BIOLOGY
LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: Stadler T, Smrckova J. 2016 Estimating shifts in diversification rates based on higher-level phylogenies. *Biol. Lett.* **12**: 20160273.
http://dx.doi.org/10.1098/rsbl.2016.0273

Received: 5 April 2016
Accepted: 9 September 2016

Subject Areas:
evolution, bioinformatics

Keywords:
phylogenetic tree, inference, speciation, extinction, incomplete sampling

Author for correspondence:
Tanja Stadler
e-mail: tanja.stadler@bsse.ethz.ch

An invited contribution to the special feature 'Biology of extinction: inferring events, patterns and processes'.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3473625.

THE ROYAL SOCIETY
PUBLISHING

Evolutionary biology

Estimating shifts in diversification rates based on higher-level phylogenies

Tanja Stadler[1,2] and Jana Smrckova[3]

[1]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland
[2]Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland
[3]Department of Zoology, Faculty of Science, University of South Bohemia, 37005 Ceske Budejovice, Czech Republic

TS, 0000-0001-6431-535X

Macroevolutionary studies recently shifted from only reconstructing the past state, i.e. the species phylogeny, to also infer the past speciation and extinction dynamics that gave rise to the phylogeny. Methods for estimating diversification dynamics are sensitive towards incomplete species sampling. We introduce a method to estimate time-dependent diversification rates from phylogenies where clades of a particular age are represented by only one sampled species. A popular example of this type of data is phylogenies on the genus- or family-level, i.e. phylogenies where one species per genus or family is included. We conduct a simulation study to validate our method in a maximum-likelihood framework. Further, this method has already been introduced into the Bayesian package MrBayes, which led to new insights into the evolution of Hymenoptera.

1. Introduction

A key goal in macroevolution is to identify changes in the rates of diversification and to find causal explanations for variations in the species diversity we observe today. It has been shown that species phylogenies based on only extant taxa and no extinct lineages can be used to infer both the speciation and extinction rates, and thus in particular the diversification rate defined as speciation rate minus extinction rate [1–3].

Many organismal clades contain a vast amount of species, which makes construction of complete species phylogenies an arduous task. Although for some species groups complete or near-complete phylogenies have been already inferred [4,5], many others are still available as phylogenies on a higher taxonomic level only, meaning that only one species per higher taxonomic unit such as genus or family is included [6–10]. We call such phylogenies 'higher-level phylogenies'.

Computational methods were developed to estimate diversification rates from higher-level phylogenies [7,11–14]. While Paradis [11] and Stadler & Bokma [14] have devised a method for estimating constant speciation and extinction rates in higher-level phylogenies, Rabosky *et al*. [12], Alfaro *et al*. [7] and Rabosky [15] refined these approaches by also allowing for the computation of speciation and extinction rate variation across clades. Additionally, rate variation through time may be induced by external variables, such as climate, break-up of continents, sea-level changes or development of key innovations or competition. In this paper, we introduce a framework allowing for the estimation of changes in diversification rates through time from higher-level phylogenies. Our mathematical equations derived here have been implemented into MrBayes, and used to infer a phylogeny of Hymenoptera through a Bayesian approach [16].

In what follows, we present a maximum-likelihood method to estimate changes in diversification (= speciation − extinction) rates and turnover (= extinction/speciation) for *higher-level phylogenies* where all phylogenetic relationships are resolved up to a certain point in time, and each clade, descending a lineage at that point in time, is collapsed to one tip. We show in a simulation approach that shifts in diversification rates can be estimated reliably based on our likelihood framework. We then explain how we can transform a phylogeny on the genus- or family-level into a higher-level phylogeny to analyse empirical data.

2. Methods

(a) Birth–death–skyline model

We extend the *constant rates birth–death process* (crBDP; [17–19]), to the *birth–death–skyline process*, following Stadler [20] and Stadler *et al*. [21].

The crBDP starts with a single lineage at time x_0 in the past (stem age) and gives birth to descendant lineages with a constant rate of speciation λ and lineages die with a constant rate of extinction μ. At the present time, the process is stopped. The *reconstructed phylogenetic tree* is acquired by pruning all lineages that went extinct.

The birth–death–skyline process generalizes the crBDP by allowing for rate changes through time: time between the present and x_0 is split up through $0 = t_0 < t_1 < t_2 < \ldots < t_m < x_0$. The present day is depicted as t_0. Speciation and extinction rates are constant (λ_i and μ_i) between t_i and t_{i+1}, and may differ arbitrarily between intervals. For estimating the parameters of the birth–death–skyline process, the probability of a reconstructed phylogeny given the parameters is provided in Stadler [20].

(b) Higher-level trees

A higher-level phylogeny is obtained from a complete species phylogeny by pruning the extant descendants of every lineage at time x_{cut} in the past to one sampled lineage together with the information on the number species represented by each sampled lineage (see e.g. fig. 1*d* in [14]). Let the branching times in the phylogeny be $x_1 > x_2 \ldots > x_{n-1}$, and let the number of species represented by tip i be n_i. We derived the probability of a higher-level phylogeny, in order to estimate maximum-likelihood diversification rates and turnover.

(c) Simulations

We investigate the accuracy of parameter estimation based on simulated higher-level phylogenies. First, we simulated species trees with 2000 tips under different diversification scenarios with one rate shift and disparate rates of speciation (resp. extinction) before and after the shift using the R package TreeSim [22]. We then collapsed clades to obtain higher-level phylogenies by pruning each clade at time x_{cut} before the present to one lineage terminating at the present with the information of number of species in the clade [23,24]. We have chosen times of x_{cut} corresponding to three quartiles of the age of the tree, i.e. 25%, 50% and 75%.

In the simulated trees, λ_{y} (y for young) denotes the speciation rate between the rate shift time and the present, and λ_{o} (o for old) is the speciation rate ancestral to the rate shift time. We set parameters specifying decelerating and accelerating diversification, namely birth rates were $\lambda_{\text{y}} = 0.5$, $\lambda_{\text{o}} = 1$ and $\lambda_{\text{y}} = 1$, $\lambda_{\text{o}} = 0.5$ for trees with decreasing and increasing diversification rate, respectively. Death rate was held constant with $\mu = 0.1$ in these simulations. For both the accelerating and decelerating scenario, we run simulations with both a time of the shift at 2 myr before present (BP) and 3.5 myr BP. Second, we increased the extinction rate to $\mu = 0.4$ (and thus increased turnover) under decreasing diversification ($\lambda_{\text{y}} = 0.5$, $\lambda_{\text{o}} = 1$). For this setting, which induces a recent diversification rate of 0.1, we chose a rate shift at 8 Mya. This setting is comparable to the simulation setting with a rate shift at 2 Mya and a recent diversification rate of 0.4, as both settings induce a lineage accumulation between the rate shift at the present of 0.8 (= 8 × 0.1 = 2 × 0.4); thus, these settings induce roughly the same number of lineages at the time of the rate shift. Last, we fixed the speciation rate to $\lambda = 1.0$ under decreasing diversification ($\mu_{\text{y}} = 0.6$, $\mu_{\text{o}} = 0.1$). We chose a rate shift time of 2 Mya, again ensuring a lineage accumulation since the rate shift of 0.8 (= 2 × 0.4). Thus, we have six different settings. For each set of parameters, 100 trees were produced.

For all trees, maximum-likelihood estimates of diversification rate $\lambda - \mu$ and turnover μ/λ before and after the rate shift, together with the time of the rate shifts, have been obtained using TreePar v. 3.3 function bd.shifts.optim [25] with the 'groups' option. This function employs equation (3.1) below.

The code required to perform our analyses is provided in the electronic supplementary material.

3. Results

(a) Probability of a higher-level phylogeny

The probability density of a higher-level phylogeny $\mathcal{T}$ with n tips is derived in the electronic supplementary material, theorem 3,

$$\mathbb{P}(\mathcal{T}|\lambda,\mu,h,x_1) = \frac{p_1(x_1|\lambda,\mu)^2}{(1-p_0(x_1|\lambda,\mu))^2}\prod_{i=2}^{n-1}\lambda_{x_i}p_1(x_i|\lambda,\mu)\prod_{i=1}^{n}\frac{p_{ni}(h|\lambda,\mu)}{p_1(h|\lambda,\mu)}, \tag{3.1}$$

where $p_k(t|\lambda,\mu)$ is the probability that a lineage at time t in the past has k descendants at present time 0, and λ_{x_i} is the speciation rate at time x_i. This is equivalent to (appendix, theorem 1)

$$\mathbb{P}(\mathcal{T}|\lambda,\mu,h,x_1) = \frac{1}{F(x_1)^2}\prod_{i=2}^{n-1}f(x_i)\left(1-\frac{1}{F(h)}\right)^{m-n},$$

with, for t in $(t_i, t_{i+1}]$, and re-defining $t_{i+1} := t$ for convenient notation

$$\left.\begin{aligned}
F(t) &= 1+\sum_{k=0}^{i}G(t_k),\\
G(t_k) &= \frac{\lambda_k}{\lambda_k-\mu_k}\left(e^{(\lambda_k-\mu_k)(t_{k+1}-t_k)}-1\right)\left(e^{\sum_{j=0}^{k-1}(\lambda_j-\mu_j)(t_{j+1}-t_j)}\right),\\
f(t) &= \frac{F'(t)}{F(t)^2}\\
\text{and}\quad F'(t) &= \lambda_i e^{(\lambda_i-\mu_i)(t-t_i)}e^{\sum_{j=0}^{i-1}(\lambda_j-\mu_j)(t_{j+1}-t_j)},
\end{aligned}\right\}$$

i.e. we do not need to specify how many species belong to each tip, but we need to know only the total number of species m. In fact, even if we knew how many species belong to each tip, that would not improve parameter estimates.

(b) Diversification rate estimates

In all simulations, diversification rates, turnover and times of the shifts were estimated correctly when analysing fully sampled species-level phylogenies (figures 1 and 2; electronic supplementary material, figure S1–S4). With the complete

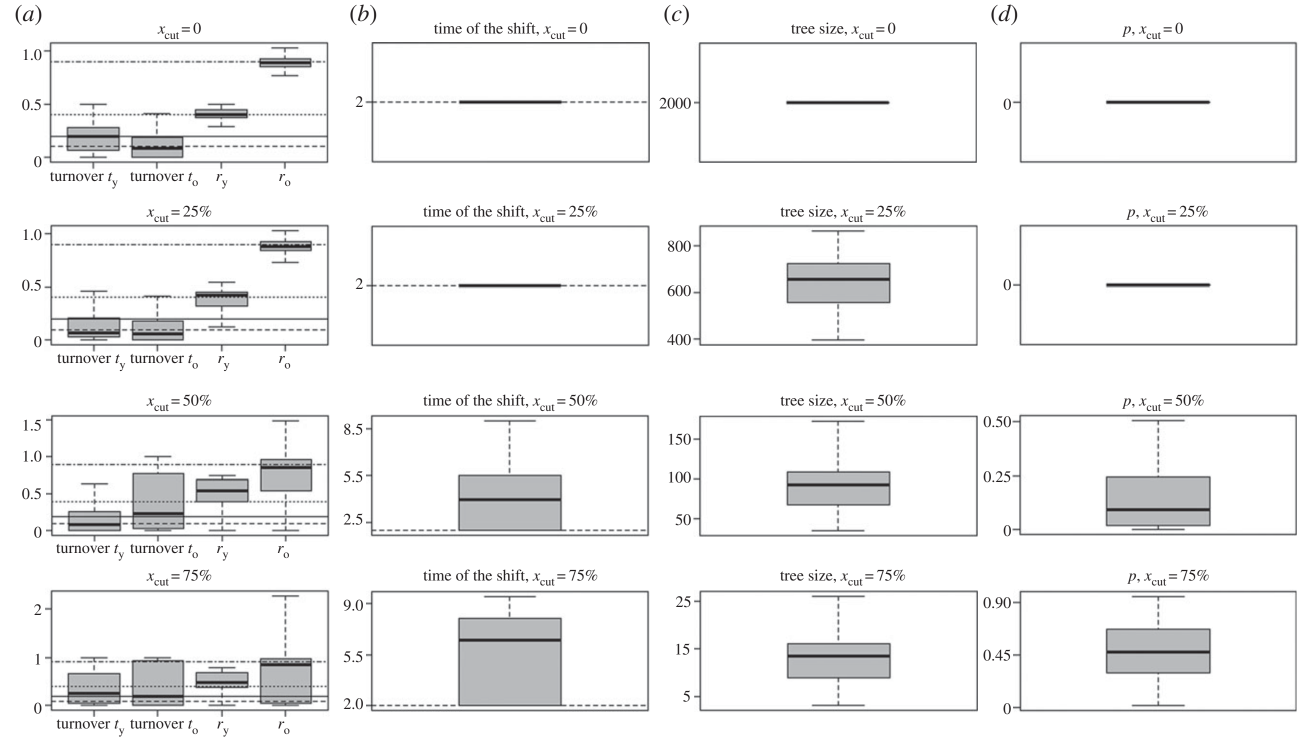

Figure 1. Results of a simulation study for trees with constant extinction rate (0.1) and decreasing diversification rate (from 0.9 to 0.4) with a rate shift at 2 myr BP. Central tendency in boxplots is median, vertical lines indicate original values used in simulations. Rows show results for trees with increasing x_{cut}. Plots in column (*a*) depict the estimated maximum-likelihood turnover and diversification rate parameters. Parameters between the present and the time of the shift and before the shift are denoted as t_n, r_y, and t_o and r_o, respectively. Plots in column (*b*) estimated maximum-likelihood shift times in myr, (*c*) sizes of the analysed trees and (*d*) *p*-values of the likelihood ratio test (using the Chi-squared distribution with 3 degrees of freedom, for speciation rate, extinction rate and shift time) comparing a single rate shift with a constant rate model.

trees, and x_{cut} at 25% with decreasing diversification and constant extinction rate, we consistently identified one significant rate shift. With increasing x_{cut}, we lost significance, and the times of the (potentially not significant) shifts were estimated to be older than initially simulated.

The better performance of constant speciation rate (figure 1) versus constant extinction rate (electronic supplementary material, figure S4) under the same diversification rates is most likely owing to the latter having a higher turnover, thus a more pronounced pull-of-the-past, and thus fewer lineages at the time of the rate shift causing less strong signal.

Simulation, in general, showed much better precision of estimates retrieved for incomplete trees with decelerating diversification than for accelerating scenario. In phylogenies with diversification rate slowing down towards the present, precise values were retrieved for x_{cut} up to 50% of tree age, whereas for trees with accelerating evolutionary rate, reliable estimates have been obtained for complete species-level phylogenies only. This result makes intuitive sense: most branching events in accelerating trees occur close to tips of the tree, i.e. most of the branching events are removed even for small x_{cut}, hence providing limited information for maximum-likelihood estimation.

4. Discussion

Not accounting for the sampling of higher-level taxa can lead to severe biases in parameter estimation, in particular an underestimation of extinction rate and turnover (see fig. 2 in [14]). We have formulated an inference framework for estimating shifts in diversification rate in higher-level phylogenies where all higher levels have the same age. Our simulations reveal that the method can estimate past diversification patterns from these higher-level phylogenies on extant species. Any phylogeny can be converted into a higher-level phylogeny by collapsing all clades descending from a lineage at time x_{cut}, where x_{cut} is a time point prior to which the phylogeny is fully resolved.

It has been shown that incorporating fossils will dramatically improve the quality of diversification rate estimates: in particular, the extinction rate can be estimated far better [26]. While Silvestro *et al*. [27] applied the model presented here to only fossil data, Zhang *et al*. [16] combined equation (3.1) of this paper with a fossilization process. Thus, coherent analysis of fossils and extant species data became possible. The resulting so-called fossilized birth–death process [28] has been implemented into MrBayes and a higher-level phylogeny with fossils of Hymenoptera has been inferred using the total-evidence-dating. The analyses of Zhang *et al*. [16] revealed that not accounting for the higher-level phylogeny structure, but assuming each species was sampled uniformly at random, has a drastic effect on tree inference. Thus, it is important to use appropriate diversification models not only for quantifying diversification rates, but also for inferring the phylogenies in the first place.

Here, we accounted for only one rate shift, even though our mathematical expression allows for an arbitrary number. The reason for this is that maximizing over multiple rate shift times is numerically very hard, and often leads the optimization tool to be stuck in local optima. In Stadler [20], a greedy approach for finding rate shifts was implemented, meaning the optimizer first searches for the best rate shift time, and with fixing this first rate shift time, it finds the

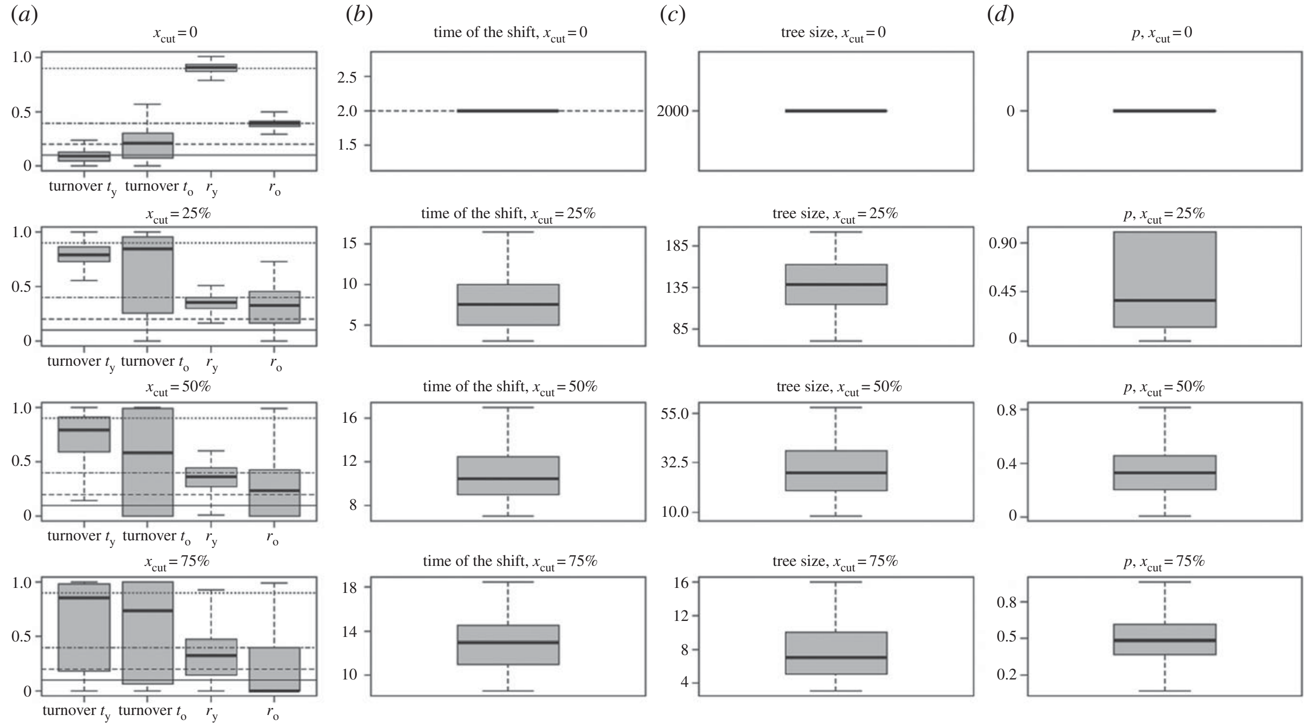

Figure 2. Results of a simulation study for trees with constant extinction rate (0.1) and increasing diversification rate (0.4–0.9) with a rate shift at 2 myr before present.

second best rate shift time, etc. However, with the implementation of our mathematical equation into the Bayesian framework MrBayes, we can directly infer the posterior distribution of rate shifts and thus do not rely on a greedy approximation.

The birth–death–skyline process still makes a number of limiting assumptions, in particular that all species are assumed to be indistinguishable, hence all have the same speciation and extinction rates and the same probabilities of being sampled at any given time. Accordingly, a birth–death–skyline model cannot allow us to explicitly test scenarios of heritable rates [29] or clade-dependent rates (Medusa: [7], BAMM: [15]), although it can be used as a null model for testing more complex patterns of diversification. It remains a future challenge to combine complex models of rate variation across clades and through time, for inference of diversification rates based on higher-level phylogenies with fossils.

Authors' contributions. Both authors contributed to the design of the paper, performed analyses and wrote the paper. Both authors approve the final version of the manuscript, and agree to be held accountable for the content therein.

Funding. This study is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

Competing interests. We declare we have no competing interests.

Acknowledgements. We thank Nick Matzke for providing the scripts for clade pruning.

References

1. Hey J. 1992 Using phylogenetic trees to study speciation and extinction. *Evolution* **46**, 627. (doi:10.2307/2409633)
2. Nee S, HolmesEC, May RM, Harvey PH. 1994 Extinction rates can be estimated from molecular phylogenies. *Phil. Trans. R. Soc. Lond. B* **344**, 77–82. (doi:10.1098/rstb.1994.0054)
3. Sanderson MJ, Donoghue MJ. 1996 Reconstructing shifts in diversification rates on phylogenetic trees. *Evolution* **11**, 15–20. (doi:10.1016/0169-5347(96)81059-7)
4. Bininda-Emonds ORP *et al.* 2007 The delayed rise of present-day mammals. *Nature* **446**, 507–512. (doi:10.1038/nature05634)
5. Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012 The global diversity of birds in space and time. *Nature* **491**, 444–448. (doi:10.1038/nature11631)
6. Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE. 2006 Phylogeny of the ants: diversification in the age of angiosperms. *Science* **312**, 101–104. (doi:10.1126/science.1124891)
7. Alfaro ME, Santini F *et al.* 2009 Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl Acad. Sci. USA* **106**, 13 410–13 414. (doi:10.1073/pnas.0811087106)
8. Bell CD, Soltis DE, Soltis PS. 2010 The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97**, 1296–1303. (doi:10.3732/ajb.0900346)
9. Couvreur TLP, Forest F, Baker WJ. 2011 Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. *BMC Biol.* **9**, 44. (doi:10.1186/1741-7007-9-44)
10. Meredith RW *et al.* 2011 Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521524. (doi:10.1126/science.1211028)
11. Paradis E. 2003 Analysis of diversification: combining phylogenetic and taxonomic data. *Proc. R. Soc. Lond. B* **270**, 2499–2505. (doi:10.1098/rspb.2003.2513)
12. Rabosky DL, Donellan SC, Talaba AL, Lovette IJ. 2007 Exeptional among-lineage variation in diversification rates during the radiation of Australia's most diverse vertebrate clade. *Proc. R. Soc. B* **274**, 2915–2923. (doi:10.1098/rspb.2007.0924)
13. Paradis E, Tedesco PE, Hugueny P. 2013 Quantifying variation in speciation and extinction rates with clade data. *Evolution* **67**, 3617–3627. (doi:10.1111/evo.12256)

14. Stadler T, Bokma F. 2013 Estimating speciation and extinction rates for phylogenies of higher taxa. *Syst. Biol*. **62**, 220–230. (doi:10.1093/sysbio/sys087)

15. Rabosky DL. 2014 Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* **9**, e89543. (doi:10.1371/journal.pone.0089543)

16. Zhang C, Stadler T, Klopfstein S, Heath TA, Ronquist F. 2016 Total-evidence dating under the fossilized birth–death process. *Syst. Biol*. **5**, 228–249. (doi:10.1093/sysbio/syv080)

17. Kendall DG. 1948 On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* **35**, 6–15. (doi:10.1093/biomet/35.1-2.6)

18. Feller W. 1968 *An introduction to probability theory and its applications*, 3rd edn, Vol. I. New York, NY: John Wiley & Sons Inc.

19. Nee SC, May RM, Harvey P. 1994 The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B* **344**, 305–311. (doi:10.1098/rstb.1994.0068)

20. Stadler T. 2011 Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl Acad. Sci. USA* **108**, 6187–6192. (doi:10.1073/pnas.1016876108)

21. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. 2012 Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* **110**, 228–233. (doi:10.1073/pnas.1207965110)

22. Stadler T. 2011 Simulating trees with fixed number of extant species. *Syst. Biol*. **60**, 676–684. (doi:10.1093/sysbio/syr029)

23. Matzke N. 2011 http://ib.berkeley.edu/courses/ib200b/scripts/_R_tree_functions_v1.R.

24. Matzke N. 2011 http://ib.berkeley.edu/courses/ib200b/scripts/_genericR_v1.R.

25. Stadler T. 2011 Inferring speciation and extinction processes from extant species data. *Proc. Natl Acad. Sci. USA* **108**, 16 145–16 146. (doi:10.1073/pnas.1113242108)

26. Didier G, Royer-Carenzi M, Laurin M. 2012 The reconstructed evolutionary process with the fossil record. *J. Theor. Biol*. **315**, 26–37. (doi:10.1016/j.jtbi.2012.08.046)

27. Silvestro D, Schnitzler J, Liow LH, Antonelli A, Salamin N. 2014 Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Syst. Biol*. **63**, 349–367. (doi:10.1093/sysbio/syu006)

28. Heath, TA, Huelsenbeck JP, Stadler T. 2014 The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proc. Natl Acad. Sci. USA* **111**, E2957–E2966. (doi:10.1073/pnas.1319091111)

29. Rabosky D. 2009 Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Syst. Biol*. **58**, 629. (doi:10.1093/sysbio/syp069)