EVOLUTION

INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

PUBLISHED BY

THE SOCIETY FOR THE STUDY OF EVOLUTION

DETECTING THE HISTORICAL SIGNATURE OF KEY INNOVATIONS USING STOCHASTIC MODELS OF CHARACTER EVOLUTION AND CLADOGENESIS

RICHARD H. REE

Department of Botany, Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, Illinois 60605
E-mail: rree@fieldmuseum.org

Abstract.—Phylogenetic evidence for biological traits that increase the net diversification rate of lineages (key innovations) is most commonly drawn from comparisons of clade size. This can work well for ancient, unreversed traits and for correlating multiple trait origins with higher diversification rates, but it is less suitable for unique events, recently evolved innovations, and traits that exhibit homoplasy. Here I present a new method for detecting the phylogenetic signature of key innovations that tests whether the evolutionary history of the candidate trait is associated with shorter waiting times between cladogenesis events. The method employs stochastic models of character evolution and cladogenesis and integrates well into a Bayesian framework in which uncertainty in historical inferences (such as phylogenetic relationships) is allowed. Applied to a well-known example in plants, nectar spurs in columbines, the method gives much stronger support to the key innovation hypothesis than previous tests.

Key words.—*Aquilegia*, Bayesian inference, character mapping, diversification rate, macroevolution, phylogeny.

Received June 11, 2004.   Accepted November 10, 2004.

The tempo of evolution is a subject of general interest to evolutionary biologists, from nucleotide mutation rates within populations to the proliferation of branches on the tree of life. At the macroevolutionary level, key innovation hypotheses (Sanderson and Donoghue 1994) have been put forth to explain unusual disparities in species number between clades. A key innovation is commonly regarded as a biological trait that promotes lineage diversification via mechanisms that increase the rate of speciation and/or decrease the rate of extinction (e.g., Hodges and Arnold 1995). (Hereafter, ''diversification'' will be used to convey the net change in clade diversity as a result of speciation and extinction.) Changes in the rate of lineage diversification are expected to leave an imprint in the phylogeny of the affected species, underscoring the importance of historical inference in studies of evolutionary innovations.

Key innovation hypotheses are most frequently studied by correlation, which measures the degree to which multiple origins of a trait are associated with clades of unusual size. Correlations are appealing because they use phylogenetically independent datapoints, typically comparisons of sister clades, to demonstrate a general, repeatable effect on diversity by the candidate trait. Increasing the sample size in this way reduces the potential impact of other factors that affect diversification rate. A popular approach has been to compare the relative sizes of sister groups using nonparametric sign tests, for example, in studies of insect phytophagy (Mitter et

al. 1988), plant latex and resin canals (Farrell et al. 1991), floral nectar spurs (Hodges 1997), and flower symmetry (Sargent 2004). Alternatively, parametric approaches make it possible to identify unusually large clades by fitting data to predictions of null models of stochastic cladogenesis, and this can increase the statistical power associated with trait correlations (Slowinski and Guyer 1993; Sims and McConway 2003; McConway and Sims 2004).

In contrast, it is more difficult to evaluate unique origins of key innovations. If a trait arises only once, distinguishing its effect on diversification from the effects of other, coincidental, factors becomes problematic (see Discussion). Despite this, general methods for detecting shifts in diversification rate on phylogenies can be used to associate an inferred rate shift with the origin of the trait of interest. These use stochastic branching models to evaluate the observed data, either the distribution of clade sizes on a phylogenetic tree (e.g., Slowinski and Guyer 1989; Sanderson and Donoghue 1994; Sims and McConway 2003; McConway and Sims 2004; Moore et al. 2004), or the distribution of the number of lineages through time (e.g., Nee et al. 1992, 1994; Paradis 1997). The dichotomy of approaches has led to them being termed topological and temporal methods, respectively (Chan and Moore 2002).

Key innovation tests based on topological data are appealing for a number of reasons. They are intuitive in that shifts in diversification rate will eventually lead to disparities

257

in clade size. They are also pragmatic in generally requiring a minimum of taxon sampling in phylogenetic analyses, so long as all relevant lineages are represented and the sizes of those lineages are known. Furthermore, they can be applied when no information about branch lengths is available (Moore et al. 2004). However, methods for detecting rate shifts that rely solely on clade size data are not generally well-suited for cases in which the hypothesized key innovation has arisen recently, with insufficient time having passed for the shift in diversification rate to affect species diversity to an extent that is statistically detectable. In these situations, they lose inferential power as a result of small sample sizes. One exception to this is the so-called whole-tree approach (Chan and Moore 2002), which examines tree symmetry in the context of a Markovian branching model and can detect rate shifts even in small trees.

A potentially limiting aspect of all current key innovation tests is their reliance on unreversed synapomorphies, meaning they tend to assume that once arisen, key innovations are never lost. While this may be a safe assumption when a trait is diagnostic of an entire clade, it is certainly reasonable to expect cases in which key innovations are secondarily lost, especially considering that the trait's adaptive value may be contingent on a particular ecological setting or association that is transient (e.g., de Queiroz 2002). In theory, losses provide additional information relevant to the key innovation hypothesis, as they should be associated with decreases in diversification rate just as gains are expected to cause increases in rate. Thus, it is clearly desirable to have a phylogenetic method for testing key innovation hypotheses that accommodates bidirectional evolution in the trait of interest.

A related problem with existing tests concerns their tendency to rely on only the most parsimonious reconstructions of character change, that is, minimum-length solutions that are known to be biased (see Harvey and Pagel 1991) and that also ignore the variance of the estimate (Nielsen 2002). Historical inference invariably introduces some uncertainty into considerations of character evolution, the magnitude of which depends on how much information we have about the past. The increasing prevalence of Bayesian statistics and stochastic estimation methods is enabling uncertainty to be incorporated into many aspects of phylogenetic inference (e.g., see Huelsenbeck et al. 2000). Here, the same approach is applied to the study of key innovations.

In summary, existing phylogenetic tests of key innovation hypotheses are not ideal for studying singular instances, smaller clades, or traits that exhibit homoplasy. Key innovations are not conceptually precluded from such circumstances, but to detect them, focus is best directed not on the most commonly used metric, clade size contrasts, but instead on the correspondence between the evolutionary history of the character and the frequency of cladogenesis events (phylogenetic branching) through time. Lineages with the hypothesized innovation are expected to have shorter waiting times between cladogenesis events compared to lineages without it, and historical reconstructions of character evolution should therefore show higher rates of cladogenesis during the time in which the innovation was present. In this paper, I describe a simple test for this prediction that uses stochastic models of character evolution and cladogenesis in continuous time.

Testing Whether Character Evolution Covaries with Diversification Rate

The test requires an estimate of the phylogeny of the species of interest (i.e., those with the putative innovation) and their relatives. This can be a single tree, or it may be a set of trees representing plausible candidates (e.g., from bootstrap replicates or Bayesian samples from the posterior distribution of trees). For simplicity, consider first the case of a single tree. The terminal nodes on the tree are assumed to represent extant species. Because we are interested in the tempo of cladogenesis, the tree's branches should have lengths that are proportionate to time. A variety of statistical tools are available for estimating the temporal durations of branches from molecular sequence data, of which the simplest (and least often defensible) is to enforce the assumption of a molecular clock (Felsenstein 1985; Hasegawa et al. 1985). However, while branch lengths should represent relative temporal durations, they should be expressed in units of evolutionary change in the character of interest, such that the length of the tree is the total amount of change in the character expected over the whole phylogeny.

The steps involved in the test can be summarized as follows. First, the history of the character of interest is mapped on the tree, identifying partitions of the phylogeny over which the candidate trait was present and absent. Next, the number of branching events and relative time in each partition are used to estimate the difference in rates of diversification between character states. This difference in rates, averaged over alternative histories in proportion to their posterior probabilities, is taken as the test statistic. Finally, the significance of the test statistic is evaluated by generating its expected density under a null model of waiting times between cladogenesis events.

Consider the two hypothetical phylogenies in Figure 1. Each has an identical binary character of which state 1 is hypothesized to be a key innovation. The trees have the same topology, but differ in their branch lengths. By convention, the tree topology is denoted τ, the vector of branch lengths v, and the character data at the tips x. Given these data, the objective is to determine in each case whether a higher diversification rate is associated with the history of state 1, the candidate innovation, compared to state 0. Our first clue to the outcome is found in the relative durations of terminal branches having different states: in Figure 1A, terminals with state 1 are all much more recently split from their ancestor than those with state 0. This can be interpreted to mean that extant species with state 0 have apparently persisted much longer without undergoing speciation than those with state 1. In contrast, in Figure 1B there appears to be little difference between species with respect to character state.

The first step is to infer the evolutionary sequence of character state changes. This is accomplished using stochastic mapping (Nielsen 2002; Huelsenbeck et al. 2003), a technique that models discrete state changes in continuous time as a Poisson process, parameterized by a matrix of instantaneous rates of change between states ($\mathbf{Q}$), a vector of sta-

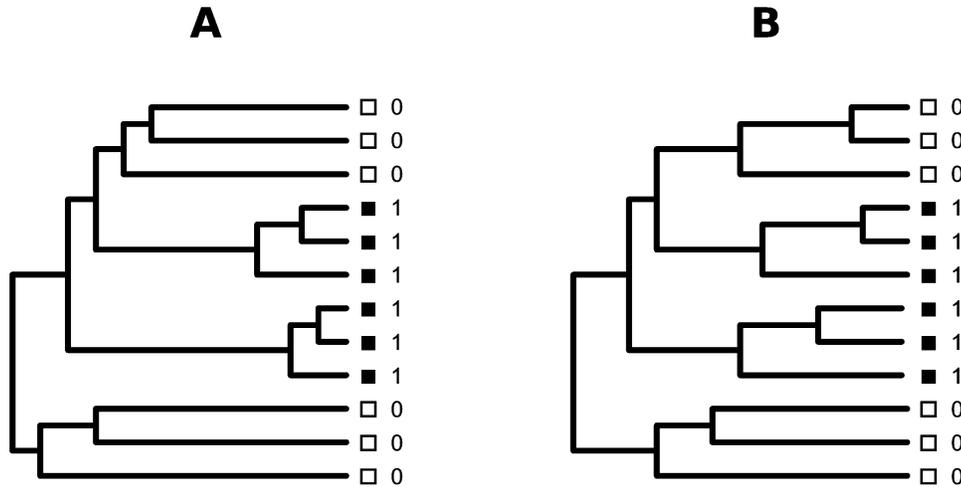**A**   **B**

FIG. 1. Hypothetical phylogenies with a candidate key innovation (black boxes, state 1). Absence of the trait, state 0, is denoted by open boxes. The relative durations of terminal branches suggest that the character covaries with diversification rate in (A) but not in (B).

tionary frequencies of states (π), and tree length, measured as the amount of expected change over the tree in the character of interest. Examples of character histories generated by stochastic mapping on the tree in Figure 1A are shown in Figure 2. Simulating many random histories consistent with the observed data is a Monte Carlo method for integrating over uncertainty about the historical details of character evolution (Nielsen 2002).

Each simulated mapping yields a history partitioned between the two character states. If the total tree length is s, let s_0 and $s_1 = s - s_0$ be the proportions of time spent in each state over a given history. A simple measure of the diversification rate associated with state i is $\lambda_i = n_i/s_i$, where n_i is the number of cladogenesis events observed in partition i of the history. The root node is not included in this calculation because we have no information about its history prior to the branching event it represents. The parameter of interest is the difference between state-specific diversification rates ($\lambda_1 - \lambda_0$), taking into account uncertainty associated with the tree, branch lengths, and inferences of character evolution. For a single tree, the average difference in rate over many draws from the posterior distribution of character mappings is

$$d(\tau, v, x) = \frac{1}{K} \sum_{k=1}^{K} \lambda_1^{(h_k)} - \lambda_0^{(h_k)}, \qquad (1)$$

where h_k is the kth simulated history given the topology, branch lengths, and character data. Here d is the test statistic.

Evaluating the significance of d involves estimating the probability of obtaining a difference in diversification rate from the observed character data equal to or greater than d if no covariation exists between the rate of cladogenesis and the character of interest. This requires estimating the distribution $E(d)$, the expected difference in diversification rate between character states under a null model of tree growth, in which waiting times between branching events on the tree topology are generated by a process independent of the character of interest. A simple model for this process is the well-known Yule model of cladogenesis (Yule 1924). Under the

Yule model, waiting times between branching events are exponentially distributed according to a rate parameter, λ. For a bifurcating tree of N species the maximum-likelihood estimate of λ is simply $(N - 2)/s$, where s is the tree length (Hey 1992; Nee 2001). Yang and Rannala (1997) provide a method for sampling and assigning node ages to a phylogeny under a generalized birth-death model, of which the Yule is a special case involving no extinction. This method is used to generate sets of branch lengths for the given tree topology with λ estimated from the observed tree and branch lengths. For each set of branch lengths (v^*) sampled from the null distribution, the average rate difference $d(\tau, v^*, x)$ is calculated as above from many simulated character histories. Ideally the same number of histories used in estimating d for the original set of branch lengths will be used here, but in practice a smaller number should suffice, as less precision is required. The expected distribution of d for the tree and character data under a Yule model is obtained by integrating over the null expectation of branch lengths, v^*:

$$E(d \mid \tau, \lambda, x) = \int_{v^*} d(\tau, v^*, x) \, dv^*, \qquad (2)$$

and the posterior predictive P-value for d is the proportion of $E(d)$ that is equal to or more extreme than d:

$$P = \Pr[E(d \mid \tau, \lambda, x) \geq d(\tau, v, x)]. \qquad (3)$$

Stochastic mapping was performed on the trees and character data in Figure 1. Uncertainty in the stationary frequencies of character states was modeled as a uniform prior distribution for π_0 over $\{0, 1\}$. The length of each tree was set equal to 2, the minimum number of changes required to explain the data at the tips. For these parameter values, 1000 simulations on each tree yielded test statistic values of 1.653 and 0.641, respectively. Figure 3 shows the null distributions $E(d \mid \tau, \lambda, x)$, which were generated by taking the average rate difference of 10 simulated character histories on each of 1000 sets of random branch lengths. In example A, $P = 0.031$, demonstrating a significant ($\alpha = 0.05$) historical relationship
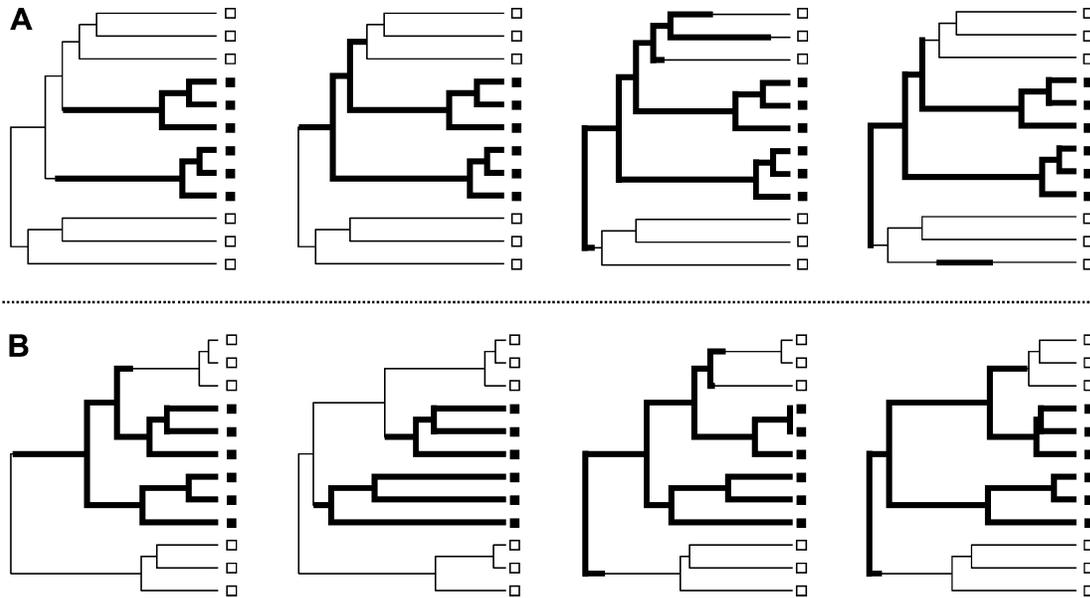
FIG. 2. Example character histories generated by stochastic mapping, with thick lines representing time spent in state 1 (black boxes). (A) Four simulated histories of the character on the tree shown in Figure 1A. (B) One history simulated on each of four different trees that are topologically identical to (A) but have random branch lengths as predicted by the Yule model of cladogenesis. Histories involving fewer character state changes are generated more frequently than those with more changes. For each history, the proposed test measures the state-specific diversification rate (λ_i) as the number of branching events associated with state i, divided by the amount of time spent in that state. Many histories such as those shown in (A) are used to obtain the the test statistic, d. To generate a null distribution of expected values of d, the procedure is repeated for many random sets of branch lengths and histories such as those shown in (B).

between the candidate trait and faster cladogenesis, whereas in example B, the result is nonsignificant ($P = 0.428$). Source code for running the test is available by contacting the author or via http://www.phylodiversity.net/rree.

These examples, while contrived, are illustrative in demonstrating that the method yields results that accord with intuition. They show that it is possible to detect the signature of a key innovation even on relatively small phylogenies (but note that for Fig. 1A, statistical significance is marginal, despite the apparently large effect of state 1 on diversification rate). They also show that the method can differentiate between trees that are topologically identical and have the same pattern of character data at the tips, but that differ in the effect of that character on diversification rate. These features distinguish this test from existing methods that rely only on comparisons of clade size.

Thus far, the test has been presented as it might be applied to a single phylogenetic tree. In real-world cases, it will generally be preferable to apply the test while accounting for uncertainty in the tree topology and branch lengths. One way to accomplish this is to use Markov chain Monte Carlo (MCMC) methods (Li 1996; Mau 1996; Rannala and Yang 1996; Yang and Rannala 1997; Mau et al. 1999; Huelsenbeck and Ronquist 2001; Drummond et al. 2002) to randomly draw a sample of trees from the posterior probability distribution. Next, calculate the test statistic and null distribution for each tree. The overall posterior predictive P-value is then simply the average of P over all the samples:

$$P = \frac{1}{K} \sum_{k=1}^{K} \Pr[E(d \mid \tau_k, \lambda_k, x) \geq d(\tau_k, v_k, x)]. \quad (4)$$

AN EMPIRICAL EXAMPLE:
FLORAL NECTAR SPURS IN *AQUILEGIA*

Perhaps the best-known example of a key innovation in plants is the floral nectar spurs of *Aquilegia* (columbines). In this angiosperm clade, nectar spurs have been hypothesized to increase the rate of diversification by promoting ecological speciation, characterized by reproductive isolation arising from pollinator specialization and shifts between hawkmoth and hummingbird pollinators (Hodges 1997). Compared to related taxa in Ranunculaceae, DNA sequence divergence in *Aquilegia* is extremely low, a pattern suggestive of a rapid radiation (Hodges and Arnold 1994a, 1995). There is considerable variation in spur length, color, and orientation among the approximately 70 species of *Aquilegia*, and functional studies in *Aquilegia* (Hodges and Arnold 1994b) and other groups (e.g., Nilsson 1988; Johnson and Steiner 1997) demonstrate the importance of spur characters in pollination and reproductive success. This has been put forth as evidence that speciation in *Aquilegia* has been driven by pollinator-mediated natural selection for divergence in floral traits (Hodges 1997).

Phylogenetic evidence that nectar spurs represent a key innovation in *Aquilegia* includes the observation that its sister group is a single species, *Semiaquilegia adoxoides* (Hodges and Arnold 1995). The method of Slowinski and Guyer (1993) applied to the 70:1 diversity ratio yields a P value of 0.014. Hodges and Arnold (1995) also applied Sanderson and Donoghue's (1994) method to the *Aquilegia* + *Semiaquilegia* clade, with the results ruling out a constant diversification rate as well as rejecting a shift in rate within one of the
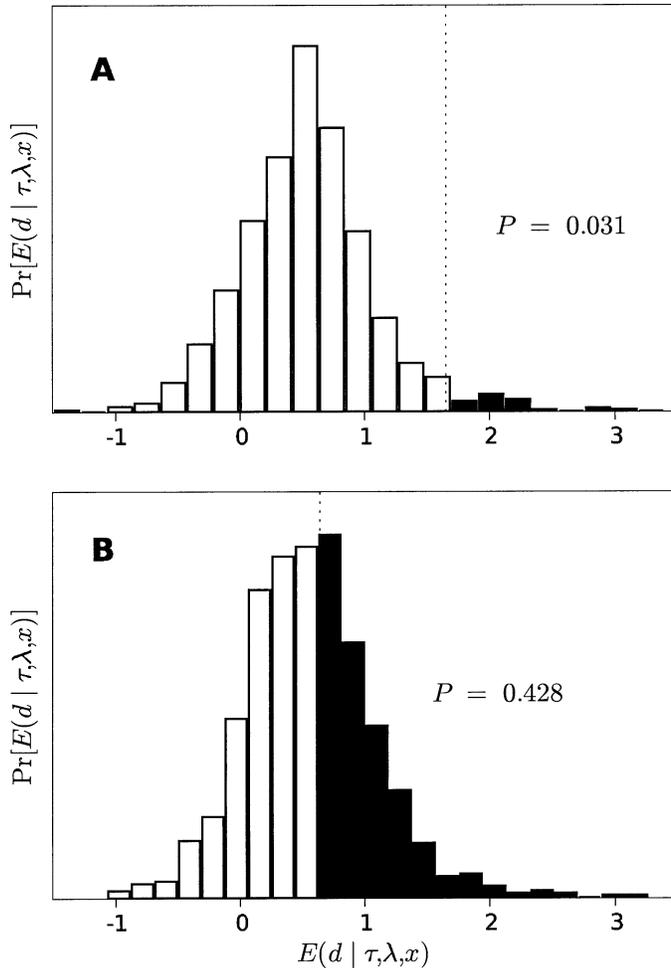
FIG. 3. Distribution of the difference in diversification rate between character states under a null model of cladogenesis, $E(d \mid \tau, \lambda, x)$, for the examples in Figure 1. The null model was used to randomly sample 1000 sets of branch lengths while holding the tree topology constant. For each set, the average difference in rate between states was calculated over 10 simulated character histories (see Fig. 2). The position of the test statistic (d) was calculated from 1000 simulated histories and is shown by the vertical dashed line. The associated posterior predictive P-value is represented by the proportion of the histogram to the right of d (filled bars). A significantly higher diversification rate is associated with state 1 for example A but not example B ($\alpha = 0.05$).

subclades of *Aquilegia*, that is, after the origin of spurs. However the latter analysis rested on the assumption of a basal split in *Aquilegia* of 47:23 species (Hodges and Arnold 1995), a relationship for which the DNA sequence data offer no statistical support.

The most striking pattern in the *Aquilegia* phylogeny from Hodges and Arnold (1995) is not in the topology of the tree, but in the low level of sequence divergence (and hence rapid cladogenesis) within *Aquilegia* relative to its outgroups. This is reflected in a null-model-based analysis of the dataset of Hodges and Arnold (1995), in which branching events through time reveal a recent upswing in diversification rate (Wollenberg et al. 1996; but see Paradis 1997). I evaluated the phylogenetic evidence for the nectar spur hypothesis with the test proposed here, using the nuclear ribosomal ITS da-

taset of Hodges and Arnold (1994a) that included 14 species of *Aquilegia*, to which I added the publicly available ITS sequence for the nonspurred species *A. ecalcarata* (GenBank: U75657). The ITS sequence of *S. adoxoides* from Hodges and Arnold (1995) was unfortunately not available for reanalysis. I excluded from the dataset the more distant relatives of *Aquilegia*, as these were sampled very sparsely, but retained the closest relatives, species of *Isopyrum*. All of the four extant species of *Isopyrum* are included in the dataset. The computer program BEAST (Drummond and Rambaut 2003) was used to conduct an MCMC analysis (Drummond et al. 2002) of the alignment for 1×10^6 generations, of which the first 1×10^5 were discarded as burn-in (the initial period of sampling during which the Markov chain is not at stationarity, i.e., stably sampling parameters from their posterior distributions). The analysis employed the HKY85 nucleotide substitution model (Hasegawa et al. 1985) with a molecular clock enforced. After apparent chain stationarity, trees were sampled every 100 generations, and for computational load considerations in the next step, this set was then thinned to 3000 trees (Fig. 4). The branch lengths of all trees were scaled to yield a total tree length of 2. The proposed key innovation test was applied using this sample of trees, with a uniform prior distribution for π_0, and the result was highly significant ($P = 0.0006$; Fig. 5).

In this analysis, the level of support for the nectar spur hypothesis is important to note, because it illustrates that incorporating branch length information reveals a much stronger signal of a key innovation than was previously detected by tests that focus on diversity contrasts. It is highly unlikely that the missing *Semiaquilegia* sequence would have had much impact on the result, especially considering that only 15 of the approximately 70 species of *Aquilegia* were included in the analysis. Taxon sampling in this case is biased, with spurred species underrepresented; the bias therefore favors rejection of the key innovation hypothesis, assuming that the *Aquilegia* species not sampled have similar levels of sequence divergence. In other words, supposing there were only 15 species of *Aquilegia*, the data would still indicate a strong signal of disparate diversification between spurred and nonspurred lineages. That the clade is in fact much larger means that the P-value obtained in this analysis is certainly an overestimate.

The *Aquilegia* example highlights another difference between the proposed test and comparisons of clade size, namely their sensitivities to the precision of taxon sampling. Before the discovery that the monotypic group *Semiaquilegia* is the sister group to *Aquilegia* (Hodges and Arnold 1995), tests based on clade diversity comparisons were unable to detect a significant pattern consistent with the key innovation hypothesis. Here, even without data on the position of *Semiaquilegia* (i.e., without knowing about the 70:1 diversity contrast between the sister groups) there is still an extremely strong signal consistent with the key innovation hypothesis. Tests relying on comparisons of clade size may generally require less taxon sampling overall than this test, so long as the standing diversities of the included terminal taxa are known. However, this analysis of *Aquilegia* makes clear that clade size comparisons also require greater precision in knowledge about the phylogeny: for sister group comparisons
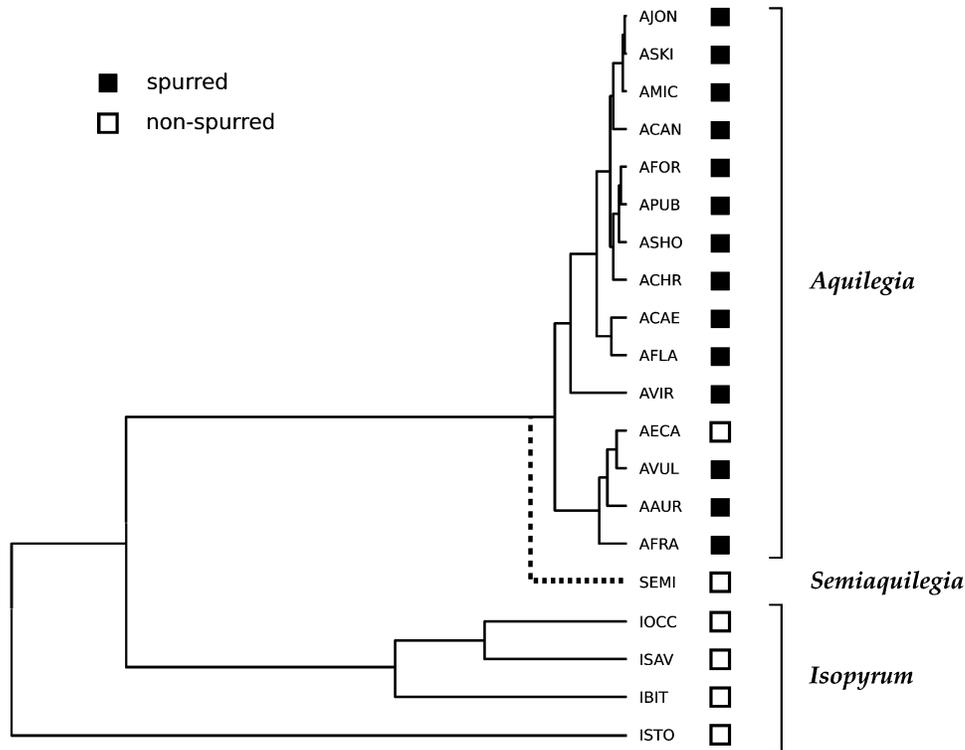
FIG. 4. ITS phylogeny of *Aquilegia* and *Isopyrum* showing one of approximately 10,000 trees sampled by Markov chain Monte Carlo under the assumption of a molecular clock. The hypothesized key innovation, floral nectar spurs, is designated by black boxes. The approximate position of *Semiaquilegia*, not included in this analysis but found by Hodges and Arnold (1995) to be the sister group of *Aquilegia*, is shown by the dotted line.
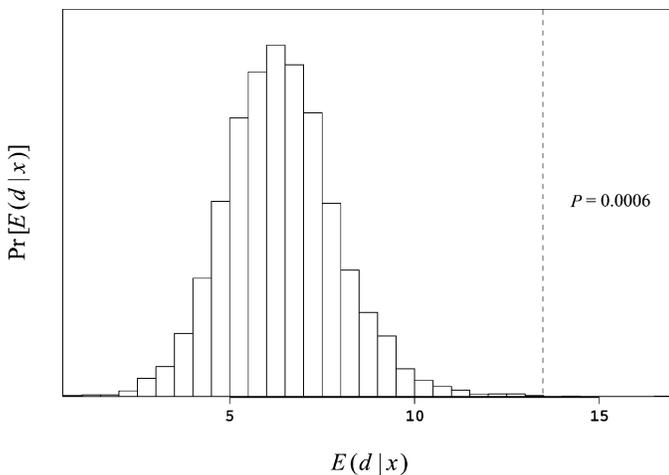

FIG. 5. Distribution of the difference in diversification rate between character states under a null model of cladogenesis, $E(d \mid x)$, for the *Aquilegia* dataset (Fig. 4) showing the test statistic (d) as a dashed line and the posterior predictive P-value. In this case d and $E(d)$ were calculated over 3000 trees sampled from the posterior distribution by Markov chain Monte Carlo. For each tree, d was calculated from 1000 simulated histories and $E(d)$ from 1000 random sets of branch lengths, with ten histories per set. Allowing for uncertainty in the phylogeny in this manner, a significantly higher diversification rate is associated with floral nectar spurs.

to be useful, one must be fairly certain about what the sister group relationships are. This does not simply imply the need for strong statistical support for clades; more importantly, it highlights the need for adequate taxon sampling. This requirement is relaxed for the method proposed here, insofar as the taxa sampled yield representative state-specific diversification rates (see below).

DISCUSSION

Existing phylogenetic tests of key innovation hypotheses generally focus on differences in species diversity and only indirectly associate those differences with the hypothesized innovation. In contrast, this test directly associates estimates of the rate of cladogenesis with the evolutionary history of the candidate trait, and it can be applied in cases where diversity-based tests cannot, such as the hypothetical example (Fig. 1). Tests comparing clade sizes fail in that case because lineages with the innovation are not more species-rich, either overall or in any sister-pair comparison not involving homoplasy.

In this paper I focus attention on a direct consequence of higher rates of diversification, namely shorter waiting times between cladogenesis events, that may result from biological innovations. It is not the only consequence: tree symmetry, or nodal balance, is also affected by diversification rate (Chan and Moore 2002). In any event, shifts in rate eventually lead to significant differences in species diversity between lineages with the innovation and lineages without it. Disparate

clade sizes could thus be considered a secondary phylogenetic effect that becomes increasingly pronounced with the length of time that lineages possess the innovation.

In general, the historical signature of a key innovation will have both a time component and a diversity component (Sanderson and Donoghue 1996), but the relative contribution of each to the overall signal will depend on how long ago the key innovation originated. Tests that focus on clade size comparisons are less likely to succeed in detecting recent key innovations than the test proposed here, because for recent shifts in diversification rate, the effect on waiting times will be stronger than on clade sizes; insufficient time will have passed for the rate differential to translate into clade size disparity. Conversely, this test may be less useful than clade size comparisons in detecting ancient key innovations, where enough time has passed to yield a predominant signal of larger clades being associated with the key innovation. This is partly due to the practical difficulty imposed by large clades of obtaining a sufficient density of taxon sampling and molecular sequence data to apply the proposed test. In contrast, diversity comparisons generally require relatively little phylogenetic information other than the relationships and sizes of clades in the vicinity of the origin(s) of the putative innovation. Moreover, as the time scale over which inferences are drawn increases, one expects greater noise in the data from factors that influence diversification rate but are unrelated to the trait of interest (e.g., mass extinctions, other innovations, colonization of new areas). For example, in *Halenia*, another angiosperm clade with nectar spurs, biogeographic events such as dispersal to South America appear to have had a more substantial impact on the tempo of diversification than the possession of spurs (von Hagen and Kadereit 2003). Over long time scales, these extraneous variables may disproportionately reduce the detectable signal in waiting times compared to the signal in clade sizes, but this is an empirical question that remains to be addressed.

Because this method focuses solely on branch lengths, statistical inferences are likely to be conservative, because only part of the total information available is used. A more comprehensive method would incorporate information on both waiting times and tree topology. Ideally, it would also allow for homoplasy in the candidate trait, and accommodate uncertainty in all historical inferences (e.g., tree topology, branch lengths); in this regard the method proposed here is perhaps a step in the right direction. A potentially fruitful pursuit might be to incorporate key innovation hypotheses into the phylogeny estimation process itself, for example, by using models that allow state-specific waiting times between cladogenesis events in MCMC proposal functions for trees.

Another reason why this test might tend to be conservative relates to the reliance on estimates of phylogenetic branch lengths that are proportionate to time. Obtaining such information from extant species commonly involves combining molecular sequence data with temporal constraints on divergence times, as can be provided by fossils or prior knowledge about historical biogeographic events. However, even with reliable calibrations, the separation of evolutionary rate and time in phylogenetic inference continues to be a difficult statistical problem (Thorne and Kishino 2002). If rates of molecular evolution and lineage diversification are positively

correlated, as found in some studies (e.g., Barraclough and Savolainen 2001; Webster et al. 2003), then temporal durations of branches could be systematically overestimated for lineages possessing a key innovation, despite statistical corrections. The power of this test to detect the key innovation would be consequently reduced.

An important question is whether the Yule model of cladogenesis is appropriate for generating predictions about the distribution of waiting times between lineage branching events. A potential liability of using the Yule pure-birth process springs from the assumptions of complete taxon sampling and no extinction. These are indeed stringent assumptions, but I suspect that to the extent that the effects of incomplete, nonrandom taxon sampling and heterogeneous extinction rates are unbiased with respect to the candidate trait, the test itself should be similarly unbiased. I also predict that the predominant effect of incomplete, nonrandom sampling and variable extinction should be a reduction in power of the test. However, a thorough and systematic analysis of the effects of these variables on the performance of this method is clearly needed, particularly to clarify when the risk of Type II error is exacerbated. Additional work on how more complex models of cladogenesis, that is, those that include birth, death, and sampling parameters (Yang and Rannala 1997), could be used to generate more realistic predictions about branch lengths is also a priority.

In considering appropriate models for the null distribution and other issues of statistical power and performance, it is worthwhile to also consider possible alternative formulations of the test itself. As presently structured, the method makes inferences by holding constant the character data, tree topology, and tree length, and generates the expected distribution of d from a null model for branch lengths. Another approach might hold the observed branch lengths constant, but allow the character data to vary, and generate $E(d)$ by simulating histories on the observed tree that are not constrained to match the observed character pattern. Alternative formulations of the basic approach outlined in this paper may differ in statistical behavior such that one is preferable over the others under particular circumstances, making this an interesting avenue for further study.

As for all inferences of evolutionary rate from phylogenies, results from this method condition on the bounds of the phylogenetic neighborhood sampled. For example, support for nectar spurs as a key innovation in *Aquilegia* might be different if the phylogeny were expanded to include *Thalictrum*, a closely related clade with 330 species, because those data would influence estimates of the nonspurred diversification rate. A question that naturally arises is how much of the phylogenetic neighborhood surrounding a key innovation is best to include in an analysis. A reasonable goal at the outset might be to sample enough clades to equalize the representation of each state with respect to the average number of inferred speciation events, the sum of reconstructed branch lengths, or the number of terminal taxa.

A common complaint about key innovation hypotheses is that historical inferences cannot distinguish among phylogenetically congruent traits as to which one is the real key innovation responsible for the observed pattern. For example, in *Aquilegia*, how do we know that it is actually nectar spurs,

and not some other trait shared by *Aquilegia* species, that is driving the result? A reasonably cautious view is that key innovation hypotheses are best judged from different angles, with phylogenetic tests like the one proposed here representing just one line of evidence that can be brought to bear on biological causes of shifts in diversification rate (e.g., Heard and Hauser 1995). It is important to remember that phylogenetic tests are generally not the best tools for evaluating hypotheses about mechanisms. Distinguishing between alternative traits that show the same phylogenetic pattern requires experiments and other tests of the mechanisms by which each trait is hypothesized to bring about higher diversification rates. However, as a corollary to the above, if alternative traits are not consistently correlated, support for a particular trait being a key innovation in a given case is strengthened if it can be shown to be generally true, that is, it has the same effect in other, phylogenetically independent, instances. This is the rationale for using correlations to test key innovation hypotheses.

In correlations, each independent origin of the trait is treated as a single data point. What are the datapoints used by this test? Because it allows homoplasy, a single analysis may include multiple origins (and losses) of the key innovation. If homoplasy is high, one could argue that a single analysis using this test is more similar to a correlation study than to a single observation (like the 70:1 ratio in *Aquilegia + Semiaquilegia*). However, the issue is not entirely clear, and the fact remains that a single analysis yields inferences from only one phylogenetic neighborhood. If general conclusions about a trait are to be drawn, multiple inferences from widely separated branches of the tree of life may be required.

The expected generality of key innovations is itself an issue that deserves critical appraisal. Should we expect a trait to have the same effect on diversification rate each and every time it evolves? One could argue that key innovations will tend to be highly contextual, requiring specific biological correlates or ecological preconditions to influence lineage diversification (e.g., de Queiroz 2002). If this is true, it is not so imperative to examine multiple cases of trait origin if they differ in this regard. At the very least, it encourages more detailed examination of each case brought forth in support of a general conclusion.

Stochastic models of character evolution and cladogenesis allow aspects of the evolutionary process to be inferred in the face of uncertainty about the details of past events. Here I demonstrate their use in detecting differences in diversification rate between character states, a fundamental parameter of key innovation hypotheses. This contrasts with other methods that only indirectly associate inferred shifts in diversification rate with origin of the character of interest. This method complements those methods and hopefully broadens the range of cases in which key innovation hypotheses may be tested in a phylogenetic context.

ACKNOWLEDGMENTS

LITERATURE CITED

Barraclough, T. G., and V. Savolainen. 2001. Evolutionary rates and species diversity in flowering plants. Evolution 55:677–683.

Chan, K. M. A., and B. R. Moore. 2002. Whole-tree methods for detecting differential diversification rates. Syst. Biol. 51: 855–865.

de Queiroz, A. 2002. Contingent predictability in evolution: key traits and diversification. Syst. Biol. 51:917–929.

Drummond, A. J., and A. Rambaut. 2003. BEAST Ver. 1.0. Available from http://evolve.zoo.ox.ac.uk/beast/.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Farrell, B. D., D. E. Dussourd, and C. Mitter. 1991. Escalation of plant defense: Do latex and resin canals spur plant diversification? Am. Nat. 138:881–900.

Felsenstein, J. 1985. Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34:152–161.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford Univ. Press, Oxford, U.K.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.

Heard, S. B., and D. L. Hauser. 1995. Key evolutionary innovations and their ecological mechanisms. Hist. Biol. 10:151–173.

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46:627–640.

Hodges, S. A. 1997. Floral nectar spurs and diversification. International J. Plant Sci. 158(Suppl. 6):S81–S88.

Hodges, S. A., and M. L. Arnold. 1994a. Columbines: a geographically widespread species flock. Proc. Natl. Acad. Sci. USA 91: 5129–5132.

———. 1994b. Floral and ecological isolation between *Aquilegia formosa* and *Aquilegia pubescens*. Proc. Natl. Acad. Sci. USA 91:2493–2496.

———. 1995. Spurring plant diversification: Are floral nectar spurs a key innovation? Proc. R. Soc. Lond. 262:343–348.

Huelsenbeck, J. P., and F. R. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:745–755.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288:2349–2350.

Huelsenbeck, J., R. Nielsen, and J. Bollback. 2003. Stochastic mapping of morphological characters. Syst. Biol. 52:131–158.

Johnson, S. D., and K. E. Steiner. 1997. Long-tongued fly pollination and the evolution of floral spur length in the *Disa draconis* complex (Orchidaceae). Evolution 51:45–53.

Li, S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. thesis, Ohio State Univ., Columbus, OH.

Mau, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. thesis, Univ. of Wisconsin, Madison, WI.

Mau, B., M. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

McConway, K. J., and H. J. Sims. 2004. A likelihood-based method for testing for nonstochastic variation of diversification rates in phylogenies. Evolution 58:12–23.

Mitter, C., B. Farrell, and B. Wiegmann. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? Am. Nat. 132:107–128.

Moore, B. R., K. M. A. Chan, and M. J. Donoghue. 2004. Detecting diversification rate variation in supertrees. Pp. 487–533 *in* O. R. P. Bininda-Emons, ed. Phylogenetic supertrees: combining information to reveal the tree of life. Kluwer Academic, Dordrecht, The Netherlands.

Nee, S. 2001. Inferring speciation rates from phylogenies. Evolution 55:661–668.

Nee, S., A. O. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. USA 89:8322–8326.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. B 344: 305–311.

Nielsen, R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51:729–739.

Nilsson, L. A. 1988. The evolution of flowers with deep corolla tubes. Nature 334:147–149.

Paradis, E. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. Proc. R. Soc. Lond. B 264:1141–1147.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Sanderson, M. J., and M. J. Donoghue. 1994. Shifts in diversification rate with the origin of angiosperms. Science 264: 1590–1593.

———. 1996. Reconstructing shifts in diversification on phylogenetic trees. Trends Ecol. Evol. 11:15–20.

Sargent, R. D. 2004. Floral symmetry affects speciation rates in angiosperms. Proc. R. Soc. Lond. B 271:603–608.

Sims, H. J., and K. J. McConway. 2003. Nonstochastic variation of species-level diversification rates within angiosperms. Evolution 57:460–479.

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. Am. Nat. 134(6):907–921.

———. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. Am. Nat. 142:1019–1024.

Thorne, J., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51(5):689–702.

von Hagen, K. B., and J. W. Kadereit. 2003. The diversification of *Halenia* (Gentianaceae): ecological opportunity versus key innovation. Evolution 57:2507–2518.

Webster, A. J., R. J. H. Payne, and M. Pagel. 2003. Molecular phylogenies link rates of evolution and speciation. Science 301: 478.

Wollenberg, K., J. Arnold, and J. C. Avise. 1996. Recognizing the forest for the trees: testing temporal patterns of cladogenesis using a null model of stochastic diversification. Mol. Biol. Evol. 13:833–849.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of J. C. Willis, F. R. S. Proc. R. Soc. Lond. B 213: 21–87.

Corresponding Editor: A. Yoder