doi:10.1111/evo.12256

QUANTIFYING VARIATION IN SPECIATION AND EXTINCTION RATES WITH CLADE DATA

Emmanuel Paradis,[1,2] Pablo A. Tedesco,[3] and Bernard Hugueny[3]

[1]*Institut de Recherche pour le Développement, ISEM UMR 226/5554 – UM2/CNRS/IRD, Jl. Taman Kemang 32B, Jakarta 12730, Indonesia*

[2]*E-mail: Emmanuel.Paradis@ird.fr*

[3]*UMR Biologie des ORganismes et des Écosystémes Aquatiques (UMR BOREA, IRD 207-CNRS 7208-UPMC-MNHN), Département Milieux et Peuplements Aquatiques, Muséum National d'Histoire Naturelle, 43 rue Cuvier 75231, Paris, cedex, France*

Received June 1, 2013
Accepted August 14, 2013

High-level phylogenies are very common in evolutionary analyses, although they are often treated as incomplete data. Here, we provide statistical tools to analyze what we name "clade data," which are the ages of clades together with their numbers of species. We develop a general approach for the statistical modeling of variation in speciation and extinction rates, including temporal variation, unknown variation, and linear and nonlinear modeling. We show how this approach can be generalized to a wide range of situations, including testing the effects of life-history traits and environmental variables on diversification rates. We report the results of an extensive simulation study to assess the performance of some statistical tests presented here as well as of the estimators of speciation and extinction rates. These latter results suggest the possibility to estimate correctly extinction rate in the absence of fossils. An example with data on fish is presented.

KEY WORDS: **Birth–death models, extinction, maximum likelihood, speciation, stem ages.**

The study of the tempo and mode of evolution has experienced a new wave of interest from evolutionists using new mathematical and statistical tools to analyze molecular phylogenies (Sanderson and Donoghue 1996; Ricklefs 2007). Following some initial breakthrough (e.g., Nee et al. 1992, 1994), significant progress has been achieved in biologically relevant statistical modeling of diversification, such as quantifying temporal variation in diversification (Paradis 2011; Hallinan 2012) or assessing the effects of biological traits on speciation and extinction rates (Maddison et al. 2007; FitzJohn et al. 2009; FitzJohn 2010). Recent advances have also been accomplished in integrating molecular and fossil data (e.g., Morlon et al. 2011; Didier et al. 2012).

Most of these recent statistical developments have focused on analyzing complete phylogenies. Incomplete phylogenies are often treated as a separate case to take missing data into account (Pybus et al. 2002; FitzJohn et al. 2009; Stadler 2011). The most common form of such data is a phylogeny resolved at a high level accompanied by the number of species associated to each tip of the tree. On the other hand, the ages of clades together with the numbers of species (named here "clade data") have been a neglected source of data in the analysis of diversification. Magallón and Sanderson (2001) provided some methods for the analysis of such data and applied them to angiosperms. They particularly developed various estimators of the (net) rate of diversification of a clade giving its age and number of species.

The relative lack of interest toward clade data may come from the fact that, for a given clade, its complete phylogeny contains more information than the pair of values "age + number of species." However, for a collection of clades, such data are a valuable source of information for several reasons. First, clades defined by higher-level taxa (e.g., families, orders) are clearly identified for almost all groups of living beings and their numbers of species are in many cases already known. Second, phylogenetic relationships among higher-level taxa have been much more

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

studied than within them, so it is more straightforward to date the age of a clade rather than the divergences among its species. Third, the fossil record is generally more informative on the origin of higher-level taxa compared to species or other low-level taxa. Fourth, it is easier to examine the impact of the species concept on the definition of clade data rather than on a phylogeny because in the former the species concept will mostly affect the number of species, whereas in the latter it will be often hard to infer different phylogenies under those distinct species definitions. Clade data have also some disadvantages: the inherent lack of temporal resolution within each clade makes it impossible to study the variation in diversification within them.

In the present article, we extend the approach presented by Magallón and Sanderson (2001) and present statistical tools for the inference of diversification patterns and processes with clade data. Our approach assumes that each clade, instead of having its own speciation and extinction rates, comes from a "statistical population of clades" so that maximum likelihood inference is straightforward. With this rationale, we show how to make inference on variation in diversification parameters among clades using different modeling tools, including testing the effects of life-history traits and environmental variables and the case where variation is a priori unknown. We also present the results of a simulation study to assess the statistical performance of several tests and estimators presented in this article, and finally, we apply our approach on a dataset of fish.

Statistical Modeling Approach

Throughout this article, we assume that diversification proceeds with speciation (λ) and extinction (μ) rates, which are the probabilities that a species splits into two daughter species or goes extinct during a very short time. We denote as X_t the number of species in a clade of age t where this may be either the stem age of the clade (divergence time of the clade from its sister-clade) or its crown age (time to the most recent common ancestor of the species belonging to the clade). Specifically, using equation (8) from Kendall (1948), we can write the probability that X_t takes a specific integer value x:

$$\Pr(X_t = x|\theta, X_0 = 1) = \eta_t(1-\eta_t)^{x-1} \qquad x \geq 1, \tag{1}$$

where θ is a vector of parameters specifying how speciation and extinction rates vary through time and η_t is a function of these parameters. The conditioning on $X_0 = 1$ emphasizes that in this article we consider stem groups. For the case of crown groups ($X_0 = 2$), the probabilities must be summed on all possible combinations. In most applications, stem groups are considered because the origin of a group is inferred from its relationships with its sister group. On the one hand, deriving the crown age of a group requires to estimate the age of the most recent common ancestor of its species, which is usually more complicated because it requires to sample all species in the clade. On the other hand, inferring stem ages requires one species from the clade and one from its sister-clade.

Various forms exist for these probabilities depending on the parameterization of θ and whether we wish to condition them on survival of the lineage until present or not. For instance, if extinction rate is zero and speciation rate is constant, then $\eta_t = e^{-\lambda t}$. This is the Yule (1924) model. Models with a nonnull extinction rate are called birth–death models (Kendall 1948).

The point of conditioning on no extinction is important when analyzing data on actual groups because total extinction of these groups did not occur. Thus, the probabilities must be modified accordingly, otherwise this would result in underestimated extinction rates (Rabosky et al. 2007).

Let us consider for the moment the simple Yule model. The expected number of species at time t is given by $E(X_t) = e^{\lambda t}$. From this expectation, a simple estimator of λ based on the method of moments is $\hat{\lambda} = \ln(x)/t$ (Magallón and Sanderson 2001). When considering a single clade, and in the absence of more detailed information, it does not seem possible to go further in the inference. When considering more than one group (e.g., the families within an order or a class), researchers usually estimate λ separately for each group, then proceed with standard statistics (e.g., McPeek 2008). This approach assumes that each clade is characterized by its own speciation rate. On the other extreme, one may assume that speciation rate is the same in all groups so that the observed data are independent outcomes of the same diversification process. Thus, it is possible to use maximum likelihood inference using equation (1). The likelihood function is

$$\prod_i \Pr(x_i|\lambda), \tag{2}$$

where $\Pr(x|\lambda)$ is a simplified notation of equation (1). We may expect less bias in the estimates from this approach, but also the possibility to test hypotheses based on fitting alternative models.

The assumption of equal speciation rates among clades is, certainly in most cases, unrealistic (Purvis et al. 1995; Paradis 2005; Alfaro et al. 2009). However, because we have several observations we may model the variation in this parameter with a statistical modeling approach. We explore several such approaches below. First, we consider approaches based on deterministic variation between two or more groups of clades. Second, we consider how temporal variation in speciation and extinction rates can be modeled and assessed. Third, we develop an approach handling unknown variation based on mixture modeling, including the combination of mixtures with a linear modeling of the

speciation rate. Finally, we attack the problem of estimating extinction rates.

VARIATION AMONG CLADES

A simple way to model variation in diversification among clades is to assume that there are two categories: some clades diversify with speciation rate λ_1 and the others with rate λ_2. The data are made of n_1 and n_2 clades in each category, respectively. The likelihood function is

$$\prod_{i_1=1}^{n_1} \Pr(x_{i_1}|\lambda_1) \prod_{i_2=1}^{n_2} \Pr(x_{i_2}|\lambda_2).$$

Note that each clade is assigned to a category a priori, although there is no assumption on whether λ_1 is greater, or smaller, than λ_2. The null hypothesis $\lambda_1 = \lambda_2$ can be tested by fitting this model and the null model whose likelihood is given by equation (2): the likelihood ratio test (LRT) comparing these two models follows a χ^2 distribution with df $= 1$. An alternative is to use the Akaike information criterion (AIC, Akaike 1973).

The present approach is easily generalized to more than two categories: let us denote the number of categories as K, then the likelihood function would become the product of K products:

$$\prod_{j=1}^{K} \prod_{i_j=1}^{n_j} \Pr(x_{i_j}|\lambda_j),$$

where n_j is the number of clades in the jth category. The LRT comparing this model with the null model of homogeneous diversification follows a χ^2 with df $= K - 1$.

These models assume, mostly for simplicity, that there is no extinction ($\mu = 0$); however, variation in extinction rate can be incorporated in a straightforward way. For instance, a model with two categories diversifying with the same λ but with different extinction rates has the following likelihood function:

$$\prod^{n_1} \Pr(x_{i_1}|\lambda, \mu_1) \prod^{n_2} \Pr(x_{i_2}|\lambda, \mu_2),$$

which could be compared with the null model with $\mu > 0$ whose likelihood is:

$$\prod^{N} \Pr(x_i|\lambda, \mu),$$

with $N = n_1 + n_2$. This test is related, but not identical, to the tests of equal diversification using sister-clades where the ages of clades are not needed (Paradis 2012b).

The Supporting Information provides annotated R code explaining how to build and fit any model following the present approach.

LINEAR MODELING

Following the previous section, two extreme models can be defined: the simplest one where all clades diversify at the same rate, and the most complex one where each clade has its own parameter(s). This second model will be overparameterized for a likelihood approach. Nevertheless, it is possible to model variation in diversification parameters with linear models. For instance, we may know a priori some variables that are likely to affect the value of speciation rate (e.g., body size), and a model that relates such "covariates" to speciation rate may be an appropriate candidate to model the variation in diversification among clades. We use here a standard strategy to model variation in a rate with respect to a covariate z

$$g(\lambda_i) = \beta z_i + \alpha,$$

where λ_i is the speciation rate in clade i, g is a function used to transform the rate to linearize the relationship, and β and α are two parameters. Here β controls the effect of z on λ: if $\beta > 0$, then species with large values of z will speciate faster than those with small values of z (and inversely if $\beta < 0$). It is possible to consider more than one predictor in which case the number of parameters is equal to the number of predictors plus one. Nonlinear models can also be considered. Each clade has its own speciation rate given by (with g^{-1} being the inverse transformation of g):

$$\lambda_i = g^{-1}(\beta z_i + \alpha), \tag{3}$$

which is used to calculate the likelihood defined by equation (2): the likelihood function is then maximized to estimate β and α (see code in the Supporting Information). A common choice for g is the logit function, $\ln(\lambda_i/(1-\lambda_i))$, so g^{-1} gives

$$\lambda_i = \frac{1}{1 + e^{-(z_i\beta+\alpha)}},$$

The null model is defined by fixing $\beta = 0$ in which case $\lambda = 1/(1 + e^{-\alpha})$ for all clades. The logit function is well suited for parameters varying between 0 and 1, which is the case for speciation rates considered on geological time scales (million of years). However, speciation rates may be larger than one on shorter scales. Other transformations can be used such as the one used below.

It must be noted that the variation among clades as modeled in the previous section is a special case of linear models where the membership of a clade to a category is coded with a discrete variable and this variable is entered as a predictor into the linear model after coding it into binary 0 and 1 variable(s) (see appendix in Paradis 2005, for details). Therefore, continuous and categorical predictors can be combined in the linear model.

TEMPORAL VARIATION

Kendall (1948) studied the birth–death model in a very general way, including the cases where λ and μ vary through time. Thus,

it is possible to derive the probability density of the distribution of the x_i's when diversification changed through time. The likelihood can be defined and fit in the same way as above. Such a temporal model can be compared with the null model of constant diversification with a χ^2 test whose df will be equal to the number of additional parameters in the first model. As before, temporal variation may reflect speciation and/or extinction rate(s). The simplest temporal model has two rates before and after a given time point in the past, so it has one additional parameter than the null model. Note that if the time point is unknown, it could be estimated from the data so there would be two additional parameters. However, a wide variety of temporal models can be defined in ape (Paradis et al. 2004) using the function dbdTime where the temporal variation is defined by the user with a standard R function.

UNKNOWN VARIATION

The above models assume that diversification parameters vary in relation to some known variables, either categorical or continuous. On the other hand, it is possible that these variables are not observable. Such unknown variation can be modeled with two approaches depending on whether we assume that the diversification parameters vary in a discrete or continuous manner.

A mixture of distributions is based on the assumption that observations come from two or more categories each characterized by its own distribution, but the assignment of an observation to a particular category is unknown (see Flury et al. 1992, for a biological example). As a simple example, consider a mixture of two Yule processes, then the likelihood function will be

$$\prod_{i=1}^{N} f \Pr(x_i|\lambda_1) + (1 - f) \Pr(x_i|\lambda_2), \tag{4}$$

where f is the proportion of clades in the first category. This model has three parameters (λ_1, λ_2, and f) and can be compared with the null model of homogeneous speciation with an LRT with $\mathrm{df} = 2$. The idea is easily generalized to more than two mixtures: a mixture with K Yule models would have $2K - 1$ parameters. As above, the mixture may involve speciation and/or extinction rate(s). In contrast to the situation above where clades were assigned to categories a priori, there is no assignment a priori. On the other hand, assignment a posteriori is possible by calculating the relative contributions to the likelihood function.

The idea may even be further generalized to include mixtures of linear models. Suppose, we know that one variable, say body size, has a significant effect on speciation rate, but there is some other, unknown, variation in this parameter that we want to model with a mixture. Then, it is possible to calculate the λ_i's with equation (3) and use them to compute the likelihood with equation (4). Each category would have its own parameters β and α, so a model with K categories has $3K - 1$ parameters.

The second approach assumes that, in the case of a Yule model, λ varies continuously across clades following a specified distribution whose parameters are estimated from the data. A transformation of λ is useful so that it follows a normal distribution: $g(\lambda) \sim \mathcal{N}(\mu_\lambda, \sigma^2_\lambda)$. A useful transformation here is the complementary log–log transformation: $g(\lambda) = \ln(-\ln(\lambda))$. As above, we do not know the value of λ for a given clade, but this time instead of a discrete sum, we have to do a continuous integration. The likelihood function is thus:

$$\prod_{i=1}^{N} \int_{-\infty}^{\infty} f_{\mathcal{N}}\left(u|\mu_\lambda, \sigma^2_\lambda\right) \Pr(x_i|g^{-1}(u))\mathrm{d}u,$$

where $f_{\mathcal{N}}$ is the density function of the normal distribution. A graphical representation of the variation in λ is obtained with the inverse transformation $g^{-1}(u) = \exp(-\mathrm{e}^u)$ with the dentity of u computed with the normal distribution and the estimates $\hat{\mu}_\lambda$ and $\hat{\sigma}^2_\lambda$.

ESTIMATING EXTINCTION RATES

The estimation of extinction rates in the absence of fossil data has appeared to be a complicated issue (Paradis 2004, 2011; McPeek 2008; Aldous et al. 2011; Morlon et al. 2011; Didier et al. 2012; Hallinan 2012). To try to tackle this problem, we implemented a procedure that fits a birth–death model estimating λ and μ simultaneously. These estimates are denoted as $\hat{\lambda}_{\mathrm{BD}}$ and $\hat{\mu}_{\mathrm{BD}}$.

Simulation Study

The present statistical modeling approach offers many possibilities and it would take a large number of simulations to assess the statistical properties of all of them. Instead, we focus on a few key questions. What is the statistical power to detect a difference in diversification between two groups of clades? How powerful is the test to detect temporal variation in diversification? What is the statistical power to detect unknown variation in diversification between two groups of clades using mixtures? Finally, what is the precision of the λ and μ estimators?

To address these four questions, we ran four sets of simulations. First, we considered a simple two-category scenario with n_1 and n_2 clades simulated with rates λ_1 and μ_1 and λ_2 and μ_2, respectively. The times of evolution were drawn from a uniform distribution: $t_i \sim \mathcal{U}(10, 20)$. A phylogeny was simulated under a birth–death process during a time t_i using ape starting from a single species. The number of species surviving at time, t_i, x_i, was extracted and the pairs (x_i, t_i) were analyzed as described above using a Yule model. The LRT testing the null hypothesis of homogeneous diversification was computed, and the rejection rate was assessed under different sets of parameter

values: $n_1 = n_2 = \{1, 3, 5, 10, 20\}$, $\lambda_1 = \{0.1, 0.15, 0.2\}$, $\lambda_2 = 0.1$, $\mu_1 = \{0, 0.05\}$, $\mu_2 = \{0, 0.05\}$.

Second, we performed simulations under three scenarios with different values of diversification rates before and after 30 time units. We first generated 100 values of t from a uniform distribution between 10 and 50. We then simulated clades with constant, increasing, or decreasing diversification rate. The number of species was extracted as before, and two models were fitted: the null Yule model of constant diversification, and an alternative model assuming different speciation rates before and after 30 time units (as above, $\mu = 0$ was assumed). The rejection rates of the LRTs comparing both models were computed.

Third, a scenario similar to the first one was considered: the difference is that the simulated clades were not identified to a particular category so the data were analyzed with a mixture of Yule models. We used $K = 2$, $n_1 = n_2 = \{10, 20, 50\}$, and $t_i \sim \mathcal{U}(10, 20)$. Four combinations of speciation and extinction rates were used as follows: (1) the null hypothesis is true and there is no extinction: $\lambda_1 = \lambda_2 = 0.1$, $\mu_1 = \mu_2 = 0$; (2) the null hypothesis is false and there is no extinction: $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\mu_1 = \mu_2 = 0$; (3) the null hypothesis is false but only μ varies: $\lambda_1 = \lambda_2 = 0.2$, $\mu_1 = 0$, $\mu_2 = 0.1$; and (4) same than before with stronger variation in μ: $\lambda_1 = \lambda_2 = 0.2$, $\mu_1 = 0$, $\mu_2 = 0.15$.

Finally, we performed an assessment of the precision of the estimators of speciation and extinction rates using five combinations of λ and μ: (0.1, 0), (0.1, 0.03), (0.1, 0.06), (0.2, 0.1), and (0.2, 0.15). Here $t_i \sim \mathcal{U}(10, 30)$ and $n = 100$.

The simulations were replicated 1000 times. Annotated R (R Development Core Team 2012) code is available in the Supporting Information with guidelines on how to run these simulations so that the readers can adapt them to their own problems. Besides, we did not attempt to compare our method with previous ones because some scenarios considered here cannot be analyzed by the latter (e.g., the third scenario does not seem to be tractable with Magallón and Sanderson's method).

Application to Fish Data

We used the data from Vega and Wiens (2012) who compiled the number of species, stem age, and percentage of marine fish species for 22 orders and superorders and for 97 families. They also provided a phylogeny for the 22 higher taxa that allowed to compare our estimates with those obtained from the combined analysis of phylogeny and species richness data (Paradis 2003). All data were unmodified from the original publication and are available at http://dx.doi.org/10.1098/rspb.2012.0075. With this dataset, we explored the variation in diversification using different mixtures of Yule and birth–death models. We also tried to assess whether this variation is due to differences in the speciation or in the extinction rates.

Table 1. Rejection rate for the test of equality of diversification rate between two categories with n_1 and n_2 (= n_1) clades.

				n_1				
λ_1	μ_1	λ_2	μ_2	1	3	5	10	20
0.1	0	0.1	0	0.044	0.060	0.066	0.054	0.056
0.1	0	0.1	0.05	0.038	0.085	0.108	0.140	0.203
0.15	0	0.1	0	0.094	0.177	0.232	0.418	0.711
0.15	0	0.1	0.05	0.112	0.297	0.446	0.759	0.958
0.2	0	0.1	0	0.174	0.497	0.707	0.943	0.998
0.2	0	0.1	0.05	0.236	0.643	0.855	0.993	1.000
0.1	0.05	0.1	0.05	0.040	0.054	0.048	0.054	0.061
0.15	0.05	0.1	0	0.050	0.075	0.083	0.123	0.209
0.15	0.05	0.1	0.05	0.069	0.129	0.201	0.387	0.653
0.2	0.05	0.1	0	0.119	0.240	0.384	0.693	0.928
0.2	0.05	0.1	0.05	0.143	0.407	0.618	0.878	0.995

Results

SIMULATION STUDY

The first set of simulations showed that, overall, the LRT testing for different diversification rates between two categories of clades had satisfactory statistical properties (Table 1). The type I error rate (rejection rate when the null hypothesis is true, i.e., $\lambda_1 = \lambda_2$ and $\mu_1 = \mu_2$) was, as expected, close to 5% (first and seventh lines in Table 1). However, when $\lambda - \mu$ was the same in both categories, the rejection rate was greater than 5% (eighth line in Table 1) showing that the present test does not test for equal diversification rate. In the cases where the null hypothesis was not true, the rejection rate varied as expected: it was greater for larger sample sizes (n_1) and for larger contrast in the speciation or extinction rate. Interestingly, if one category of clades had smaller μ while λ was the same, then the test was able to detect this difference; however, the statistical power was less when the same contrast in diversification was due to different λ (compare the second and third lines in Table 1).

In the second set of simulations, the test for temporal variation rejected the null hypothesis more than 90% when $\mu = 0$ and λ varied, either it was an increase or a decrease (third to sixth lines in Table 2). On the other hand, the results were contrasted when $\mu > 0$. When there was no temporal variation in the parameters, the type I error rates were inflated in relation to the value of μ (seventh and eighth lines in Table 2). When μ varied through time, the test behaved very differently depending on the direction of this variation: it did not reject the null hypothesis in most cases when μ increased (ninth line in Table 2), whereas it rejected it in 68% of the cases when μ decreased (tenth line in Table 2). To further

Table 2. Rejection rate for the test of temporal variation in diversification.

Ancient		Recent		
λ	μ	λ	μ	Rejection Rate
0.01	0	0.01	0	0.029
0.1	0	0.1	0	0.038
0.1	0	0.05	0	0.917
0.1	0	0.01	0	1.000
0.05	0	0.1	0	0.923
0.01	0	0.1	0	1.000
0.1	0.025	0.1	0.025	0.105
0.1	0.05	0.1	0.05	0.248
0.1	0.025	0.1	0.075	0.057
0.1	0.075	0.1	0.025	0.682

The null model was a Yule model with constant rate, and the alternative model was a Yule model with λ allowed to take different values before and after 30 time units. The first two pairs of columns give the parameter values used for the simulations (ancient and recent: values before and after 30 time units).

Table 3. Same than in Table 2 but the null model was a birth–death model with constant rates, and the alternative model was a model with λ constant and μ allowed to take different values before and after 30 time units.

Ancient		Recent		
λ	μ	λ	μ	Rejection Rate
0.1	0.05	0.1	0.05	0.019
0.1	0.075	0.1	0.025	0.080
0.1	0.025	0.1	0.075	0.121
0.1	0	0.1	0.08	0.313
0.1	0.08	0.1	0	0.211

investigate this point, we repeated some of these simulations, but this time the null model was a birth–death model with λ and μ constant through time, and the alternative model was with λ constant and μ allowed to vary before and after 30 time units. In this situation, the test behaved as expected: the rejection rate was less than 5% when μ was constant, whereas it varied between 8% and 31% when the null hypothesis was false (Table 3). It is noteworthy that the present test to detect time-dependent extinction rate is not very powerful: it was necessary to simulate a strong contrast in μ to reach a statistical power greater than 0.2.

The third set of simulations showed that the mixture-based LRT was able to detect heterogeneous diversification among two unknown categories of clades (Table 4). The test was more powerful when the contrast was due to different λ compared to different μ. Otherwise, the test showed satisfactory statistical performance:

Table 4. Rejection rate for the test of equality of diversification rate between two unknown categories using mixtures with *n* clades in each category.

		n		
λ	μ	10	20	50
0.1	0	0.011	0.006	0.011
(0.1, 0.2)	0	0.235	0.548	0.929
0.2	(0, 0.1)	0.057	0.143	0.369
0.2	(0, 0.15)	0.199	0.423	0.829

Table 5. Results of fitting models to the fish data using mixtures of Yule processes with *K* from two to seven.

	Orders		Families	
K	ln *L*	AIC	ln *L*	AIC
2	−171.207	348.414	−599.846	1205.691
3	−158.381	326.763	−599.846	1209.691
4	−171.207	356.414	−599.846	1213.691
5	−171.207	360.414	−599.846	1217.691
6	−171.207	364.414	−599.846	1221.691
7	−171.207	368.414	−599.846	1225.691

its power increased with sample size and/or contrast in the parameters.

The distribution of the estimates of speciation rate under the Yule model, $\hat{\lambda}_{\text{Yule}}$, shows that this estimator appeared unbiased when $\mu = 0$ (Fig. 1A). On the other hand, when $\mu > 0$, it was negatively biased although it can be observed that $\hat{\lambda}_{\text{Yule}} > \lambda - \mu$ so this cannot be actually taken as an estimator of the net diversification rate. The estimator based on the birth–death model, $\hat{\lambda}_{\text{BD}}$, appears less biased, even though the presence of extinctions seems to induce a slightly more dispersed distribution of the estimates (Fig. 1B). The estimates of extinction rate based on the birth–death model, $\hat{\mu}_{\text{BD}}$, were almost unbiased (Fig. 1C).

APPLICATION TO FISH DATA

The fit of the Yule model to the fish data at the higher level ($N = 22$) resulted in a global estimate $\hat{\lambda}_{\text{Yule}} = 0.058$ (SE = 0.002; AIC = 456). We tried to fit a birth–death model which led to a much improved fit (AIC = 376); however, the likelihood function had a pronounced ridge on the line $\lambda = \mu$ (not shown). The fit of mixtures of Yule models with increasing number of categories (K) showed that the best fit was with three categories (Table 5). The parameter estimates were $\hat{\lambda}_1 = 0.041$, $\hat{\lambda}_2 = 0.080$, $\hat{\lambda}_3 = 0.013$, $\hat{f}_1 = 0.65$, and $\hat{f}_2 = 0.10$. The analysis of the combined taxonomic and phylogenetic data (Paradis 2003) gave $\hat{\lambda} = 0.056$ and $\hat{\mu} = 1.83 \times 10^{-7}$.

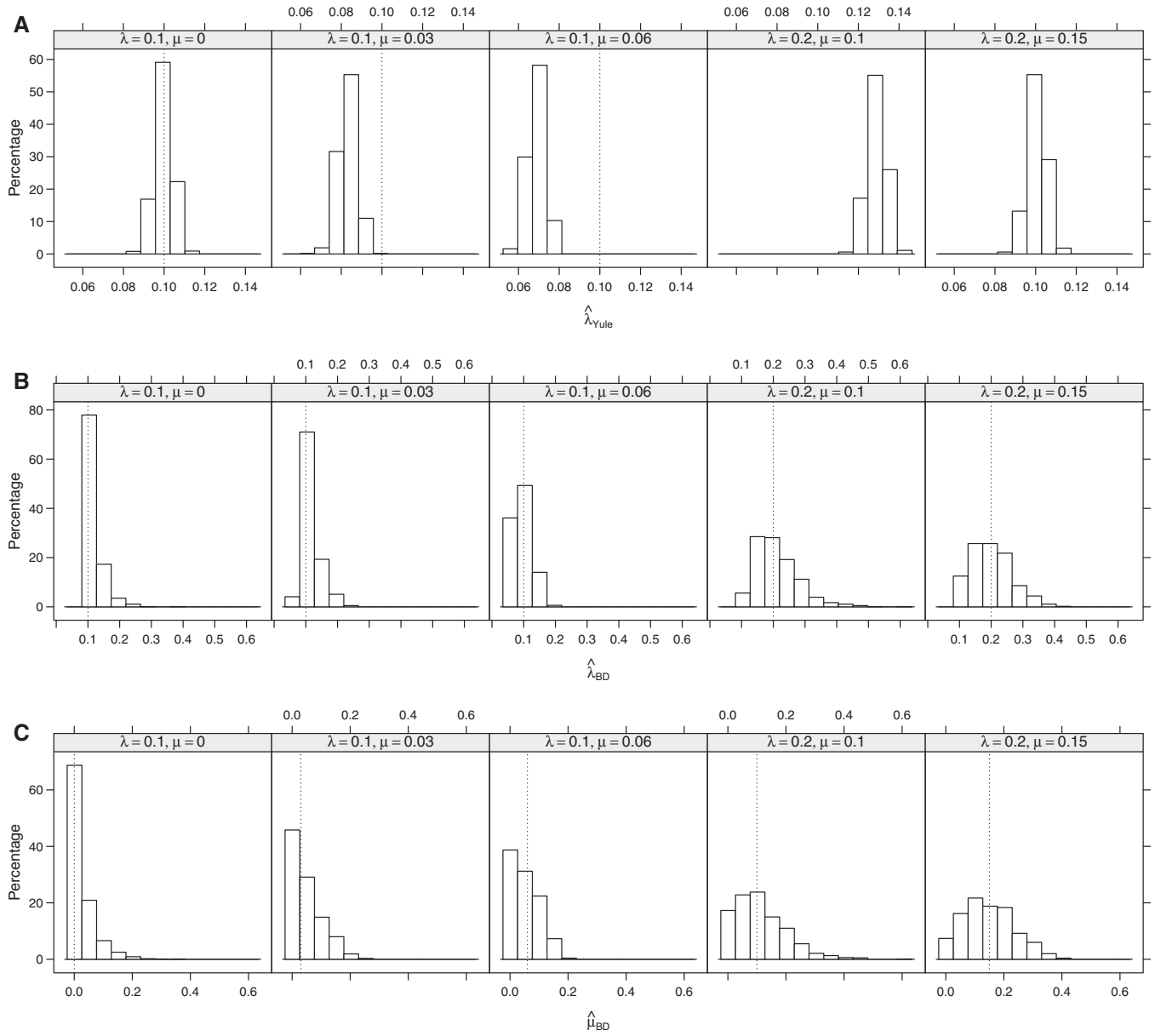

Figure 1. Distribution of the estimates of λ and μ with (A) the Yule model ($\hat{\lambda}_{Yule}$) and (B and C) the birth–death model ($\hat{\lambda}_{BD}$ and $\hat{\mu}_{BD}$) under five sets of parameters (values are given in the strips). Note the different scales of the x-axes. The vertical dotted lines indicate the values of λ (A and B) or μ (C) used in the simulation (not visible if outside the range of the x-axis). In all cases, $n = 100$ clades.

The analysis at the level of the families ($N = 97$) gave for the Yule model, $\hat{\lambda}_{Yule} = 0.0756$ (SE = 0.0016; AIC = 1483). Like above, the fit of the birth–death model resulted in a likelihood surface with a ridge on the line $\lambda = \mu$. The mixture of Yule models with the best fit had two categories (Table 5); the parameter estimates were as follows: $\hat{\lambda}_1 = 0.099$, $\hat{\lambda}_2 = 0.036$, $\hat{f} = 0.42$.

The analysis with a model assuming continuous variation in λ across clades gave close results for both taxonomic levels. In both cases, the model fitted well and the AIC values were smaller than for any of the previous models (Table 6). Figure 2 shows the distribution of λ inferred with the estimated parameters. Trying to introduce μ did not result in successful fits and the estimates of this parameter were close to zero.

Table 6. Results of fitting a model of continuous variation in speciation rate across orders ($N = 22$) and families ($N = 97$) of fish.

	AIC	$\hat{\mu}_\lambda$ (SE)	$\hat{\sigma}_\lambda$ (SE)
Orders	320.396	1.221 (0.039)	0.163 (0.032)
Families	1137.066	1.086 (0.026)	0.224 (0.022)

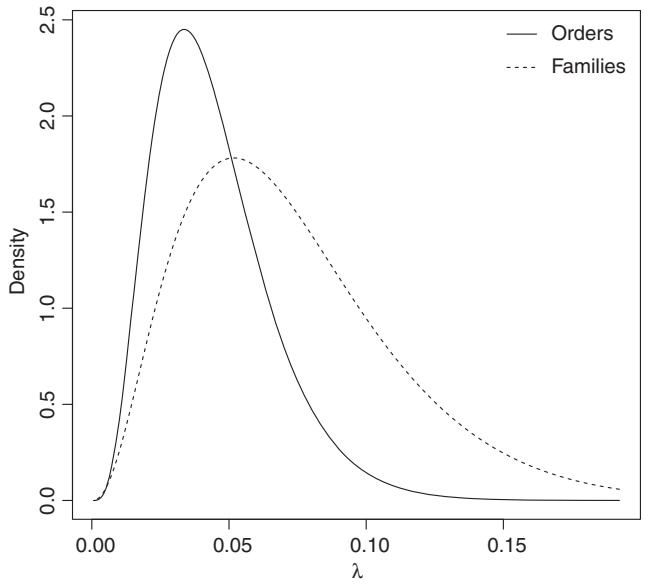

Figure 2. Inferred distribution of speciation rate among orders and families of fish.

Vega and Wiens (2012) reported the percentage of marine and freshwater species at both taxonomic levels. This was distributed very asymmetrically with most orders and families having only marine or freshwater species. Thus, we split the data into two groups whether they had more or less than 50% of marine species. A test of different speciation rates between these groups was performed. For orders, the difference was significant (LRT: $\chi^2_1 = 28.09$, $P < 0.001$) with a larger estimate for marine orders ($\hat{\lambda} = 0.063$, SE $= 0.002$) compared to the freshwater ones ($\hat{\lambda} = 0.046$, SE $= 0.002$). An examination of the data suggested that this result was dependent on Percomorpha, which is one of the youngest clades in this dataset and includes 16,625 species (Fig. 3A). Removing this clade resulted in a nonsignificant test ($\chi^2_1 = 2.10$, $P = 0.147$, $N = 21$). For families, an analogous result was found with a significant test (LRT: $\chi^2_1 = 5.58$, $P = 0.018$) comparing marine families ($\hat{\lambda} = 0.079$, SE $= 0.002$) and freshwater ones ($\hat{\lambda} = 0.071$, SE $= 0.002$). This result was dependent on two families older than 200 million years (Fig. 3B): the Amiidae (one species) and Polypteridae (12). Removing these two families led to a nonsignificant test: $\chi^2_1 = 2.02$, $P = 0.155$ ($N = 95$).

Discussion

The analysis of phylogenetic diversification with molecular data is enjoying a remarkable success in the literature. Some spectacular results have been accomplished using complete phylogenies (e.g., Goldberg et al. 2010; Hugall and Stuart-Fox 2012; Penney et al. 2012). Although complete phylogenies, possibly supplemented with fossil data, are probably the best way to investigate evolutionary diversification, the goal of our study was to show the merit of an alternative approach based on the analysis of clade data.

Our modeling approach is based on the assumption that each clade is characterized by its diversification parameters and variation among these parameters can be quantified in a statistical way. Bokma (2003) and Paradis (2003) developed a method to combine information from high-level phylogenies with clade data: both authors considered the simple constant-rate birth–death model. Alfaro et al. (2009) used similar combined data to assess variation among clades of vertebrates using a stepwise procedure (see details in Paradis 2012a). Thus, the approach in the present article complements previous methodological developments. The possibility to quantify variation among clades with linear models seems

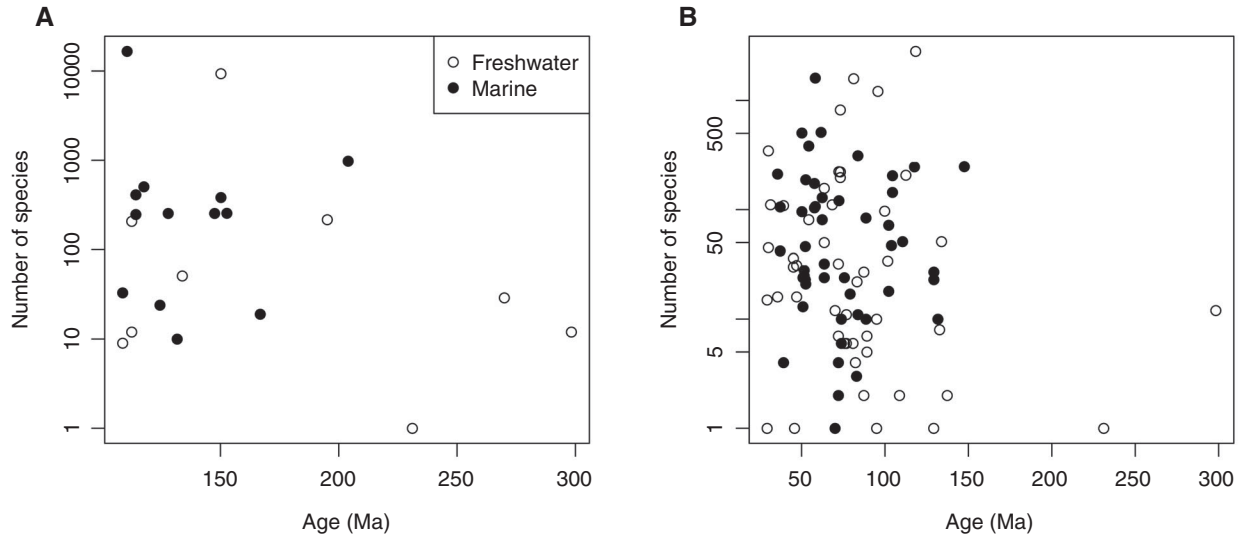

Figure 3. Number of species with respect to stem clade age for (A) orders and some superorders and (B) families of fish.

a fruitful way to avoid overparameterization. Future applications will reinforce the relative merits of this approach.

Recently, Stadler and Bokma (2013) developed alternative likelihood functions with respect to the way higher taxa are defined. They showed that the estimation of speciation and extinction rates vary substantially depending on these definitions. Although they considered only the constant-rate birth–death model, it seems possible and interesting to include their sampling scheme into the developments presented in the present article.

Our modeling approach ignores the background phylogeny of the clades, the set of branches that link the clades together to make a higher-level phylogeny. There are two reasons for this. First, using information from the background phylogeny is straightforward when the rates of speciation and extinction are constant and homogenous, but when this assumption is relaxed, it is not simple how one assumes changes in rates in the background tree. It is clear that if a well-supported background phylogeny is available, this might give additional information that can be combined with clade data (e.g., Paradis 2003). However, this extra information will in most cases require its own model because it relates to older diversification events compared to clade data. On the other hand, ignoring backbone phylogeny and assuming that the clades are independent units simplifies the definition of alternative models as done in this article. Second, although some higher-level phylogenies are available (mammals, birds), we believe these are still exceptions rather than the rule. For instance, the basal relationships of reptiles, amphibians, or fishes are still debated. Therefore, having the possibility to analyze their clade data without the need of a background phylogeny is of some general application. Furthermore, the present approach can be used when analyzing sets of clades across different phyla, for instance arthropods, echinoderms, vertebrates, etc., where the background phylogeny would not be very informative because this would branch at the origin of metazoa.

The use of mixtures as an approach to analyze heterogeneity in diversification rates is not limited to clade data. For instance, one could model speciation and extinction rates on a fully resolved phylogeny assuming that these parameters vary among its branches, though we do not know a priori which sections of the tree evolved fast and which others evolved slowly. Furthermore, the mixture approach can also be used to model variation in rates of trait evolution along a phylogeny. In that case, the variation may be among branches (as in the previous example), or among traits where some traits are assumed to evolve faster but we do not know which ones.

Some subtle but important facts come from the results of the simulation study. Even though most of the tests considered here assumed $\mu = 0$, they appeared not to be tests of equal diversification. If the net diversification rates ($\lambda - \mu$) were equal among clades, the tests rejected the null hypothesis in more than 5% (see eighth row of Table 1). On the other hand, if λ was equal among clades, the tests detected differences in μ. It is clear that results based only on the Yule model must be interpreted with caution.

The tests of temporal variation showed some contrasted but interesting results. When the extinction rate was zero, these tests performed very well and were able to detect either a decrease or an increase in speciation rate. However, when extinction rate was not null, the tests based on the Yule model showed poor performance with an increased type I error rate and a high type II error rate (frequency of accepting the null hypothesis when it is false) when μ decreased through time. These poor performances were corrected if the assumption $\mu = 0$ was relaxed (i.e., if a null birth–death model was used in place of the Yule one), although the test had low power. Some of these results make sense: the increased type I error rate obtained with the Yule model is clearly due to the fact that a pattern of accelerated speciation can be created under a diversification process with extinction, when old lineages are mostly extinct (e.g., Paradis 2011). On the other hand, the high type II error rate of the same model when extinction rate increased through time is somehow surprising considering the widely reported results of slowing down diversification (Rabosky and Lovette 2008a,b; Morlon et al. 2011; Etienne and Rosindell 2012, among others). Obviously, the same test was not used in these studies, so this clearly requires further investigation. Besides, the result that the test based on a birth–death model shows statistically consistent results (i.e., the null hypothesis was rejected in less than 5% when μ was constant and in more than 5% when this parameter varied through time) is encouraging and will also be further investigated. Interestingly, this test was more powerful when the extinction rate increased through time.

A particularly interesting result comes from the precision of the estimator of extinction rate, $\hat{\mu}_{BD}$, which appears to have a very small bias, even when the data were simulated with a relatively large value of μ. This contrasts with previous studies showing that the estimator of extinction rate based on complete phylogenies is, overall, inaccurate except if it is small compared to the speciation rate (Paradis 2004; Didier et al. 2012). This result is important because several authors have cast doubt on the possibility to estimate with some precision extinction rates without fossils (Aldous et al. 2011; Paradis 2011).

The analysis with the fish data were essentially illustrative, but the results call for several comments. The present method seems successful in quantifying variation in diversification rates from a sample of clades. The difference in the results from both taxonomic levels makes sense because we expect more variation among families than among orders. The AIC values evidence that the model assuming continuous variation in λ across clades fits better than a model with discrete variation in this parameter. Because similar tests have not been done with other data, this

clearly calls for further analyses before concluding whether diversification varies continuously or discretely across clades.

The apparent failure to estimate the extinction rate, μ, of fishes is disappointing because our simulation study showed that this parameter can be estimated correctly with the present approach. The fossil record shows many episodes of radiations, extinctions, and turnover during the evolutionary history of fishes (Friedman and Sallan 2012). So, the reality is very different from the homogeneous scenario used in our simulations. Our results combined with previous studies (e.g., Aldous et al. 2011) suggest that the estimators of μ are far more complex when rate heterogeneity is present which is likely the case with most real dataset.

Vega and Wiens (2012) addressed the paradox of equivalent species diversity between marine and freshwater fishes despite the fact that freshwater environments occupy a considerably smaller fraction of the Earth's surface than oceans. In particular, they wondered whether this could be related to differences in diversification rates. Our results are in agreement with these authors' who tested their hypothesis by correlating the proportion of marine species in a clade with the method-of-moment estimator from Magallón and Sanderson (2001). We found significant differences in λ between marine and freshwater clades from the raw data; however, the small difference in $\hat{\lambda}$ between both groups suggested the influence of one or two clades. Hopefully, the analysis of a more comprehensive dataset with the statistical tools introduced in this article will help to solve the paradox of less biological diversity in the ocean (Mora et al. 2011).

ACKNOWLEDGMENTS

We are grateful to four anonymous reviewers, the Associate Editor, and L. Kubatko for their constructive comments on previous versions of our manuscript. Financial support was provided by grant ANR-09-PEXT-008.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pp. 267–281 *in* B. N. Petrov and F. Csaki, eds. Proceedings of the Second International Symposium on Information Theory. Akadémia Kiado, Budapest.

Aldous, D. J., M. A. Krikun, and L. Popovic. 2011. Five statistical questions about the tree of life. Syst. Biol. 60:318–328.

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. USA 106:13410–13414.

Bokma, F. 2003. Testing for equal rates of cladogenesis in diverse taxa. Evolution 57:2469–2474.

Didier, G., M. Royer-Carenzi, and M. Laurin. 2012. The reconstructed evolutionary process with the fossil record. J. Theor. Biol. 315:26–37.

Etienne, R. S., and J. Rosindell. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61:204–213.

FitzJohn, R. G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–633.

FitzJohn, R. G., W. P. Maddison, and S. P. Otto. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

Flury, B. D., J.-P. Airoldi, and J.-P. Biber. 1992. Gender identification of water pipits (*Anthus spinoletta*) using mixtures of distributions. J. Theor. Biol. 158:465–480.

Friedman, M., and L. C. Sallan. 2012. Five hundred million years of extinction and recovery: a Phanerozoic survey of large-scale diversity patterns in fishes. Palaeontology 55:707–742.

Goldberg, E. E., J. R. Kohn, R. Lande, K. A. Robertson, S. A. Smith, and B. Igić. 2010. Species selection maintains self-incompatibility. Science 330:493–495.

Hallinan, N. 2012. The generalized time variable reconstructed birth–death process. J. Theor. Biol. 300:265–276.

Hugall, A. F., and D. Stuart-Fox. 2012. Accelerated speciation in colour-polymorphic birds. Nature 485:631–634.

Kendall, D. G. 1948. On the generalized "birth-and-death" process. Ann. Math. Stat. 19:1–15.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.

Magallón, S., and M. J. Sanderson. 2001. Absolute diversification rates in angiosperm clades. Evolution 55:1762–1780.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. Am. Nat. 172:E270–E284.

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How many species are there on Earth and in the ocean? PLoS Biol. 9:e1001127.

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108:16327–16332.

Nee, S., A. Ø. 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. Phil. Trans. R. Soc. Lond. B 344:305–311.

Paradis, E. 2003. Analysis of diversification: combining phylogenetic and taxonomic data. Proc. R. Soc. Lond. B 270:2499–2505.

———. 2004. Can extinction rates be estimated without fossils? J. Theor. Biol. 229:19–30.

———. 2005. Statistical analysis of diversification with species traits. Evolution 59:1–12.

———. 2011. Time-dependent speciation and extinction from phylogenies: a least squares approach. Evolution 65:661–672.

———. 2012a. Analysis of phylogenetics and evolution with R (second edition). Springer, New York.

———. 2012b. Shift in diversification in sister-clade comparisons: a more powerful test. Evolution 66:288–295.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Penney, H. D., C. Hassall, J. H. Skevington, K. R. Abbott, and T. N. Sherratt. 2012. A comparative analysis of the evolution of imperfect mimicry. Nature 483:461–464.

Purvis, A., S. Nee, and P. H. Harvey. 1995. Macroevolutionary inferences from primate phylogeny. Proc. R. Soc. Lond. B 260:329–333.

Pybus, O. G., A. Rambaut, E. C. Holmes, and P. H. Harvey. 2002. New inferences from tree shape: numbers of missing taxa and population growth rates. Syst. Biol. 51:881–888.

R Development Core Team. 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at http://www.R-project.org.

Rabosky, D. L., and I. J. Lovette. 2008a. Density-dependent diversification in North American wood warblers. Proc. R. Soc. Lond. B 275:2363–2371.

———. 2008b. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? Evolution 62:1866–1875.

Rabosky, D. L., S. C. Donnellan, A. L. Talaba, and I. J. Lovette. 2007. Exceptional among-lineage variation in diversification rates during the radiation of Australia's most diverse vertebrate clade. Proc. R. Soc. Lond. B 274:2915–2923.

Ricklefs, R. E. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22:601–610.

Sanderson, M. J., and M. J. Donoghue. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. Trends Ecol. Evol. 11:15–20.

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. USA 108:6187–6192.

Stadler, T., and F. Bokma. 2013. Estimating speciation and extinction rates for phylogenies of higher taxa. Syst. Biol. 62:220–230.

Vega, G. C., and J. J. Wiens. 2012. Why are there so few fish in the sea? Proc. R. Soc. Lond. B 279:2323–2329.

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

Associate Editor: M. Rosenberg

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Data File S1

Data File S2