Syst. Biol. 62(2):220–230, 2013

DOI:10.1093/sysbio/sys087
Advance Access publication October 25, 2012

Estimating Speciation and Extinction Rates for Phylogenies of Higher Taxa

*Department of Environmental Systems Science, ETH Zürich, 8092 Zürich, Switzerland; and †IceLab and Department of Ecology and Environmental Science, Umea University, 90187 Umea, Sweden
*Correspondence to be sent to: Institute for Integrative Biology, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland; E-mail: tanja.stadler@env.ethz.ch

Received 25 January 2012; reviews returned 12 June 2012; accepted 11 October 2012
Associate Editor: Olivier Gascuel

Abstract.—Speciation and extinction rates can be estimated from molecular phylogenies. Recently, a number of methods have been published showing that these rates can be estimated even if the phylogeny is incomplete, that is, if not all extant species are included. We show that the accuracy of such methods strongly depends on making the correct assumptions about how the sampling process was performed. We focus on phylogenies that are incomplete because some subclades (e.g., genera and families) are each represented as a single lineage. We show that previous methods implicitly assumed that such subclades are defined by randomly (or in an extreme deterministic way) choosing the edges that define the subclades from the complete species phylogeny. We show that these methods produce biased results if higher taxa are defined in a different manner. We introduce *strict higher level phylogenies* where subclades are defined so that the phylogeny is fully resolved from its origin to time x_{cut}, and fully unresolved thereafter, so that for all subclades, stem age $> x_{\text{cut}} >$ crown age. We present estimates of speciation and extinction rates from a phylogeny of birds in which this subclade definition was applied. However, for most higher level phylogenies in the literature, it is unclear how higher taxa were defined, but often such phylogenies can be easily transformed into strict higher level phylogenies, as we illustrate by estimating speciation and extinction rates from a near-complete but only partly resolved species-level phylogeny of mammals. The accuracy of our methods is verified using simulations. [Birth–death process; higher taxa; macroevolution; phylogenetics.]

Introduction

Molecular phylogenies represent hypotheses about the historical relationships of species in the form of a bifurcating tree. Because it is rarely possible to obtain molecular data from fossil remains, these trees are typically pruned of extinct species. Nevertheless, it has been shown that the time axis of phylogenies (i.e., the times of bifurcations) provides information about both speciation and extinction rates despite the extinct species being pruned (Nee et al. 1994). In other words, if an appropriate model is assumed for evolutionary diversification, speciation and extinction rates can be estimated from phylogenies of present-day species.

The first method to estimate speciation and extinction rates from molecular phylogenies required phylogenies to be complete at the species level (Nee et al. 1994; Kubo and Iwasa 1995). That is, they require all extant members of a monophyletic group (a clade) to be represented on the tree. Unfortunately, only a small fraction of published phylogenies meets that requirement: for some well known or small groups of species, phylogenies may contain all species, but the large majority of published phylogenies remain incomplete and can therefore not be used to estimate speciation and extinction rates. As ever more molecular phylogenies were published, and interest in estimating speciation and extinction rates increased, new methods were developed to estimate rates also from incomplete phylogenies (Yang and Rannala 1997; Paradis 2003; Rabosky et al. 2007; Alfaro et al. 2009; FitzJohn et al. 2009; Stadler 2009; Höhna et al. 2011; Morlon et al. 2011; Stadler 2011a).

Incomplete phylogenies can be roughly grouped into 2 types: first, a fraction of all species might be missing in the phylogeny, meaning that the lineages belonging to missing species are pruned from the complete species tree. Second, phylogenies might be only resolved up to a higher taxonomic level (e.g., family or order), meaning that each higher taxon in the complete species tree is collapsed to one lineage. Typically, the number of species within each higher taxon is known.

In recent articles (Cusimano and Renner 2010; Brock et al. 2011; Höhna et al. 2011), it was shown that when considering the first type of incomplete phylogenies, it is crucial to have accurate information on how species were sampled, in order to obtain accurate speciation and extinction rate estimates. In the current article we will show that, when considering higher level phylogenetic trees, it is crucial to have accurate information on how higher taxa were defined. For both types of incomplete phylogenies, incorrect assumptions typically bias the extinction rate estimates to be zero.

When estimating speciation and extinction rates, we need to assume a model for the macroevolutionary process. The simplest model used for inferring macroevolutionary rates is a *constant rate birth–death process* (crBDP, Kendall 1948a; Feller 1968); in this model each species has the same constant rate of speciation and the same constant rate of extinction. Although the crBDP may be inappropriate in particular for large and old phylogenies, it has substantially improved our understanding of evolutionary diversification by serving as a null model in many evolutionary and paleontological studies (Raup et al. 1973; Foote et al. 1999). Here, we will focus on the crBDP and show how the rate estimates obtained under the crBDP are very sensitive toward the assumption of how higher taxa were defined. In other words, the speciation and extinction rate estimates are biased if an inappropriate assumption of how higher taxa were defined is used. Because the crBDP is a special case of the more complex models

proposed in the literature (Rabosky et al. 2007; Alfaro et al. 2009; FitzJohn et al. 2009; Etienne et al. 2012; Höhna et al. 2011; Morlon et al. 2011; Stadler 2011a), this bias may also be present in these more complex methods.

When considering higher level phylogenies, we need to assume a model describing how higher taxa are chosen from a complete species-level phylogeny. The first kind (i) of higher level phylogenies that we consider are *random higher level phylogenies* where the species phylogeny is partitioned into subclades (higher taxa) as follows: each edge has the same probability s of giving rise to a subclade (higher taxon). If an edge gives rise to a higher taxon, we replace all species descending this edge by a single edge. If a chosen edge is ancestral of another chosen edge, the more recent subclade is ignored. We assume that the number of extant species in each subclade is known, that is, each pendant edge has associated with it the number of extant species represented by that edge. As an example, the chosen edges of a species phylogeny in Figure 1b are denoted in bold with a circle at the start of the edge. The resulting random higher level phylogeny is displayed in Figure 1c.

The second kind (ii) of higher level phylogeny for which we present estimation functions is the *strict higher level phylogeny*. We define a strict higher level phylogeny as a species phylogeny fully resolved from its origin up to a certain point in time (x_{cut}), and fully unresolved after that point. Hence, all extant species are grouped in subclades, where each subclade has a crown age younger and a stem age older than x_{cut}. Again, we assume that the species numbers of these subclades are known. As an example, a time x_{cut} is displayed in the species phylogeny in Figure 1b with a dashed line. The resulting strict higher level phylogeny is displayed in Figure 1d.

While in the random higher level phylogeny, each edge has the same probability of giving rise to a higher taxon (with nested edges being ignored), in the strict higher level phylogeny, each edge existing at time x_{cut} gives rise to a higher taxon with probability 1, and all other edges do not give rise to a higher taxa with probability 1. The 2 scenarios (i) and (ii) can thus be seen as 2 extremes, where in reality higher taxa are defined with some (unknown) intermediate process.

Using simulations, we demonstrate that our derived functions for estimating speciation and extinction rates give correct results if the correct definition of data selections (i) and (ii) is assumed, and that results are incorrect when the assumptions are not met. In particular, extinction rates are underestimated, if scenario (i) is assumed while scenario (ii) is correct.

A well-known example of a strict higher level phylogeny is the *tapestry* phylogeny of families of birds by Sibley and Ahlquist (1990) who applied a strict genetic-distance-based definition of families. Therefore, we use this phylogeny to illustrate the estimation of speciation and extinction rates using the higher taxa definitions (i) and (ii). We show that the inappropriate scenario (i) and a previous method for estimating speciation and extinction rates based on higher level phylogenies (Paradis 2003) produce results significantly different from those obtained using the appropriate scenario (ii). We analyze the mammalian phylogeny by Bininda-Emonds et al. (2007) to illustrate that scenario (ii) can also deal with more general phylogenies (including recent polytomies, nonrandom sampling), by altering the phylogeny in an appropriate way. Our new speciation and extinction rate estimation method is implemented into the R package TreePar (Stadler 2011a) available on CRAN.

Methods

Throughout this article we will assume the (crBDP) (Kendall 1948b; Feller 1968) as a model for the evolutionary diversification of clades. Therefore, we first define the crBDP of speciation and extinction and review some known results for the crBDP that are needed for our probability density derivations.

The crBDP starts with one species at some *time of origin* x_0 in the past. This and every subsequent species may give birth to new species with rate λ. Species may cease to exist at a rate μ, which, just like λ, is equal for all species and constant over time. A crBDP that evolves between a time x_0 in the past and the present induces a species tree of age x_0 with extinct lineages being included (Fig. 1a). We obtain a *reconstructed tree* by pruning all extinct lineages (Fig. 1b).

Throughout this article, we consider *oriented trees* where the 2 descendants of a branching event are distinguished to be l and r. The derivation of probability density functions is easier using oriented trees than using the more common *labeled trees* where each leaf is labeled with a unique name. We recall that the likelihood used for parameter inference is the probability density function up to any normalization constants. For parameter estimation, the likelihood for oriented trees and the likelihood for labeled trees yield the same results (for details see the Supplementary Information; Dryad doi:10.5061/dryad.b0c8470m).

Under a crBDP, the probability that a lineage leaves n descendants after time t, $p_n(t)$, is (Kendall 1949):

$$p_0(t) = \frac{\mu\left(1-e^{-(\lambda-\mu)t}\right)}{\lambda-\mu e^{-(\lambda-\mu)t}},$$

$$p_1(t) = \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{\left(\lambda-\mu e^{-(\lambda-\mu)t}\right)^2},$$

$$p_n(t) = (\lambda/\mu)^{n-1} p_1(t)\left[p_0(t)\right]^{n-1}.$$

It should be noted that the time of origin x_0 of a tree is a parameter of the birth–death process. If this time would be known, we could fix the parameter x_0 and estimate the parameters of interest λ and μ. However, for a reconstructed phylogeny, usually no precise information about x_0 is available. We can instead assume a uniform prior on $(0,\infty)$ for the time of origin of the process. However, in that case, the probability of obtaining a finite

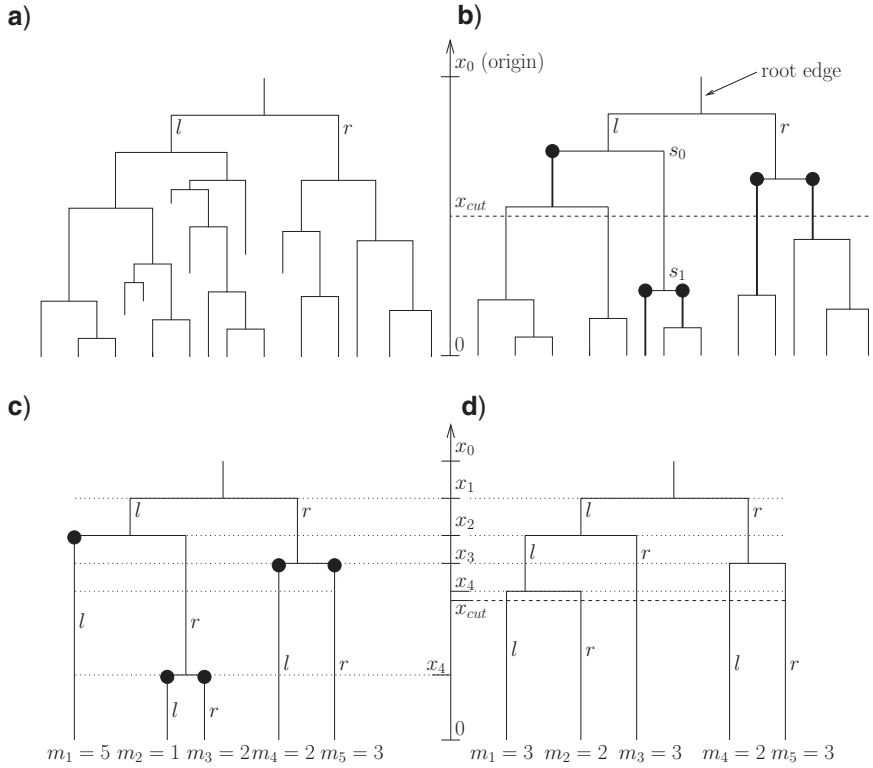

FIGURE 1. a) Species tree with extinct lineages, and b) the corresponding reconstructed species tree induced by the crBDP of age x_0. Always the left edge descending a speciation event has orientation l and the right edge has orientation r. The orientation l and r is only given for the first speciation event for easier readability. The bold edges with the black circles denote higher taxa defined under scenario (i). The line at x_{cut} denotes the cutoff under scenario (ii). c) Random higher level phylogeny (scenario (i)) obtained from the reconstructed tree in Figure 1b where the higher taxa are defined by the bold edges with the black circles. Note that $t_1 = x_2, t_2 = x_4, t_3 = x_4, t_4 = x_3$, and $t_5 = x_3$. d) Strict higher level phylogeny (scenario (ii)) obtained from the reconstructed tree in Figure 1b where the higher taxa are defined by the species extant at time x_{cut}.

tree is 0, so the process must be conditioned on obtaining m extant species (Aldous and Popovic 2005; Gernhard 2008). Alternatively, instead of assuming a prior for x_0 (e.g., because we have no good prior assumptions), one can condition on the age of the root (i.e., the first split), x_1, which is directly available from the reconstructed tree. This means that 2 birth–death processes, started at time x_1, together give rise to the observed extant species. Note that conditioning on x_1 implicitly puts a prior on x_0.

Overall, the 3 conditionings are very natural: the time of origin x_0, the time since the most recent common ancestor x_1, or the number of extant species in conjunction with an (improper) uniform prior on $(0, \infty)$ for x_0. We will present probability density functions for higher level phylogenies under all these 3 conditions in the next sections.

From a reconstructed phylogeny, we obtain a *higher level phylogeny* ($\mathcal{T}$) by replacing non-nested subclades by single pendant edges, each associated with the number of species in the subclade.

A higher level phylogeny $\mathcal{T}$ with n leaves (i.e., n higher taxa) has $n-1$ branching times, which will be denoted by $x_1, \ldots, x_{n-1}$. We measure time before present (where the present is at time 0), thus we have $x_i > x_{i+1}$ for all i. The n leaves consist of $m_1, m_2, \ldots, m_n$ species ($m \geq 1$), such that in total there are $m = \sum_{i=1}^{n} m_i$ species, and the taxa are attached to the tree at times $t_1, t_2, \ldots, t_n$ where $t_i \in \{x_1, \ldots, x_{n-1}\}$ for $i \in \{1, \ldots, n\}$. Thus, the stem age of a subclade is the length of the edge that represents it.

In the following, we derive the probability density of a higher level phylogeny under 2 different definitions of how higher taxa are defined. These probability density functions will be used to estimate maximum-likelihood speciation and extinction rates based on simulated and empirical higher level phylogenies. Since our model produces ultrametric trees in units of calendar time, we require empirical trees which are dated, that is, which were inferred using some molecular clock. This can be done by inferring first a nonclock tree using either a neighbor-joining, parsimony or maximum-likelihood approach, followed by inferring branching times using, for example, Paml (Yang 2007). Alternatively, a calendar time tree can be inferred directly from sequence data using an MCMC approach assuming, for example, a strict or relaxed molecular clock; this can be done, for example, with Beast (Drummond and Rambaut 2007).

Random Higher Level Phylogenies

In this section, we consider the case where subclades (higher taxa) are defined in a random way as follows. Each edge in the reconstructed phylogeny is sampled with constant probability s. If an edge is sampled it defines a subclade (higher taxon), meaning that all descending edges are pruned and represented by a single pendant edge in the higher level phylogeny, together with the number of extant species belonging to the subclade (Fig. 1c). If a chosen edge is ancestral to another chosen edge, the more recent subclade is ignored. Let k be the number of pendant edges which only represent one species in the higher level phylogeny; in Figure 1c, we have $k=1$. The probability density of a random higher level phylogeny is provided in the following theorem. A proof is found in the Supplementary Information.

Theorem 1 *The probability density of the random higher level phylogeny $\mathcal{T}$ conditioned on the time since origin is*

$$p(\mathcal{T}|x_0)=\lambda^{n-1}s^{n-k}(1-s)^{n-1}\prod_{i=0}^{n-1}p_1(x_i)\prod_{i=1}^{n}\left(\frac{\lambda}{\mu}p_0(t_i)\right)^{m_i-1}. \quad (1)$$

Now, assume that we do not know x_0. We are interested in $\mathcal{T}$ conditioned on the time since the first speciation event (x_1). The probability density of $\mathcal{T}$ conditioned on the time since the first speciation event x_1 is composed of (a) the probability density of the left subtree descending x_1 conditioned on both its time since origin being x_1 and its survival until today, times (b) the probability density of the right subtree descending x_1 conditioned on both its time since origin being x_1 and its survival until today. Note that the probability density of $\mathcal{T}$ conditioned on the time since origin (Equation 1) provides (a) and (b) without conditioning on survival. Conditioning on survival is obtained through dividing by the probability of survival of the 2 lineages descending from the root, $(1-p_0(x_1))^2$. Therefore, the probability density of $\mathcal{T}$ conditioned on the time since the first speciation event being x_1 is

$$p(\mathcal{T}|x_1)=\lambda^{n-2}s^{n-k}(1-s)^{n-2}\frac{p_1(x_1)^2}{(1-p_0(x_1))^2}\prod_{i=2}^{n-1}p_1(x_i)\prod_{i=1}^{n}\left(\frac{\lambda}{\mu}p_0(t_i)\right)^{m_i-1}. \quad (2)$$

Finally, we are interested in $\mathcal{T}$ conditioned on the number of species (m) in the tree, while assuming a uniform prior for x_0. In the Supplementary Information, we prove:

Theorem 2 *The probability density of the random higher level phylogeny $\mathcal{T}$ conditioned on the number of species, m, with a uniform prior for the time of origin of the tree is*

$$p(\mathcal{T}|m)=m\lambda^{n-1}s^{n-k}(1-s)^{n-2}\frac{p_1(x_1)^2}{(1-p_0(x_1))}\prod_{i=2}^{n-1}p_1(x_i)\prod_{i=1}^{n}\left(\frac{\lambda}{\mu}p_0(t_i)\right)^{m_i-1}. \quad (3)$$

This result had been established for $m=n$ in Gernhard (2008).

Strict Higher Level Phylogenies

Above we considered higher level phylogenies with randomly selected subclades. In this section, we will consider another kind of higher level phylogeny, where the subclades are not selected randomly, but where we use a specific time x_{cut} and collapse all bifurcations that are more recent than x_{cut}, so that x_{cut} defines all the subclades (Fig. 1d). Hence, the age of each subclade is at least x_{cut}.

A well-known example of such a strict higher level phylogeny is the DNA–DNA hybridization phylogeny of birds by Sibley and Ahlquist (1990).

First, let us again consider the probability density of the higher level phylogeny $\mathcal{T}$ conditioned on the time since its origin, x_0.

Theorem 3 *The probability density of the strict higher level phylogeny $\mathcal{T}$ is*

$$p(\mathcal{T}|x_0)=\lambda^{n-1}\left(\frac{\lambda}{\mu}p_0(x_{cut})\right)^{m-n}\prod_{i=0}^{n-1}p_1(x_i). \quad (4)$$

A proof is found in the Supplementary Information. Note that if we replace t_i by x_{cut} in Theorem 1 and ignore the term $s^{n-k}(1-s)^{n-1}$ (which does not influence the maximum-likelihood estimates), we obtain Theorem 3.

In order to obtain the probability density of the higher level phylogeny conditioned on the time since the first split, x_1, we follow the same logic as for the random subclades above: the probability density $p(\mathcal{T}|x_1)$ is composed of the probability densities of the left and right subtrees descending from the first split conditioned on both the time since origin being x_1 (Equation 4) and survival until today, $(1-p_0(x_1))^2$. Therefore,

$$p(\mathcal{T}|x_1)=\lambda^{n-2}\frac{p_1(x_1)^2}{(1-p_0(x_1))^2}\left(\frac{\lambda}{\mu}p_0(x_{\text{cut}})\right)^{m-n}\prod_{i=2}^{n-1}p_1(x_i). \quad (5)$$

Finally, we establish the density of the tree conditioned on the number of extant species m, with a uniform prior for the time of origin of the tree. This density is derived just like the density of the higher level phylogeny with randomly selected subclades.

Theorem 4 *The probability density of the higher level phylogeny $\mathcal{T}$ conditioned on m is*

$$p(\mathcal{T}|m) = m\lambda^{n-1}\frac{p_1(x_1)^2}{(1-p_0(x_1))}\left(\frac{\lambda}{\mu}p_0(x_{\text{cut}})\right)^{m-n}\prod_{i=2}^{n-1}p_1(x_i). \quad (6)$$

It should be noted that the probability densities of the strict higher level phylogenies, where a cutoff time x_{cut} is used to define subclades, do not depend on the species numbers of the individual subclades $m_1, \ldots, m_n$, but only on the total number of species $m = \sum_{i=1}^{n} m_i$. In fact, it had already been shown by Farris (1976) (see also Nee et al. 1994) that given n lineages give rise to m lineages, all possible distributions of progeny for the n lineages are equally likely. This result was used, for example, in Purvis et al. (1995) for testing the rate homogeneity assumption in higher level phylogenies.

Simulations.—In order to evaluate the performance of the likelihood functions that we presented above, we used them to estimate speciation and extinction rates from 4 sets of simulated reconstructed phylogenies. We simulated trees using TreeSim (Stadler 2011b) with 4500 extant species, and speciation rate $\lambda = 0.1$, and 3 different extinction rates: $\mu = 0.025$, $\mu = 0.05$, and $\mu = 0.075$. The fourth set of simulated phylogenies had $\lambda = 0.1$ and $\mu = 0.05$ but 1000 instead of 4500 extant species. For each of these 4 parameter combinations, we simulated 100 trees.

For each simulated phylogeny, we first estimated the speciation and extinction rates from the complete tree, that is, using all branching times and using likelihood equations from Gernhard (2008). Second, we randomly selected 25 edges from the complete tree defining non-overlapping subclades, collapsed the lineages descending from each edge, and estimated the speciation and extinction rates from the species numbers in the 25 subclades and the remaining branching times using Equations (1–3). Third, we applied strict cutoff times x_{cut} so as to create strict higher level phylogenies and estimated the speciation and extinction rate using Equations (4–6).

Finally, we also estimated the speciation and extinction rates from the strict higher level phylogenies with the inappropriate method, that is, assuming randomly selected subclades (i.e., using Equations [1–3]). Vice versa, we estimated the speciation and extinction rates from the random higher level phylogenies with the inappropriate method, that is, assuming a strict cutoff (at the time of the youngest stem age) using Equations (4–6).

Because results obtained under the 3 conditions (time of origin, time of first split, and number of extant species) are very similar, we show only results conditioned on the number of species.

Results

Simulations

We obtained accurate estimates of speciation and extinction rates from the simulated higher level phylogenies (Fig. 2). When subclades are arbitrarily selected, estimation precision decreases with the number of subclades, as fewer branching times remain when more species are included in subclades. For the trees we simulated, with 1000 or 4500 species, choosing 25 subclades still provides reliable estimates (Fig. 2). Similarly, estimation of speciation and extinction rates from strict higher level phylogenies becomes less precise when the cutoff time that defines the higher taxa approaches the origin of the tree, as more lineages are included in subclades, and fewer branching times remain. For the trees we simulated, when x_{cut} is placed at 75% of x_0, the rates cannot be reliably estimated any more.

When speciation and extinction rates are estimated from strict higher level phylogenies inappropriately assuming random subclade selection, the rates will be underestimated. Figure 2 shows that, in particular, the extinction rates are estimated to be zero, and the speciation rate approximates the net speciation rate ($\lambda - \mu$).

The other way round, estimation from a higher level phylogeny with randomly selected subclades erroneously assuming a strict higher level phylogeny yields only slightly biased results (the rates are slightly overestimated).

Sensitivity of Estimates Toward Cutoff Time

We investigated the sensitivity of the parameter estimates toward different cutoff times using a phylogeny of 2 tips and $x_1 = 1$, using Equation (5). As expected, if the cutoff is more recent, then for $m > 2$, the turnover is estimated to be high, as many recent speciation events correspond to a strong pull of the present effect (Fig. 3). Since confidence intervals for recent cutoffs are contained within confidence intervals for early cutoffs, we suggest to use an early cutoff for cases when the cutoff time is not known accurately.

Application to Empirical Data

As the simulations confirmed that the probability density functions return reliable results if their assumptions are met, we used Equations (1–6) to estimate speciation and extinction rates for phylogenies that we retrieved from the literature. The 2 chosen examples (birds and mammals) consist of more species than the number of species in our simulations, thus the empirical results should be at least as reliable as the simulation results, if the models are appropriate.

In order to illustrate estimation of speciation and extinction rates from higher level phylogenies, we used the phylogeny of 23 avian orders from Sibley and

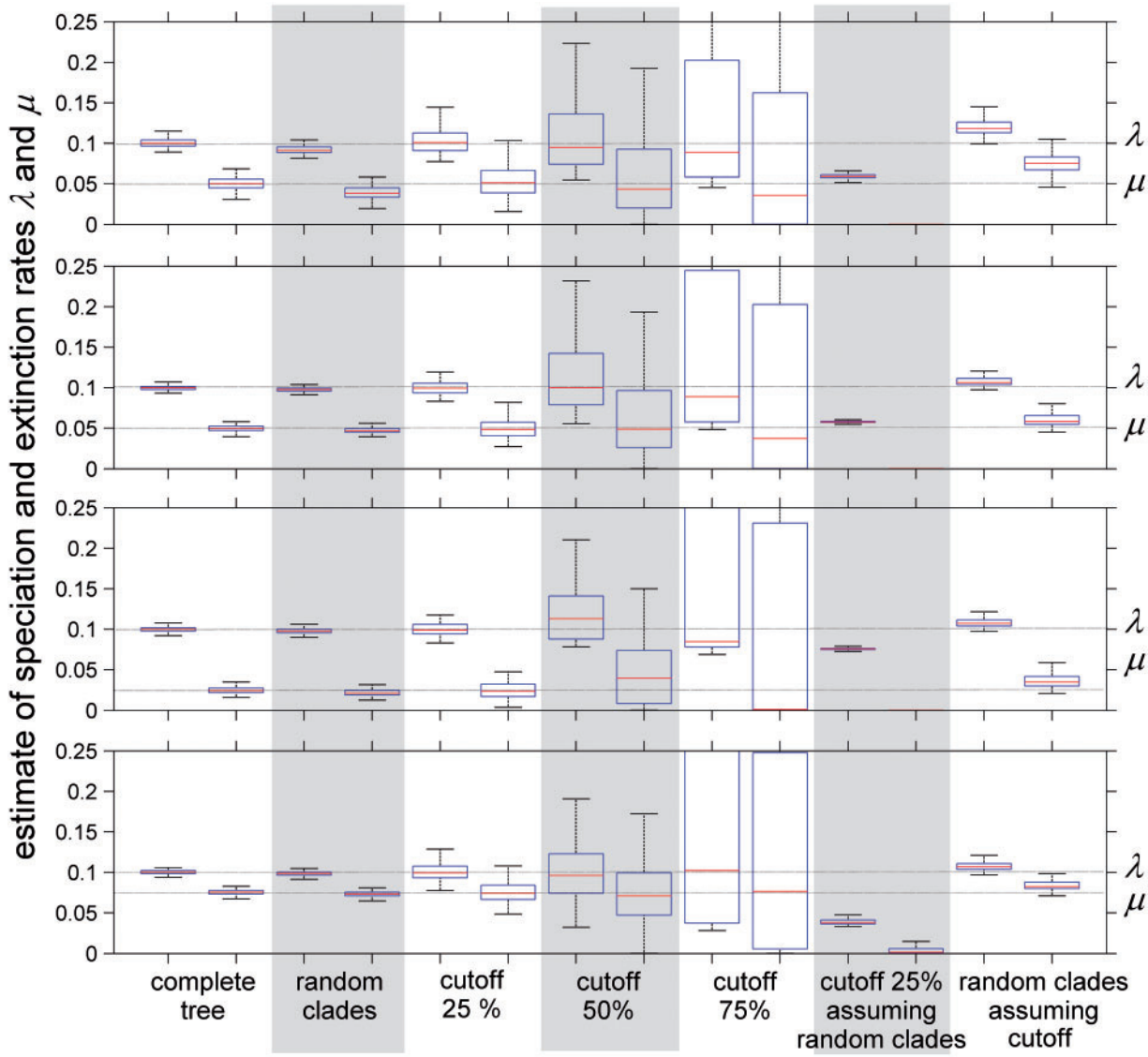

FIGURE 2. Estimates of speciation and extinction rates from 100 simulated higher level phylogenies on 1000 (upper panel) and 4500 (other panels) extant species. True values of the speciation and extinction rates are indicated by dotted lines. Boxes show median and quartiles, whiskers extend to the most extreme values within 1.5 times the interquartile range. Complete tree: estimates from the complete reconstructed species-level phylogenies. Random clades: estimates from the random higher level phylogenies with 25 randomly selected non-overlapping subclades. Cutoff: estimates from strict higher level phylogenies with cutoff time at 25% (respectively 50% and 75%) of x_0. Cutoff 25% assuming random clades: estimates obtained from the strict higher level phylogenies erroneously assuming randomly selected subclades. Random clades assuming cutoff: estimates obtained from the random higher level phylogenies erroneously assuming a strict cutoff at the youngest stem age.

Ahlquist (1990) (Supplementary Fig. S1), where orders are defined with a genetic distance of 20 units, the first split in the avian tree being at 28 units. Assuming the first split occurred $\approx$135 Ma, avian orders are then defined by a cutoff at $x_{\text{cut}} = 96.43$ Ma. We also estimated speciation and extinction rates from the same phylogeny at the family level (Sibley and Ahlquist 1990). This phylogeny, where families are defined with a genetic distance of 9 units, corresponding to $\approx$43.39 Ma, contains 135 families (Supplementary Fig. S2). (We strictly enforced $x_{\text{cut}} =$ 43.39 by lumping *Glareolidae* + *Laridae* and *Vireonidae* + *Corvidae*). Because we do not know the time of origin of the avian phylogeny, we could not obtain results conditioned on x_0. The estimates of speciation and extinction rates differ only slightly depending on whether we condition on the number of species or on the timing of the first split. However, estimates from the family-level tree are much higher than estimates from the order-level phylogeny (Table 1), with the family tree confidence region being much smaller (Fig. 4). Assuming random subclade selection instead of the strict subclade selection leads to very different rate estimates and confidence intervals, in particular to zero estimates of the extinction rate for avian orders.

We also estimated speciation and extinction rates from the almost complete species-level phylogeny of

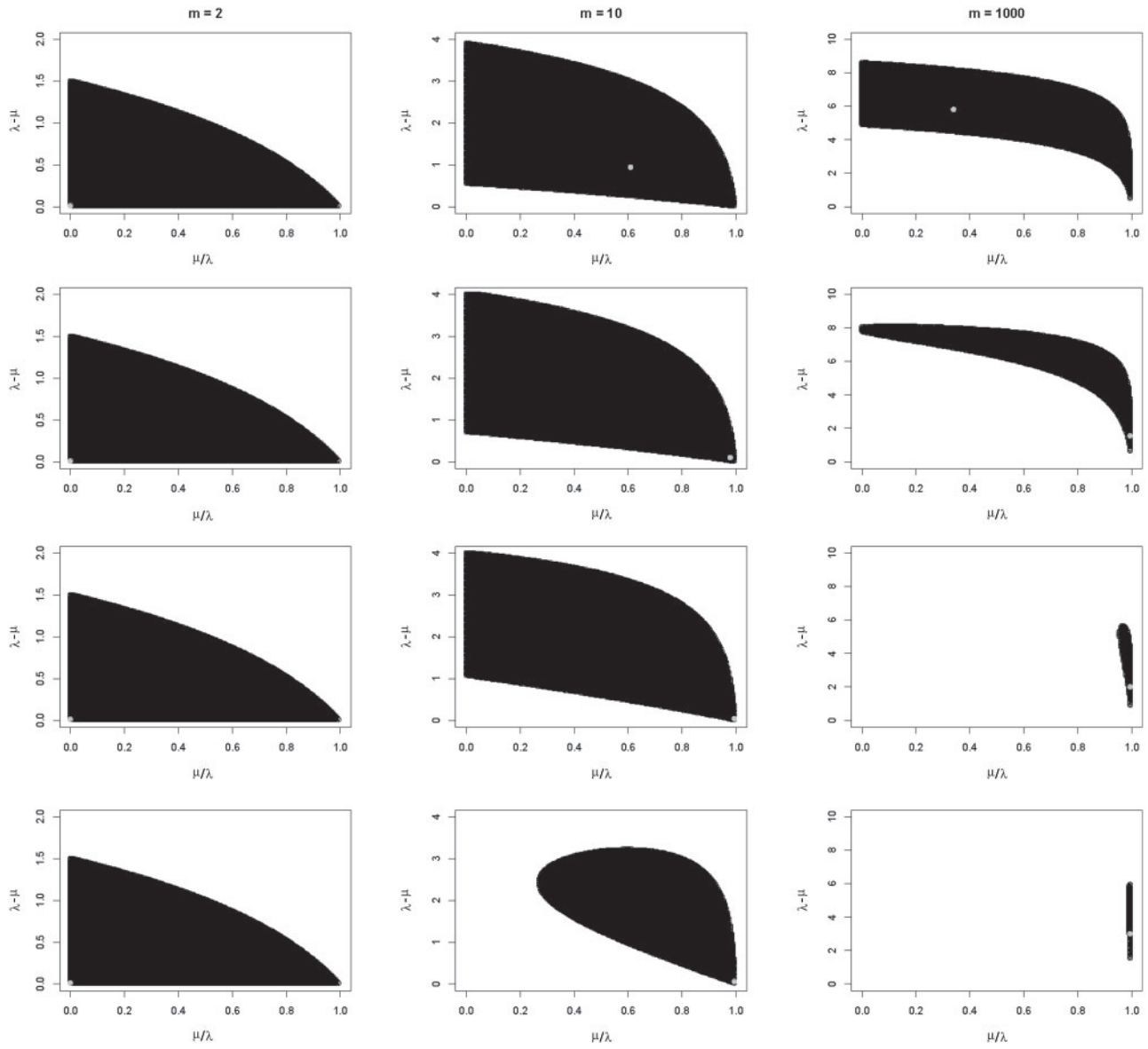

FIGURE 3. Maximum-likelihood speciation and extinction rate estimates (gray points) together with the 95% confidence regions (black) from a tree with 2 tips, $x_1 = 1$ and $m = 2, 10, 1000$ (columns from left to right). Rows correspond to a cutoff at $1, 0.75, 0.5, 0.25$ timesteps in the past (from top to bottom). We conditioned on the time of the mrca.

mammals by Bininda-Emonds et al. (2007). This tree contains almost all ≈4500 present-day mammalian species, but it is not fully resolved as it contains many polytomies. We used this tree to create strict higher level phylogenies, applying 2 cutoff times: $x_{\text{cut}} = 43.39$ Ma as for the avian family-level analysis and $x_{\text{cut}} = 68$ Ma (as also used for mammalian orders, see below). With these cutoff times, most polytomies disappear in subclades (the remaining polytomies can be interpreted as a binary tree with zero edge lengths at the polytomies). We estimated a speciation rate of ≈0.17 and an extinction rate of ≈0.12 with both cutoff times (Table 2). As expected, the confidence region for the tree with $x_{\text{cut}} = 43.39$ is almost fully contained within the confidence region for the tree with $x_{\text{cut}} = 68$ (Fig. 4). Again the analysis assuming a random subclade selection yields very different results, in particular zero estimates for the extinction rates (Fig. 4).

We also estimated speciation and extinction rates from the phylogeny of 16 mammalian orders presented by Paradis (2003). These mammalian orders were not defined with a strict cutoff, but probably also not using random subclade selection (discussed below), but we report estimates under both assumptions (Table 2). Then, after analyzing the tree as originally presented (Paradis 2003), we lumped orders with exceptionally recent splitting times so as to make the phylogeny more appropriate for analysis assuming a strict higher level phylogeny with $x_{\text{cut}} = 68$. In either case, estimates of

TABLE 1. Maximum-likelihood estimates of speciation and extinction rates λ and μ (in units: per million years) from the avian phylogeny by Sibley and Ahlquist (1990) at the order and family level (cutoff at 96.43 and 43.4 Ma, respectively)

Birds					
		Orders		Families	
Method	Condition	λ_{ML}	μ_{ML}	λ_{ML}	μ_{ML}
Cutoff	Given x_1	0.102	0.046	0.764	0.733
	Given # species	0.082	0.023	0.755	0.724
No cutoff	Given x_1	0.059	0.000	0.346	0.303
	Given # species	0.059	0.000	0.342	0.298
Paradis		0.060	0.000	0.083	0.000

Notes: Parameters were estimated assuming a cutoff, and (incorrectly) assuming no cutoff but random subclade selection. Estimates are shown conditioned on the time since the first speciation event x_1 and on the number of extant species. For comparison, we present the results of Paradis (2003), which are based on the same data.

extinction rates are zero for this phylogeny, as also suggested by Paradis (2003).

DISCUSSION

This study shows that speciation and extinction rates can be estimated from higher level phylogenies using likelihood methods. Intuitively, the method infers macroevolutionary rates as follows. From the rate of lineage accumulation in the tree prior to the times of collapsing clades to single tips, the diversification rate $\lambda-\mu$ is estimated (in expectation, these early lineages accumulate with rate $\lambda-\mu$ (Harvey et al. 1994)). In order to separate speciation and extinction rates, the size of the collapsed clades is considered. The clade sizes are slightly bigger than expected under a species accumulation rate $\lambda-\mu$, due to the pull-of-the-present effect (Harvey et al. 1994): in expectation lineages in the most recent past accumulate with rate λ, instead of $\lambda-\mu$. Quantifying by how much bigger the collapsed clades are than expected if lineages accumulate with rate $\lambda-\mu$ provides λ and μ.

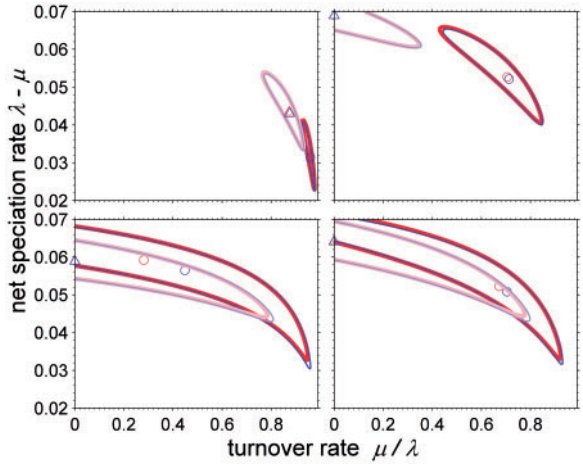

FIGURE 4. Maximum-likelihood estimates of speciation and extinction rates (points) with corresponding 95% confidence regions (lines). The left panels show estimates from the avian phylogeny by Sibley and Ahlquist at the family (upper panel, cutoff 43.4 Ma) and order (lower panel, cutoff 96.4 Ma) level. The panels at the right show estimates from the mammal phylogeny by Bininda-Emonds et al. with a cutoff at 43.4 Ma (upper panel) and 68 Ma (lower panel). We conditioned on the time of the mrca (blue), and on the number of species (red), assuming a cut-off (circles), and, inappropriately for these data but to illustrate the difference, no cutoff (triangles, and confidence regions in lighter shading). The numerical values of the maximum-likelihood estimates are stated in Table 1.

This study shows that accounting for the way in which higher taxa were defined strongly affects the estimates of speciation and extinction rates. For example, a higher taxon may have been defined as a subclade which has a crown age of <0.25 timesteps in the past, and a stem age of >0.25 timesteps in the past. If we acknowledge this higher taxon definition (i.e., use scenario (ii)), we obtain correct speciation and extinction rate estimates. However, if we use scenario (i) for a phylogeny defined under (ii), we essentially neglect the information that the crown age is younger than 0.25 timesteps in the past, and we only acknowledge the stem age (length of the lineage representing the higher taxa).

TABLE 2. Maximum-likelihood estimates of speciation and extinction rates λ and μ (in units: per million years) for mammals from 2 phylogenies

Mammals									
		P_{60}		P_{68}		$BE_{43.4}$		BE_{68}	
Method	Condition	λ_{ML}	μ_{ML}	λ_{ML}	μ_{ML}	λ_{ML}	μ_{ML}	λ_{ML}	μ_{ML}
Cutoff	Given mrca	0.171	0.090	0.086	0.000	0.181	0.129	0.172	0.121
	Given # species	0.127	0.039	0.086	0.000	0.178	0.126	0.160	0.107
No cutoff	Given mrca	0.079	0.000	0.080	0.000	0.069	0.000	0.064	0.000
	Given # species	0.079	0.000	0.080	0.000	0.069	0.000	0.064	0.000
Paradis		0.080	0.000	–	–	–	–	–	–

Notes: Estimates are shown conditioned on the time since the first speciation event x_1 and conditioned on the number of extant species. Estimates in column P_{60} are from the phylogeny presented in Paradis (2003). For comparison, we present the results of Paradis (2003) in the bottom row. Estimates in the second column (P_{68}) were obtained from the same phylogeny with a cutoff at 68 Ma, thus lumping some subclades that are far younger than the rest, so as to make the assumption of a cutoff more realistic. Estimates in column $BE_{43.4}$ are from the species-level phylogeny by Bininda-Emonds et al. (2007) with a cutoff applied at 43.4 (corresponding to avian families, Table 1), and at 68 Ma for column BE_{68}. Many polytomies in the tree by Bininda-Emonds et al. disappear when a cutoff is applied.

Using simulations, we show that neglecting the crown age information results in very biased estimates (Fig. 2, second right panel), in particular extinction rates are underestimated; typically they are estimated to be close to zero. Vice versa, if higher taxa were truly defined randomly (scenario (i)), then analysis under scenario (ii) still provides reasonable estimates (Fig. 2, right panel), meaning that scenario (ii) seems more robust toward model misspecification. Further, simulations reveal that a tree on as little as 25 higher taxa can provide accurate speciation and extinction rate estimates. This should not be confused with a tree on 25 species, from which the parameter estimates will have a very large confidence region.

Regarding scenario (i), we want to emphasize that one obtains the same likelihood equations as Equations (1–3) (except for the terms in s that do not affect estimates) if a deterministic subclade selection criterion is applied, for example, each edge with precisely N extant species descendants induces a subclade. Although it is unlikely that in reality higher taxa are ever defined in this way (e.g., the bird family sizes differ greatly in species numbers, from a handful to several thousands), it remains to be investigated which higher taxon definitions besides the random subclade selection give rise to likelihood equations (1–3).

Previous methods estimating speciation and extinction rates from higher level phylogenies (Paradis 2003; Rabosky et al. 2007; Alfaro et al. 2009; FitzJohn et al. 2009) essentially correspond to our scenario (i): those previous likelihood functions are a product of the likelihood of the branching structure (denoted by L_P in the previous papers) and the likelihood of the sizes of the higher taxa (denoted by L_T in the previous papers), which is essentially equivalent to the likelihood under our scenario (i) with $L_P = \lambda^{n-1} \prod_{i=0}^{n-1} \frac{p_1(x_i)}{p_1(t_{i+1})}$ and $L_T = \prod_{i=1}^{n} p_{m_i}(t_i)$ (up to conditioning the likelihood on different aspects such as survival and/or size of clades). Note that $s^{n-k}(1-s)^{n-1}$ only appears in our formulation, due to explicitly assuming a random subclade selection; however, as this product is independent of λ and μ, it is neglected in likelihood inference. Due to the equivalence of our scenario (i) equations and the equations in Paradis (2003), Rabosky et al. (2007), Alfaro et al. (2009), and FitzJohn et al. (2009), the previous approaches, which did not explicitly state how the selection of higher taxa was modeled, implicitly assumed our scenario (i).

There are few phylogenies published for which the particular higher taxa definition is known. One is the avian family-and order-level phylogeny by Sibley and Ahlquist (1990), who defined higher taxa strictly based on genetic distance as in our scenario (ii). Therefore, we used this tree to obtain rate estimates using our 2 higher level phylogeny definitions (i) and (ii), again highlighting the importance of choosing (i) or (ii). Our method for random subclades (i) as well as Paradis (2003) estimated a zero extinction rate for bird orders, while the method assuming a cutoff (ii) gave nonzero estimates for the extinction rate of bird orders as well as families. These differences between extinction rate estimates are similar to the biases seen in the simulations. Note that the other previous methods (Rabosky et al. 2007; Alfaro et al. 2009; FitzJohn et al. 2009) are extensions of Paradis (2003), allowing for varying rates instead of the crBDP. As we are only considering the crBDP here, we only compared our method with the previous crBDP-based method by Paradis (2003), but more complex models may be biased as well.

Our estimates of avian speciation and extinction rates based on the family level match the estimates for passerine birds well (Ricklefs 2003), with a turnover of at least 0.9 and expected lifetime of around 1–1.4 myr. However, our estimates of avian speciation and extinction rates differed substantially between the family-and order-level phylogenies (Table 1 and Fig. 4). As the density functions we presented perform well on simulated phylogenies, the reason for this discrepancy must be in the data. As Sibley and Ahlquist (1990) strictly enforced a genetic-distance-based definition of higher taxa, we can exclude the possibility that the higher taxa were selected in some other way. There remain 2 possibilities explaining the discrepancy (Paradis 2003): first, the order tree might be too small to produce accurate results. The family tree is much larger and thus produces more accurate parameter estimates, which is indeed observed when considering the 95% confidence regions (Fig. 4). However, the confidence region of the large tree is only partially contained in the confidence region of the small tree. Second, the crBDP model of evolutionary diversification may not be appropriate for the avian phylogenies. It is possible that speciation and extinction rates have not been constant across the tree (Ricklefs 2006), for example, due to differences between taxa (Bokma 2003), or have not been constant over time, for example, due to density-dependent diversification (Rabosky and Lovette 2008) or mass extinctions. In particular, Paradis (2003) suggests that speciation rate estimates for families might be inflated due to many large passerine bird families, while the heterogeneity in rates might be lessened among orders. Our analyses indicate that hyperdiverse clades induce inflated turnover, which again would be lessened among orders.

For the mammals, the estimates based on the species phylogeny (Bininda-Emonds et al. 2007) do not differ much between the 2 cutoff times, with the confidence region for the more recent cutoff being contained within the confidence region for the earlier cutoff (Fig. 4). However, these estimates do differ from the speciation and extinction rates that we estimated from the phylogeny of 16 mammalian orders from Paradis (2003). That difference may be simply due to differences between the phylogenies; further we expect less accurate parameter estimates in the small order phylogeny compared with the large species phylogeny. Interestingly, the extinction rate is estimated to be zero if we use the full species phylogeny without applying a cutoff (Stadler 2011a). No extinction is estimated based on the full species phylogeny since the species lineages-through-time plot flattens out in the very recent past,

while extinction would yield a turn-up. The recent flattening may be a bias introduced by the polytomies though (as also suggested by the analysis in Bokma (2008)), as an unresolved split is dated at the time of the polytomy, while in fact it may be much younger. Collapsing the polytomies into subclades removes this bias.

Our analyses of the bird and mammal data sets reveal large confidence regions (Fig. 4), highlighting the importance of not only focusing on the point estimates but also on the confidence regions. Typically, the diversification rate $\lambda-\mu$ is estimated with higher confidence than the turnover μ/λ: the diversification rate is informed already by the number of species and their stem/crown age (Magallon and Sanderson 2001) which is provided in higher level phylogenies, while turnover requires information about the relative timing of speciation events (Harvey et al. 1994) which is only partially available in higher level phylogenies. Interestingly, for the avian family phylogeny, we obtain a more confident estimate for the turnover than for the diversification rate. In this data set, turnover is estimated to be close to 1. This signal may be real, or may reflect rate heterogeneity: if a collapsed clade has an increased diversification rate, the pull of the present effect is bigger, yielding a higher turnover.

The future development of novel likelihood approaches acknowledging higher taxa definitions as introduced in this article, together with allowing for varying speciation and extinction rates (Rabosky et al. 2007; Rabosky and Lovette 2008; Alfaro et al. 2009; Crisp and Cook 2009; Rabosky 2009; Etienne et al. 2012; Morlon et al. 2011; Stadler 2011a), will most likely allow us to resolve the discussed discrepancies of different bird and mammal speciation and extinction rate estimates. Moreover, combining the phylogenetic trees with fossil data, requiring an extension of the methods presented here, will allow us to obtain tighter estimates for the turnover.

In addition to the considered bird phylogenies, we expect to obtain a growing number of bacterial phylogenies for which scenario (ii) will apply (Pommier et al. 2009) (note though that it may be hard to determine the precise number of species within a bacterial subclade). Most other published higher level phylogenies probably contain subclades that are defined in an intractable fashion. Currently, we do not have methods to estimate speciation and extinction rates from such phylogenies. However, it will often be possible to modify such phylogenies by applying a cutoff where the phylogeny still contains all lineages, that is, before the oldest subclade. We used that strategy when we lumped *Glareolidae* + *Laridae* and *Vireonidae* + *Corvidae* in the family-level phylogeny of birds (Sibley and Ahlquist 1990), and when we applied cutoff times to the species phylogeny of mammals (Bininda-Emonds et al. 2007), which eliminated unresolved polytomies. By applying a cutoff, we loose information (namely the branching times in the phylogeny after the cutoff) but it allows us to estimate speciation and extinction rates in cases where otherwise no appropriate likelihood functions would be available. Thus by applying a cutoff, we can avoid biases due to unknown data selection and/or polytomies.

In summary, when estimating speciation and extinction rates from incomplete phylogenies, it is very important to take into account how species were sampled. In the case of higher level phylogenies, in order to estimate speciation and extinction rates, one needs likelihood functions that account for the definition of higher taxa. We presented likelihood equations for 2 possible definitions of higher taxa: random selection of subclades (which turns out to be equivalent to a specific deterministic subclade selection) and a strict cutoff. Most published phylogenies probably fall into neither of theses categories, and it has to be decided for the particular data sets which category is more appropriate. We emphasize that our simulations revealed a smaller bias in rate estimates if the method assuming a strict cutoff was applied to random higher level phylogenies compared with the method assuming randomly selected subclades was applied to strict higher level phylogenies. Thus, the likelihood functions for the strict criterion are more robust toward model violations. If the precise cutoff in a phylogeny is not known, we suggest to use an earlier rather than later cutoff time, as in the investigated cases, confidence intervals for earlier cutoffs contain the confidence intervals for later cutoffs. Furthermore, many phylogenies can be modified by application of a cutoff after which they can be used to estimate speciation and extinction rates using the equations for a strict cutoff.

Supplementary Information

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.b0c8470m.

Funding

This work was supported by the Swiss National Science Foundation (to T.S.) and the Swedish research council (to F.B.).

Acknowledgments

We want to thank the editor, associate editor, Michael Blum, Luke Harmon, and four anonymous referees for stimulating criticism on a previous version of this article.

References

Aldous D., Popovic L. 2005. A critical branching process model for biodiversity. Adv. Appl. Probab. 37:1094–1115.

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

Bininda-Emonds O.R., Cardillo M., Jones K.E., MacPhee R.D., Beck R.M., Grenyer R., Price S.A., Vos R.A., Gittleman J.L., Purvis A. 2007. The delayed rise of present-day mammals. Nature 446:507–512.
Bokma F. 2003. Testing for equal rates of cladogenesis in diverse taxa. Evolution 57:2469–2474.
Bokma F. 2008. Bayesian estimation of speciation and extinction probabilities from (in) complete phylogenies. Evolution 62: 2441–2445.
Brock C., Harmon L., Alfaro M. 2011. Testing for temporal variation in diversification rates when sampling is incomplete and nonrandom. Syst. Biol. 60:410–419.
Crisp M., Cook L. 2009. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. Evolution 63:2257–2265.
Cusimano N., Renner S. 2010. Slowdowns in diversification rates from real phylogenies may not be real. Syst. Biol. 59:458–464.
Drummond A., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.
Etienne R., Haegeman B., Stadler T., Aze T., Pearson P., Purvis A., Phillimore A. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. Roy. Soc. B. 279:1300–1309.
Farris J. 1976. Expected asymmetry of phylogenetic trees. Syst. Biol. 25:196–198.
Feller W. 1968. An introduction to probability theory and its applications. 3rd ed. Vol. I. New York: John Wiley & Sons Inc.
FitzJohn R., Maddison W., Otto S. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595.
Foote M., Hunter J., Janis C., Sepkoski J. Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. Science 283:1310.
Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.
Harvey P.H., May R.M., Nee S. 1994. Phylogenies without fossils. Evolution 48:523–529.
Höhna S., Stadler T., Ronquist F., Britton T. 2011. Inferring speciation and extinction rates under different sampling schemes. Mol. Biol. Evol. 28:2577–2589.
Kendall D.G. 1948a. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika 35:6–15.
Kendall D.G. 1948b. On the generalized "birth-and-death" process. Ann. Math. Statist. 19:1–15.
Kendall D.G. 1949. Stochastic processes and population growth. J. Roy. Statist. Soc. Ser. B 11:230–264.
Kubo T., Iwasa Y. 1995. Inferring the rates of branching and extinction from molecular phylogenies. Evolution 49:694–704.
Magallon S., Sanderson M. 2001. Absolute diversification rates in angiosperm clades. Evolution 55:1762–1780.
Morlon H., Parsons T., Plotkin J. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108:16327–16332.
Nee S.C., May R.M., Harvey P. 1994. The reconstructed evolutionary process. Philos. Trans. Roy. Soc. Lond. Ser. B 344:305–311.
Paradis E. 2003. Analysis of diversification: combining phylogenetic and taxonomic data. Proc. Roy. Soc. B 270:2499.
Pommier T., Canbäck B., Lundberg P., Hagström Å., Tunlid A. 2009. Rami: a tool for identification and characterization of phylogenetic clusters in microbial communities. Bioinformatics 25:736–742.
Purvis A., Nee S., Harvey P. 1995. Macroevolutionary inferences from primate phylogeny. Proc. Roy. Soc. Lond. Ser. B 260:329–333.
Rabosky D. 2009. Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. Syst. Biol. 58:629.
Rabosky D., Donnellan S., Talaba A., Lovette I. 2007. Exceptional among-lineage variation in diversification rates during the radiation of Australia's most diverse vertebrate clade. Proc. Roy. Soc. B 274:2915.
Rabosky D., Lovette I. 2008. Density-dependent diversification in North American wood warblers. Proc. Roy. Soc. B 275:2363.
Raup D.M., Gould S.J., Schopf T.J.M., Simberloff D.S. 1973. Stochastic models of phylogeny and the evolution of diversity. J. Geol. 81: 449–452.
Ricklefs R. 2003. Global diversification rates of passerine birds. Proc. Roy. Soc. B 270:2285.
Ricklefs R. 2006. Global variation in the diversification rate of passerine birds. Ecology 87:2468–2478.
Sibley C., Ahlquist J. 1990. Phylogeny and classification of birds: a study in molecular evolution. New Haven: Yale University Press.
Stadler T. 2009. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.
Stadler T. 2011a. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. USA 108:6187–6192.
Stadler T. 2011b. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–684.
Yang Z. 2007. Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 17:717–724.