones that addressed the impact of missing data on likelihood and Bayesian analyses (e.g., the results of Philippe et al. 2004; Wiens 2005; Wiens et al. 2005 were not mentioned, and the latter two studies were not cited). For example, LEA conclude that “in both ML and Bayesian frameworks, among-site rate variation can interact with ambiguous data to produce misleading estimates of topology” (p. 130) and that estimation becomes problematic “when rate variation across sites is not properly modeled” (p. 141). But the simulation studies by Wiens (2005) and Wiens and Moen (2008) showed accurate estimation of topologies with incomplete taxa by Bayesian and likelihood methods when rate variation was simulated but completely ignored.

**Problematic Simulations and Conclusions of Lemmon et al. (2009)**

LEA analyzed a very limited set of simulated conditions and found results that were seemingly discordant with those of other simulation studies of the same topic. Yet, they make sweeping conclusions from their results (e.g., that their results have implications for “all analyses that rely on accurate estimates of topology or branch lengths”, p. 130). They also imply that their results overturn those of previous studies. It is therefore important to look at what they did and found more closely.

LEA examined the four-taxon case, with the simplest model of sequence evolution (Jukes–Cantor), and equal branch lengths (given that an unrooted tree is estimated). For each set of conditions, they simulated two data sets (Gene A and Gene B), one with complete data for all taxa and characters, and another lacking data for all characters for two taxa. These genes were simulated under either the same or different rates of change. They then evaluated Bayesian Pp for the single internal node for Gene A alone and for Gene A and B combined. For maximum likelihood, they evaluated the frequency with which this clade was correctly reconstructed. They assumed that the combined data would give the same results as Gene A alone because data were only present in two of the four taxa for Gene B (making Gene B uninformative under the parsimony criterion, but note that this is not necessarily true for likelihood or Bayesian analysis, see below).

They found that Bayesian Pp for the combined data sometimes differed from the observed values based on Gene A alone (but for maximum likelihood a comparable result only occurred when branch lengths were arbitrarily fixed to nonzero values). They refer to these differences as “bias.” In some cases, these biases appear to be problematic, as when Pp approaches zero for the true tree (such that the true phylogeny is not estimated). Similarly, they found some cases where Pp was very high for the true tree, even though both data sets were effectively invariable. They suggested that these biases were related to the Bayesian star-tree paradox (e.g., Lewis et al. 2005), the tendency for Bayesian analysis to strongly favor one tree when there is little information with which to choose among trees (regardless of missing data). However, they found virtually none of these extreme biases unless the characters were effectively invariant or “saturated” (i.e., used by LEA as meaning so variable so as to be effectively uninformative), and unless rate heterogeneity between genes was simulated and then ignored by failing to partition by genes. The only exception we find is in their fig. 4, for one set of conditions with very high rates in both genes and data missing in sister taxa. Thus, their results do not support their sweeping generalizations about the negative impact of missing data, especially for conditions likely to be encountered by most empirical systematists. This presumably explains why so many previous simulation and empirical studies contradict their conclusions about the negative impact of missing data on Bayesian and likelihood analyses (see above).

In some cases, they show that combined data Pp differ moderately from those for Gene A alone (e.g., their fig. 4, when rate of Gene A is low). However, arguing that these Pp are “biased” assumes that Gene B has no influence on topology estimation whatsoever (i.e., Pp for the combined analysis should be the same as for Gene A alone). Although this is true for parsimony, it is not necessarily true for Bayesian or likelihood analyses. For example, if Gene B has no influence on Bayesian estimates of topology (which are based on Pp), then how can it influence Pp at all? Clearly, the initial assumption that incomplete characters in Gene B have no impact cannot be fully correct. Furthermore, finding that combined data Pp differ from Pp for Gene A is not direct evidence that combined data Pp yield biased estimates of accuracy (i.e., in this context, the probability that the clade is correctly reconstructed by the method under a given set of conditions). Demonstrating bias would require directly examining the relationship between accuracy (the probability that the clade is correctly reconstructed by the Bayesian analysis and data missing in sister taxa. Thus, their results do not support their sweeping generalizations about the negative impact of missing data, especially for conditions likely to be encountered by most empirical systematists. This presumably explains why so many previous simulation and empirical studies contradict their conclusions about the negative impact of missing data on Bayesian and likelihood analyses (see above).

In summary, the results of LEA suggest that missing data are primarily problematic when utilizing uninformative characters and/or when failing to partition clearly heterogeneous data sets, conditions that may not be routinely encountered by most systematists. This is not to say that we think that data sets with missing data always yield accurate phylogenies with
unbiased support values, but rather that the simulation results of LEA can have a very different interpretation from their sweeping negative conclusions, if one simply considers which of their results are relevant to what phylogeneticists actually do.

Another critical issue is the addition of sets of characters with data for only two species, which are expected to have little impact on the analysis (and which presumably would not be used by empirical systematists). It is unclear if their results are specific to adding only two species or if they also apply to larger numbers of species. We have therefore performed new simulations to address the relevance of the results of LEA to larger numbers of taxa.

**NEW SIMULATIONS**

**Methods**

We addressed how adding data from a gene with incomplete taxon sampling to the one with complete taxon sampling influences the accuracy of Bayesian phylogenetics. Simulation methods generally followed Wiens and Moen (2008). We simulated 16-taxon phylogenies that were either fully asymmetric or symmetric. Following LEA, we simulated DNA data with the Jukes-Cantor model with equal branch lengths across the tree. We generally used 500 characters per gene, but also simulated 100. Data set 1 had all characters for all 16 taxa. For Data set 2, we simulated the complete data, and then a set of taxa was randomly selected in each replicate to have all their characters replaced with missing data cells. In one set of simulations, eight taxa were incomplete in Data set 2 (one way of generalizing the design of LEA to larger numbers of taxa). In another set of simulations, 14 taxa were incomplete, leaving only 2 taxa in Data set 2 with nonmissing data (an alternate way of generalizing the design of LEA to more taxa). Data were simulated under a broad range of rates of change, from very low (probability of change in a given character along a given branch of 0.0001) to relatively high (0.50), and six intermediate rates (0.001, 0.01, 0.10, 0.20, 0.30, 0.40). Initial analyses on the asymmetric tree showed that the most extreme rates gave relatively low accuracy for these conditions (30% of tree or less resolved correctly). We simulated both equal rates in each data set and many unequal rates (Fig. 1), but not every possible combination of rates. We analyzed 100 replicates for each set of conditions. We analyzed Data set 1 alone and then analyzed data sets 1 and 2 combined. Data sets were analyzed using MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001), assuming a Jukes–Cantor model with a parameter for unequal rates of change among sites (gamma), and other options set to default values. Importantly, combined analyses were partitioned, allowing a different value for gamma in each data set. Analyses were run for 50,000 generations each, sampling every 100 generations, and excluding the first 10,000 generations as burn-in. These settings provide adequate tree searches for these conditions (Wiens and Moen 2008). We evaluated accuracy for each replicate as the proportion of resolved nodes in the majority-rule Bayesian tree that are shared with the known, true topology, and overall accuracy (for a given set of conditions) is based on the mean for 100 replicates. This measure of accuracy is used in many previous simulation studies (e.g., Wiens 2003b; Wiens and Moen 2008); other measures are certainly possible, but they should also reflect the similarity between the true and estimated trees averaged across replicates. We did not directly evaluate Pp support for individual clades, but a clade will not be resolved unless its Pp is >0.50, and LEA did not directly examine the relationship between accuracy and Pp either.

**Results**

We find that across a broad range of conditions (Fig. 1), adding the data set consisting of 50% missing data (8 of 16 taxa incomplete) either increases or has no effect on accuracy, relative to analyzing the complete data set alone. Although the increases are typically small, under some conditions, the relative increase can be >20% (e.g., 0.49 vs. 60). These increases in accuracy may occur when the rates of change in the two data sets are equal, or when they are very unequal as well. When the added data set has only two complete taxa (as in the simulations of LEA), accuracy may be slightly higher or slightly lower than Data set 1 alone, and is consistently within 0.05. These latter results suggest that adding sets of characters with only two species has little influence on the overall accuracy of analyses with larger numbers of taxa, and that the design and results of LEA do not generalize to more realistic conditions.

**Discussion**

Contrary to the conclusions of LEA, we find no evidence that adding sets of characters with extensive missing data leads to misleading estimates of Bayesian phylogeny or support values (i.e., only clades with Pp > 0.50 are supported). Importantly, our results suggest that under some conditions, failing to add characters with missing data can lead to reduced phylogenetic accuracy. Thus, being overly cautious about excluding characters simply because they have missing data can lead to reduced phylogenetic accuracy. This is a critical point that LEA do not discuss.

Our simulation methods are not identical to those of LEA. For example, we assume that researchers will not choose to analyze data sets that completely lack phylogenetic information due to rates that are too fast or slow (and so we do not simulate these conditions, but we do simulate branches that are both extremely long and extremely short). We also assume that most researchers will partition data sets evolving at different rates. But most importantly, our results suggest that the misleading Bayesian estimates noted by LEA do not necessarily occur under slightly more realistic conditions (e.g., more taxa, partitioned data, and
use of variable characters). As one example, LEA suggest that Bayesian Pp may be strongly influenced by whether the taxa with nonmissing data are sister or nonsister taxa, but this simple division becomes unclear when additional taxa are included. For example, given a five-taxon tree (A, B) (C (D, E), with missing data in species C and D, nonmissing data are simultaneously present in both sister taxa (A, B) and nonsister taxa (A, E).

These simulations are also very limited and still very far from realistic. Many parameters that could have been varied were not (e.g., more complex substitution models, variation in rates within genes), in order to make the results more comparable with those of LEA (see instead Wiens 2005; Wiens and Moen 2008). Perhaps the most important oversimplification is the use of equal branch lengths throughout the tree. In order to address how missing data influence Bayesian and likelihood analyses...
under fully realistic conditions, we also perform analyses of eight empirical data sets. Before we do that, we briefly address the empirical example offered by LEA.

Reexamining the Manipulated Empirical Example of LEA

LEA analyzed an empirical data set but again made many methodological choices that make these data very different from those analyzed by empirical systematists. They analyzed data from a single mitochondrial gene from eight species of plethodontid salamanders (but for which entire mitochondrial genomes were available; Mueller et al. 2004), deliberately analyzing a very small number of characters. We could not find an explanation for why these particular characters and taxa were chosen. They then added a second data set consisting of missing data for six of the eight species and “manipulated empirical data” for the other two. These added, nonmissing data consist of resampled sites from the same gene for the same species, selected to be either all invariant or all variable between the two species. Although they found that adding this second set of characters influenced Bayesian and likelihood estimates of topology, support, and branch lengths, this analysis raises many questions about its design. Why not use actual data (e.g., another gene) instead of resampling sites from the same gene? Why only variable and invariant sites? To what extent are their results an artifact of these methodological choices?

We addressed this latter question using very similar empirical data, and our results offer a dramatic contrast to those of LEA (Fig. 2). We downloaded the same 16S data, but instead of adding only invariant or variant sets of characters from the same gene, we added unmanipulated data from another gene (the widely used cytochrome b; again from Mueller et al. 2004) to the same species to which LEA added data. Clearly, adding another gene is more relevant to what empirical systematists actually do. Instead of finding that “ambiguous characters can strongly bias estimates of topological support and branch lengths” (p. 139) we find that Bayesian and likelihood estimates of topology, support, and branch lengths are almost identical after adding cytochrome b with data missing in six species (see Fig. 2 legend for methods). As in their simulations, it appears that the results of LEA reflect artifacts of adding invariant and saturated characters (and failing to partition data sets), and therefore may have limited relevance to most empirical studies.

Apart from their example involving “manipulated” empirical data, LEA do not show any empirical studies in which missing data seem to be problematic for Bayesian or likelihood methods. Note that on p. 142, LEA state “One of us (K.S.-H.) has come across such an example of discordance among gene trees in empirical data from North American fireflies. Once ambiguous sites were excluded from the analysis, gene tree congruence increased substantially (Stanger-Hall et al. 2007).” However, the only references to missing data in that paper were the following quotes: “However, due to stretches of missing data in individual taxa (due to differences in primer binding and sequencing success) and the possibility that these unduly influence the phylogenetic analysis (Lemmon et al. unpublished data), the final alignment was reduced to 1906 bp.” (p. 36) and also (p. 42) “it seems to have a significant effect on the outcome of a ML and/or Bayesian analysis (Lemmon et al., unpublished data). This led us to exclude DNA segments with missing data for more than one taxon from our final alignment.” Thus, the Stanger-Hall et al. (2007) paper does not contain the empirical results that LEA state that it does, only references to LEA.

New Results from Empirical Data Sets

The problem of missing data is something that empirical phylogeneticists may encounter every day. LEA state that the supposed negative impacts of missing data on phylogenetic analysis are relevant to “all studies” that estimate and use phylogenies (p. 130). If this were true, we would expect to see widespread negative impacts in empirical analyses that include extensive missing data. We have previously described empirical studies that showed evidence that such impacts can be small or nonexistent (e.g., Philippe et al. 2004; Wiens et al. 2005), specifically for likelihood and Bayesian analyses. Here we present analyses of eight additional empirical data sets that show similar patterns.

Obviously, the true phylogeny is unknown in most empirical data sets. However, one can make predictions about how methods will perform with real data given the results of simulations. It is not immediately clear what specific empirical predictions can be derived from the simulations of LEA. Nevertheless, they state that extensive missing data may “positively mislead” (p. 143) estimated topologies in likelihood and Bayesian analyses. If this were the case, then we predict that highly incomplete taxa will be placed in clades that appear to be incorrect based on previous taxonomy and systematic research (i.e., assessing accuracy based on congruence). In contrast, if the hypothesis of Wiens (2003b) is correct, and if sufficient characters have been sampled, then we expect that incomplete taxa will be placed in the expected higher taxa (e.g., genera, families), and with strong support. In addition, there should be strong support for the localized placement of these species within these higher level taxa (if sufficient characters were sampled). Following Wiens et al. (2005), we test for a negative relationship between localized clade support and incompleteness of individual taxa.

Methods

We selected eight published empirical data sets (Table 1), all involving Bayesian or likelihood analyses of mostly nonoverlapping vertebrate clades. All eight
Analyses of mitochondrial DNA sequence data in plethodontid salamanders show that adding sets of characters with extensive missing data may have negligible impacts on topology, support, and branch lengths (contrast with fig. 7 of LEA). We obtained the same 16S data for the same species as LEA (aligned using MUSCLE; Edgar 2004), but instead of adding resampled invariant and variable sets of characters from 16S, we added data from another gene (cytochrome b; cyt b). Data are added for the same pairs of sister and nonsister species (boldfaced) used by LEA; all other species have missing data cells for cyt b. Bayesian analyses used MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001) with the GTR + I + Γ model and 1,000,000 generations. Likelihood analyses used RAxML version 7.2 (Stamatakis 2006) with the recommended GTR + Γ model, with 100 integrated bootstrapping and heuristic search replicates. Both analyses were partitioned by gene. Numbers adjacent to branches indicate posterior probabilities (Bayesian) or bootstrap values (likelihood). Trees are unrooted (but see Kozak et al. 2009 for justification for rooting near Ensatina and Hydromantes).

For each data set, we quantified the percentage of missing data cells in each species. Most missing data originate from the complete absence of data from one or more genes (or parts of genes) in combined analyses, but a small fraction also comes from gaps in alignments. Levels of branch support were based on whichever model-based method was used in the original study (i.e., Bayesian vs. likelihood); we arbitrarily selected likelihood for Hua et al. (2009), which used both. For the ranid and phrynosomatid data sets, we used 49% as the lowest bootstrap, as specific values <50% were unavailable; however, relatively few nodes had values <50% (14.6% for ranids, 7.3% for phrynosomatids) and using reasonable alternate values (e.g., 25%) gave identical correlation results. Detailed methods are described in the original studies. However, given that effects of missing data may depend on how a software package treats these cells, we note that maximum likelihood analyses used RAxML (Stamatakis 2006) and Bayesian analyses used MrBayes (Huelsenbeck and Ronquist 2001).

We first evaluated whether highly incomplete taxa are placed in the clades expected based on previous taxonomy, and whether they are placed in these clades with strong support. If highly incomplete taxa are generally problematic, then they should not be consistently placed in the clades predicted by previous taxonomy, or if they are, the support for these clades should be weak. For each data set, we identified a set of nonnested clades from previous taxonomy. These mostly consisted of genera, as the generic-level assignment of most of these species was previously established based on
nonmolecular data. However, for higher-level snake phylogeny, with only one species per sampled genus, we used families (and well-established subfamilies for Colubridae). For ranids, we used subfamilies, given that the taxonomy for these clades is relatively stable (e.g., Bossuyt et al. 2006; Frost et al. 2006) whereas generic-level taxonomy is not (e.g., Frost et al. 2006 vs. Wiens et al. 2005). The nonmonophyly of \( P. \) bolitoglossine salamanders, other factors besides missing data that are involved. In summary, if missing data are generally problematic and completeness is not inconsistent with the mechanism proposed by Wiens (2003b). If there are too little data to accurately place a taxon on the tree, then the support for its placement should be weak. However, the simulations of Wiens (2003b) suggest that, given enough characters, even highly incomplete taxa will be accurately placed in the phylogeny with high consistency. This should be reflected with high support values.

Finally, we note that this analysis does not necessarily address whether support values are biased by missing data, unless they are strongly biased to be consistently positive or negative (but LEA do not address moderate biases either because they did not directly test how accuracy and support values are related).

### Results

The eight data sets collectively include >1000 species and >60 higher taxa, and almost all of these species are placed in the expected higher taxa, despite many having extensive missing data (Tables 1 and 2). Furthermore, the monophyly of most of these clades is strongly supported (Bayesian PP = 1.00; likelihood bootstrap = 78%–100%, but most >90%). In the three cases in which genera are not monophyletic in these data sets, there are other factors besides missing data that are involved. In bolitoglossine salamanders, \( Pseudoeurycea \) and \( Linotriton \) both appear to be nonmonophyletic (Wiens, Parra-Olea, et al. 2007), but previous phylogenetic studies with little missing data suggest that this reflects parallel evolution and misleading taxonomy (Parra-Olea and Wake 2001). The nonmonophyly of \( T. \) seems to reflect conflict between mitochondrial and nuclear genes, not missing data per se (Wiens, Kuczynski, Arif, et al. 2010).

In summary, if missing data are generally problematic as LEA suggest, there does not seem to be any evidence for it in these eight data sets (unless the previous nonmolecular taxonomies in these groups have been misled in a way that is consistent with the misleading effects of missing data on likelihood and Bayesian analyses of DNA sequence data).

Only two of the eight studies show significant negative relationships between branch support and incompleteness (Fig. 3). These results suggest that missing data have little consistent negative impact on levels of branch support, and there is sometimes strong
Table 2. Summary of support for previously recognized higher taxa (genera, families, subfamilies) within the eight data sets, showing that almost all higher taxa are strongly supported as monophyletic, despite many of them containing one or more taxa with extensive missing data.

<table>
<thead>
<tr>
<th>Clade</th>
<th>Higher taxon</th>
<th>Method</th>
<th>Support</th>
<th>Maximum incompleteness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plethodontid salamanders</td>
<td>Batrachoseps</td>
<td>ML</td>
<td>84</td>
<td>93.5</td>
</tr>
<tr>
<td></td>
<td>Gymnophiilus</td>
<td>ML</td>
<td>100</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>Pseudotriton</td>
<td>ML</td>
<td>78</td>
<td>47.1</td>
</tr>
<tr>
<td></td>
<td>Eurycea</td>
<td>ML</td>
<td>100</td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td>Plethodon</td>
<td>ML</td>
<td>100</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td>Hydromantes</td>
<td>ML</td>
<td>100</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>Ensatina</td>
<td>ML</td>
<td>100</td>
<td>50.3</td>
</tr>
<tr>
<td></td>
<td>Anciles</td>
<td>ML</td>
<td>100</td>
<td>41.6</td>
</tr>
<tr>
<td></td>
<td>Desmognathus</td>
<td>ML</td>
<td>100</td>
<td>80.4</td>
</tr>
<tr>
<td>Bolitoglossine salamanders</td>
<td>Cryptotriton</td>
<td>BA</td>
<td>1.00</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>Dendrotriton</td>
<td>BA</td>
<td>1.00</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>Nototriton</td>
<td>BA</td>
<td>1.00</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>Oedipina</td>
<td>BA</td>
<td>1.00</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>Thorius</td>
<td>BA</td>
<td>1.00</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Chloropterotriton</td>
<td>BA</td>
<td>1.00</td>
<td>71.9</td>
</tr>
<tr>
<td></td>
<td>Pseudoeurycea</td>
<td>BA</td>
<td>Not supported</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>Ixalotriton (nested inside Pseudoeurycea)</td>
<td>BA</td>
<td>1.00</td>
<td>49.7</td>
</tr>
<tr>
<td></td>
<td>Lineotriton (nested inside Pseudoeurycea)</td>
<td>BA</td>
<td>Not supported</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>Bolitoglossa</td>
<td>BA</td>
<td>1.00</td>
<td>70.8</td>
</tr>
<tr>
<td>Treefrogs (Hyla)</td>
<td>Tlacochyla</td>
<td>ML</td>
<td>100</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Isthmohyla</td>
<td>ML</td>
<td>100</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Smilisica</td>
<td>ML</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Hyla</td>
<td>ML</td>
<td>79</td>
<td>96</td>
</tr>
<tr>
<td>Hemiphractid frogs</td>
<td>Flectonotus</td>
<td>BA</td>
<td>1.00</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>Hemiphractus</td>
<td>BA</td>
<td>1.00</td>
<td>68.7</td>
</tr>
<tr>
<td></td>
<td>Stefania</td>
<td>BA</td>
<td>1.00</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td>Gastrotheca</td>
<td>BA</td>
<td>1.00</td>
<td>77.4</td>
</tr>
<tr>
<td>Ranid frogs</td>
<td>Ptychadeniae</td>
<td>ML</td>
<td>96</td>
<td>90.8</td>
</tr>
<tr>
<td></td>
<td>Phrynobatrachinae</td>
<td>ML</td>
<td>94</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>Conrauinae</td>
<td>ML</td>
<td>100</td>
<td>44.8</td>
</tr>
<tr>
<td></td>
<td>Petropetidinae</td>
<td>ML</td>
<td>100</td>
<td>54.2</td>
</tr>
<tr>
<td></td>
<td>Pyxicephalinae</td>
<td>ML</td>
<td>86</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Micrixalinae</td>
<td>ML</td>
<td>100</td>
<td>55.2</td>
</tr>
<tr>
<td></td>
<td>Dicroglossinae</td>
<td>ML</td>
<td>96</td>
<td>89.8</td>
</tr>
<tr>
<td></td>
<td>Ranixalinae</td>
<td>ML</td>
<td>88</td>
<td>87.1</td>
</tr>
<tr>
<td></td>
<td>Ceratobatrachinae</td>
<td>ML</td>
<td>100</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>Nyctibatrachinae</td>
<td>ML</td>
<td>85</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>Mantellinae</td>
<td>ML</td>
<td>99</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>Rhacophorinae</td>
<td>ML</td>
<td>94</td>
<td>90.3</td>
</tr>
<tr>
<td></td>
<td>Rainae</td>
<td>ML</td>
<td>88</td>
<td>90.7</td>
</tr>
<tr>
<td>Emydid turtles</td>
<td>Glyptemys</td>
<td>BA</td>
<td>1.00</td>
<td>22.3</td>
</tr>
<tr>
<td></td>
<td>Terrapene</td>
<td>BA</td>
<td>1.00</td>
<td>42.6</td>
</tr>
<tr>
<td></td>
<td>Pseudemys</td>
<td>BA</td>
<td>1.00</td>
<td>33.1</td>
</tr>
<tr>
<td></td>
<td>Trachemys</td>
<td>BA</td>
<td>Not supported</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>Malaclemys</td>
<td>BA</td>
<td>1.00</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>Graptemys</td>
<td>BA</td>
<td>1.00</td>
<td>57.6</td>
</tr>
<tr>
<td>Phrynosomatid lizards</td>
<td>Holbrookia</td>
<td>ML</td>
<td>100</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>Uma</td>
<td>ML</td>
<td>100</td>
<td>81.5</td>
</tr>
<tr>
<td></td>
<td>Phrynosoma</td>
<td>ML</td>
<td>100</td>
<td>92.3</td>
</tr>
<tr>
<td></td>
<td>Uta</td>
<td>ML</td>
<td>100</td>
<td>53.2</td>
</tr>
<tr>
<td></td>
<td>Petrosaurus</td>
<td>ML</td>
<td>100</td>
<td>37.1</td>
</tr>
<tr>
<td></td>
<td>Urosaurus</td>
<td>ML</td>
<td>100</td>
<td>52.6</td>
</tr>
<tr>
<td></td>
<td>Sceloporus (including Sator)</td>
<td>ML</td>
<td>84</td>
<td>91.9</td>
</tr>
<tr>
<td>Snakes</td>
<td>Tropidophiidae</td>
<td>ML</td>
<td>100</td>
<td>16.9</td>
</tr>
<tr>
<td></td>
<td>Pythonidae</td>
<td>ML</td>
<td>100</td>
<td>51.4</td>
</tr>
<tr>
<td></td>
<td>Uropeltidae</td>
<td>ML</td>
<td>100</td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td>Boiidae</td>
<td>ML</td>
<td>100</td>
<td>47.7</td>
</tr>
<tr>
<td></td>
<td>Viperidae</td>
<td>ML</td>
<td>100</td>
<td>23.2</td>
</tr>
<tr>
<td></td>
<td>Atractaspidae</td>
<td>ML</td>
<td>100</td>
<td>68.9</td>
</tr>
<tr>
<td></td>
<td>Boodontidae</td>
<td>ML</td>
<td>100</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>Elapidae</td>
<td>ML</td>
<td>100</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>Colubridae-Xenodontinae</td>
<td>ML</td>
<td>100</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td>Colubridae-Colubrinae</td>
<td>ML</td>
<td>100</td>
<td>15.8</td>
</tr>
<tr>
<td></td>
<td>Colubridae-Natricinae</td>
<td>ML</td>
<td>100</td>
<td>58.3</td>
</tr>
</tbody>
</table>

Notes: BA = Bayesian analysis; ML = maximum likelihood.
support for the localized phylogenetic placement of taxa with >90% missing data (Fig. 3), within these expected higher taxa. Interestingly, the two data sets with significant relationships between support and completeness (plethodontids, ranids) have the largest numbers of taxa but only modest numbers of characters (Table 1). Again, we note that when too few informative characters have been sampled in a taxon, we expect only weak support for its placement in the tree.

Other Studies

In addition to these eight studies and others mentioned previously (e.g., Driskell et al. 2004; Philippe et al. 2004; Wiens et al. 2005), other recent studies have also shown similar patterns (e.g., Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010; Pyron et al. 2011). For example, Lynch and Wagner (2010) examined boid snake relationships with a Bayesian analysis of 14,417 molecular characters, with some taxa 98% incomplete and each taxon having an average of 70% missing data. Yet, their phylogeny is generally strongly supported and congruent with previous hypotheses and taxonomy (e.g., of six genera with >1 species, five are strongly supported as monophyletic with Pp > 0.98). Wiens, Kuczynski, Townsend, et al. (2010) showed that addition of >15,000 molecular characters to a data set of 363 morphological characters for squamate reptiles did not change the placement of most fossil taxa in a combined Bayesian analysis (despite the fossils having >98% missing data in this analysis) and caused no significant change in Bayesian Pp for fossil taxa. Furthermore, the placement of fossil taxa was consistent with previous taxonomy (e.g., fossil snakes placed in snakes), both before and after addition of molecular data.
Areas for Future Research

There are now many studies showing concordant support for the idea that highly incomplete taxa can be accurately placed in model-based analyses, and sweeping statements about the negative impacts of missing data are not substantiated. Nevertheless, many other aspects of the potential impact of missing data on phylogenetic analysis are still in need of further research.

Adding Characters with Missing Data

In addition to the effects of incomplete taxa, another major question is: given a complete set of characters for a set of taxa, is it useful to add a second set of characters that are incomplete (because they include data for only some of the taxa)? In other words, when do the benefits of adding more characters outweigh the potential disadvantages of increasing missing data in the matrix? Superficially, it might seem that the simulations of LEA addressed this question. However, their results may be of limited relevance to empirical studies because only two species were added. Our simulations here (Fig. 1) suggest that adding a set of characters with data for 50% of the species is generally either beneficial or harmless for Bayesian analysis. However, these simulations were not comprehensive either, and additional analyses are needed (e.g., exploring unequal branch lengths, different numbers of characters, and different levels of taxon sampling).

Other simulation and empirical studies have also found results suggesting that incomplete characters can be beneficial, but with some caveats. Wiens (1998) found that adding sets of incomplete characters can potentially increase accuracy for parsimony, but that accuracy was increased more by distributing the same amount of added data among fewer taxa and more characters (and with less missing data). He also found potential problems of long-branch attraction when a set of highly incomplete characters is added.

Wiens et al. (2005) showed that adding a set of slow-evolving characters (nuclear genes) available for only some taxa (and with much missing data) seemed to improve results relative to those from analyzing only fast-evolving characters (mitochondrial genes) for a larger number of taxa. Specifically, some taxa are apparently misplaced in the analysis of fast-evolving characters alone (based on previous taxonomy), but not in the combined analysis.

The simulations of Gouveia-Oliveira et al. (2007) showed that accuracy of likelihood analyses was much higher when sequences with gaps (i.e., missing data) are included rather than excluded. Similarly, Wiens (2009) used simulations to address whether adding molecular data improves phylogenetic accuracy for fossil taxa, in a combined analysis of molecular and morphological data, with parsimony and Bayesian analysis (where the molecular data are missing in the fossil taxa). These simulations showed that under many conditions, adding molecular data improved accuracy for fossil taxa. A review of empirical studies (Wiens 2009) showed that adding molecular data can improve resolution (i.e., resolve polytomies in consensus trees) for the placement of fossil taxa, at least in some parsimony analyses (e.g., Manos et al. 2007). An analysis of squamate reptiles (Wiens, Kuczynski, Townsend, et al. 2010) confirmed that molecular data can change the placement of some fossil taxa, in addition to increasing resolution.

Estimating Divergence Times

It would also be worthwhile to investigate the effects of missing data on estimation of divergence times. LEA state that their results on branch length estimation are relevant to this issue, but they acknowledge that their results may be an artifact of not including rate heterogeneity in the likelihood model (p. 139), and this latter hypothesis is supported by our analyses also (Fig. 2). Furthermore, their study contains no actual estimation of divergence dates. We have conducted several divergence-dating analyses using matrices that contain extensive missing data (e.g., Wiens, Parra-Olea, et al. 2007; Kozak et al. 2009; Wiens et al. 2009), using both penalized likelihood and Bayesian approaches (Sanderson 2002; Drummond et al. 2006). Yet, we have found no evidence to suggest that these estimates are generally misled by missing data. Instead, these estimates are generally similar to those for the same groups based on smaller data sets with fewer missing data cells (e.g., Bossuyt et al. 2006 vs. Wiens et al. 2009 for ranid frogs; Wiens 2007 vs. Kozak et al. 2009 for plethodontid salamanders). But again, this is an area in need of further investigation.

Other Areas

Many other areas remain to be investigated. For example, it is unclear how congruence among gene trees may interact with missing data to impact phylogenetic accuracy. All simulation studies published so far have assumed that different genes share the same history, and have been based on combined analysis of genes (either implicitly or explicitly). The impact of missing data on methods that estimate species trees without concatenation (e.g., Edwards et al. 2007) also requires study.

The impact of missing data on support values would also benefit from additional study. For examples, simulations are needed to address whether the standard interpretation of support values (e.g., likelihood bootstrap support, Bayesian Pp) remains valid for taxa with extensive missing data.

Conclusions

LEA state (p. 130) that the results of their study ‘‘have major implications for all analyses that rely on accurate estimates of topology or branch lengths, including divergence time estimation, ancestral state reconstruction, tree-dependent comparative methods, rate variation analysis, phylogenetic hypothesis testing, and
phylogeographic analysis.” However, examination of their results shows that their evidence for the negative impacts of missing data hinge largely on methodological choices that would presumably not be made by most empirical systematists (e.g., adding data sets consisting of invariant or “saturated” characters, failing to partition data sets evolving at dramatically different rates). Unless those choices are made, their sweeping generalizations are not supported by their own results. These generalizations are also contradicted by many previous simulation and empirical studies, and also by new results from simulations that incorporate larger numbers of taxa and data partitioning (Fig. 1), from reanalysis of their plethodontid salamander example (Fig. 2), and from eight empirical data sets analyzed here (Fig. 3).

In contrast to the idea of discordance among studies promoted by LEA, we argue that most results on missing data can be explained in a common theoretical framework (Wiens 2003b), and that most studies suggest that it should generally be possible to accurately place incomplete taxa in phylogenies, if enough informative characters are sampled. We think that there is a need for continued investigation of the impact of missing data on phylogenetics, and we point out specific topics in particular need of focused research. However, future studies should strive to reconcile their new results with those from previous studies in order to make real progress in this area.

FUNDING
This work was supported by the U.S. National Science Foundation (EF 0334923 to J.J.W).

ACKNOWLEDGMENTS
For initial comments on the manuscript, we thank N. Butler, M. C. Fisher-Reid, X. Hua, D. Moen, A. Pyron. For comments on the submitted version, we thank the reviewers (H. Philippe, A. Lemmon, and anonymous) and associate editor (K. Kjer).

REFERENCES


