

# Supplementary Text for: Quantifying Age-dependent Extinction from Species Phylogenies

HELEN K. ALEXANDER<sup>1</sup>, AMAURY LAMBERT<sup>2,3</sup>, AND TANJA STADLER<sup>4</sup>

<sup>1</sup>*Institute for Integrative Biology, ETH Zürich, 8092 Zürich, Switzerland;*

<sup>2</sup>*Laboratoire de Probabilités et Modèles Aléatoires CNRS UMR 7599, UPMC Univ Paris 06,  
Paris, France;*

<sup>3</sup> *Center for Interdisciplinary Research in Biology CNRS UMR 7241, Collège de France, Paris,  
France;*

<sup>4</sup>*Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland*

**Corresponding author:** Helen K. Alexander, Institute for Integrative Biology, ETH  
Zürich, CHN H.74, Universitätsstrasse 16, 8092 Zürich, Switzerland; E-mail:  
helen.alexander@env.ethz.ch

# EFFECT OF LIFETIME SHAPE PARAMETER $k$ ON EXTINCTION RATE AS A FUNCTION OF AGE

We want to prove that, given a gamma-distributed lifetime with shape parameter  $k$ , the extinction rate  $\mu(a)$  as a function of age  $a$  is decreasing if  $k < 1$  and increasing if  $k > 1$  (recall that it is constant in the exponential case when  $k = 1$ ). From Equation (2) in the main text, we get that  $\mu$  is the only solution to  $\mu'(a) = \mu(a)h(a)$  with terminal value  $\mu(\infty) = 1/\theta$ , where

$$h(a) := \frac{k-1}{a} - \frac{1}{\theta} + \mu(a).$$

Note that  $\mu'$  and  $h$  always have the same sign. Assume  $k > 1$ . Then note that if  $\mu$  is decreasing with  $a$ ,  $h$  is also decreasing with  $a$ . Further assume that there exists  $a_0 > 0$  such that  $\mu'(a_0) < 0$ , so that  $h(a_0) < 0$ . An additional consequence is that  $\mu(a_0) < 1/\theta$ , since  $k > 1$  and  $\mu(a_0) = h(a_0) - \frac{k-1}{a_0} + \frac{1}{\theta}$ . Since  $\mu(\infty) = 1/\theta$ , we cannot have  $\mu'(a) < 0$  for all  $a > a_0$ . As a consequence,  $a_1 := \inf\{a > a_0 : \mu'(a) > 0\}$  is finite, and by continuity, it satisfies  $\mu'(a_1) = 0$ . Now since  $\mu$  is decreasing on  $[a_0, a_1)$ ,  $h$  also is. This contradicts the fact that  $h(a_0) < 0$  while  $h(a_1) = 0$ , and shows that  $\mu'$  cannot take negative values when  $k > 1$ . The same reasoning shows that  $\mu'$  cannot take positive values when  $k < 1$ .

# NET DIVERSIFICATION RATE AS A FUNCTION OF SPECIATION RATE $\lambda$ AND LIFETIME SHAPE PARAMETER $k$

We prove the following two claims, given a gamma-distributed lifetime where  $k$  and  $\theta$  are assumed to vary in such a way that the mean lifetime  $\ell = k\theta$  remains constant: (i) for given  $k$ ,  $\eta$  increases (asymptotically linearly) with  $\lambda$  and (ii) for given  $\lambda$ ,  $\eta$  increases with  $k$ , approaching an asymptotic value corresponding to the case when all lifetimes are fixed equal to  $\ell$ .

First recall that the net diversification rate  $\eta$  is zero if and only if  $\lambda\ell = 1$ , and is otherwise the nonzero root of the Laplace exponent  $\psi$  defined as

$$\psi(y) := y - \lambda \int_0^\infty g(x)(1 - \exp(-yx))dx = y - \lambda \left[ 1 - \left( 1 + \frac{y\ell}{k} \right)^{-k} \right] \quad \text{for } y \geq 0.$$

Let us show (i) in the case of a general density  $g$ . It is easy to see that for any fixed density  $g$ ,  $\psi$  decreases as  $\lambda$  increases. Since  $\psi$  is a convex function with two roots, of which one is zero and the other is  $\eta$ , this ensures that  $\eta$  increases with  $\lambda$ . Let us show that this increase is asymptotically linear. Writing

$$\frac{\eta}{\lambda} = \int_0^\infty g(x)(1 - \exp(-\eta x))dx, \quad (\text{S1})$$

we see that  $\eta$  cannot tend to a finite value as  $\lambda \rightarrow \infty$ , so that  $\eta \rightarrow \infty$  as  $\lambda \rightarrow \infty$ . Now by dominated convergence, the last displayed equation yields  $\eta/\lambda \rightarrow \int_0^\infty g(x)dx = 1$ .

Let us now prove (ii). We assume that  $\lambda$  is fixed and we define the two-variable function  $\phi$  as

$$\phi(k, y) := y - \lambda \left[ 1 - \left( 1 + \frac{y\ell}{k} \right)^{-k} \right] \quad \text{for } y, k \geq 0.$$

In particular,  $\eta(k)$  is defined implicitly by  $\phi(k, \eta(k)) = 0$ . Since  $\phi$  is continuously differentiable, the implicit function theorem implies that  $\eta$  is also continuously differentiable and

$$\eta'(k) = -\frac{\partial \phi}{\partial k}(k, \eta(k)) / \frac{\partial \phi}{\partial y}(k, \eta(k)).$$

Let us compute these two partial derivatives. Elementary calculations show that

$$\frac{\partial \phi}{\partial k}(k, y) = \lambda \left( 1 + \frac{y\ell}{k} \right)^{-k} H \left( \frac{y\ell}{k} \right),$$

where  $H(z) = \frac{z}{1+z} - \ln(1+z)$ , and

$$\frac{\partial \phi}{\partial y}(k, y) = 1 - \lambda \ell \left( 1 + \frac{y\ell}{k} \right)^{-k-1}.$$

Using the fact that  $\phi(k, \eta(k)) = 0$ , we get

$$\eta'(k) = -H \left( \frac{\ell \eta(k)}{k} \right) \frac{(\lambda - \eta(k))(1 + \ell \eta(k)/k)}{J(k)},$$

where  $J(k) := 1 - \lambda \ell + \ell \eta(k)(k+1)/k$ . Now because  $\phi(k, \eta(k)) = 0$ , we have  $\eta(k) \leq \lambda$ , and it is easy to see that  $H(z)$  only takes negative values on  $z > 0$ , so that  $\eta'(k)$  is of the sign of  $J(k)$ .

Let us show that  $J$  cannot change sign. If  $J$  did change sign, there would be some value of  $k$  for which  $\eta(k) = \frac{k(\lambda\ell-1)}{\ell(k+1)}$ . Plugging this into  $\phi(k, \eta(k)) = 0$  would yield  $G(\lambda\ell) = 0$ , with

$$G(y) := \left( \frac{k+y}{k+1} \right)^{k+1} - y.$$

Now

$$G'(y) = \left( \frac{k+y}{k+1} \right)^k - 1$$

is positive iff  $y > 1$ , and since  $G(1) = 0$ ,  $G(y) = 0$  iff  $y = 1$ . Thus unless  $\lambda\ell = 1$  (in which case  $\eta(k) = 0$  for all  $k$ ), we cannot have  $G(\lambda\ell) = 0$ . As a consequence,  $J$  cannot change sign.

Now that we know that  $\eta(k)$  is monotonic, it makes sense to define  $Y := \ell\eta(\infty)$ . Letting  $k \rightarrow \infty$  in the equation  $\phi(k, \eta(k)) = 0$ , we find that  $Y$  is a solution to

$$\gamma Y - 1 + e^{-Y} = 0,$$

with  $\gamma := (\lambda\ell)^{-1}$ . By an application of the implicit function theorem to the last equality, where  $Y$  is thus seen as a function of  $\gamma$ , we get

$$Y'(\gamma) = \frac{1 - \gamma Y(\gamma) - \gamma}{Y(\gamma)} = \frac{\lambda\ell - \ell\eta(\infty) - 1}{\lambda\ell^2\eta(\infty)} = -\frac{J(\infty)}{\lambda\ell^2\eta(\infty)}.$$

But  $Y$  is easily seen (e.g., graphically) to be a decreasing function of  $\gamma$ , so that  $Y'(\gamma) < 0$  and  $J(\infty)$  has the sign of  $\eta(\infty)$ .

Recalling that  $J(k)$  cannot change sign, and that  $\eta'(k)$  has the sign of  $J(k)$ , we have the following result: In the supercritical case where  $\lambda\ell > 1$ ,  $\eta(k) > 0$  increases with  $k$  to the value  $\eta(\infty)$ , easily computable as  $\ell^{-1}Y$ , where  $Y$  is the unique positive solution to  $(\lambda\ell)^{-1}Y - 1 + e^{-Y} = 0$ . Similarly, in the subcritical case where  $\lambda\ell < 1$ ,  $\eta(k) < 0$  decreases with  $k$ , and in the critical case where  $\lambda\ell = 1$ ,  $\eta(k) = 0$  for all  $k$ .

## IDENTIFIABILITY OF THE GAMMA MODEL

It is known that the model of diversification with sampling at present and constant speciation and extinction rates is not identifiable (Stadler 2009). Here, we prove that when species lifetimes are no longer exponentially distributed, but gamma distributed (i.e., extinction rate is not constant but age-dependent), the model becomes identifiable. Let us state this in a more specific way.

**Theorem 1** *Consider two models of diversification with sampling probabilities of extant species  $p_1$  and  $p_2$  respectively, constant speciation rates  $\lambda_1$  and  $\lambda_2$  respectively, and where*

species lifetimes are gamma-distributed, with parameters  $(k_1, \theta_1)$  and  $(k_2, \theta_2)$ , respectively. If these two models induce the same distribution on reconstructed trees and if at least one of them is really age-dependent ( $k_1 \neq 1$  or  $k_2 \neq 1$ ), then  $(p_1, \lambda_1, k_1, \theta_1) = (p_2, \lambda_2, k_2, \theta_2)$ .

Put another way, a model with gamma distributed lifetimes is not identifiable if and only if the lifetimes are actually exponentially distributed. We will now prove the theorem.

To this end, we will first consider a generic age-dependent model of diversification with sampling probability  $p$ , speciation rate  $\lambda$ , and lifetime density  $g$ . We will also denote by  $V$  a random variable with density  $g$ , and by  $\varphi$  the Laplace transform of  $V$ , that is,

$$\varphi(y) := E(e^{-yV}) = \int_0^\infty g(t) e^{-yt} dt.$$

We will use later the well-known fact that if  $V$  has a finite variance, then  $\varphi$  is twice differentiable and we have

$$\varphi'(0) = -E(V) \quad \text{and} \quad \varphi''(0) = E(V^2). \quad (\text{S2})$$

Recall from main text that the node depths of the reconstructed tree all have the same distribution as some random variable  $H$ , satisfying

$$(\Pr(H > t))^{-1} = 1 - p + pW(t),$$

where  $W$  is the unique nonnegative function with Laplace transform

$$\int_0^\infty W(t) e^{-yt} dt = \frac{1}{y - \lambda + \lambda\varphi(y)}.$$

From the last two equations, we get

$$\int_0^\infty \frac{\Pr(H \leq t)}{\Pr(H > t)} e^{-yt} dt = \int_0^\infty p(W(t) - 1) e^{-yt} dt = \frac{p}{y - \lambda + \lambda\varphi(y)} - \frac{p}{y}.$$

So if we set

$$\phi(y) := \frac{p}{y - \lambda + \lambda\varphi(y)} - \frac{p}{y} = \frac{p\lambda(1 - \varphi(y))}{y(y - \lambda + \lambda\varphi(y))},$$

we get the following statement: If two models have the same reconstructed tree in distribution (i.e., the same distribution of node depths), then the function  $\phi$  is the same

for both models. Now it is easy to see that  $\lim_{y \rightarrow \infty} y^2 \phi(y) = p\lambda$ . So these two models have the same  $p\lambda$ , and consequently, they have the same function  $B$ , where

$$B(y) := \frac{p\lambda}{y\phi(y)} = \frac{y}{1 - \varphi(y)} - \lambda.$$

Now thanks to (S2), assuming that  $V$  has a finite variance, elementary calculus shows that  $B$  is continuously differentiable at 0, with

$$B(0^+) = \frac{1}{E(V)} - \lambda \quad \text{and} \quad B'(0^+) = \frac{E(V^2)}{2E(V)^2}. \quad (\text{S3})$$

Now focus on the special case when  $V$  is a  $\text{Gamma}(k, \theta)$  random variable. Recall that  $E(V) = k\theta$  and that  $\text{Var}(V) = k\theta^2$ , so that  $E(V^2) = \theta^2 k(1 + k)$ . Also,  $\varphi(y) = (1 + \theta y)^{-k} \sim (\theta y)^{-k}$ , as  $y \rightarrow \infty$ , so that

$$B(y) - y = \frac{y\varphi(y)}{1 - \varphi(y)} - \lambda \quad (\text{S4})$$

has three possible behaviors at  $+\infty$ . If  $k < 1$ , then  $B(y) - y \sim \theta^{-k} y^{1-k}$ . If  $k = 1$ , then  $B(y) - y \sim \theta^{-1} - \lambda$ . If  $k > 1$ , then  $B(y) - y \sim -\lambda$ .

Now we apply these results to the two models with gamma-distributed lifetimes of the theorem, assumed to have the same reconstructed tree in distribution. From what precedes, they have the same  $p\lambda$ , i.e.,

$$p_1 \lambda_1 = p_2 \lambda_2. \quad (\text{S5})$$

Moreover, they have the same  $B$ , so recalling the expressions for the expectation and variance of a gamma random variable, (S3) yields

$$\frac{1}{k_1 \theta_1} - \lambda_1 = \frac{1}{k_2 \theta_2} - \lambda_2 \quad \text{and} \quad k_1 = k_2. \quad (\text{S6})$$

Now by assumption, one of the two models is really age-dependent ( $k_1 \neq 1$  or  $k_2 \neq 1$ ), so  $k := k_1 = k_2 \neq 1$ . Let us use the equivalent expressions for  $B(y) - y$  as  $y \rightarrow \infty$ ,

according whether  $k < 1$  or  $k > 1$ . If  $k < 1$ , then as  $y \rightarrow \infty$ ,

$B(y) - y \sim \theta_1^{-k} y^{1-k} \sim \theta_2^{-k} y^{1-k}$ , which forces  $\theta_1 = \theta_2$ . So by (S6), we get  $\lambda_1 = \lambda_2$ , and by

(S5), we get  $p_1 = p_2$ . If  $k > 1$ , then  $B(y) - y \sim -\lambda_1 \sim -\lambda_2$ , which forces  $\lambda_1 = \lambda_2$ . So by

(S6), we get  $\theta_1 = \theta_2$ , and by (S5), we get  $p_1 = p_2$ . In any of these two cases, we can

conclude that  $(p_1, \lambda_1, k_1, \theta_1) = (p_2, \lambda_2, k_2, \theta_2)$ , which proves the theorem.

Actually, we conjecture that the same theorem should hold without even assuming that lifetimes are gamma distributed. More specifically, we suspect that if two age-dependent models of diversification have the same  $B$  (this must be the case if they have the same reconstructed trees), but not the same  $\lambda$ , then they must both be exponential.

## CHOICE OF GRID POINTS FOR NUMERICAL EVALUATION

Computation of the likelihood value requires numerical evaluation of the scale function  $W$  at discrete points, which we take as an evenly spaced grid on  $[0, T]$ , where  $T$  is the stem or crown age of the tree. Increasing the number of grid points reduces numerical error in the derivative of  $W$ , but requires more computational time and memory. In a preliminary study, we determined an appropriate number of grid points to use. Although we could instead have investigated the grid spacing, number of points provides more consistency across trees of different ages, and is thus expected to be more transferable across data sets.

For this trial we used  $\eta = 1$  and  $k = 1$ , and performed maximum likelihood inference on a fixed set of 100 simulated 1000-tip trees, using 125, 250, or 500 grid points. The results yield two main insights. Firstly, the maximum likelihood estimates of each parameter remains similar as number of grid points is varied. The distribution of actual MLEs across the 100 trees is very similar, and in particular the median shows little if any trend, as we increase the number of grid points (Supplementary Fig. S5a). Calculating the differences in the MLEs for each individual tree as we increase the number of grid points, i.e. the MLE evaluated at the higher number of points minus that evaluated at the lower number, shows that not only the overall distribution, but also the individual estimates for each tree, typically change very little as the number of grid points increases (Supplementary Fig. S5b). There are some exceptions, which importantly remain for particular trees when we go from 250 to 500 points, as well as 125 to 250 points. These are likely to correspond to trees which contain a relatively weak signal and are thus difficult to infer consistently, making them sensitive for instance to different initial points used for likelihood optimization.

The second insight is that the numerical method appears to overestimate the likelihood value, at least at its peak. This tendency can be seen directly for inference under the exponential model, where we can compare the numerically evaluated maximum log likelihood to that found using analytical computations (Supplementary Fig. S6a). The numerical value approaches the analytical value as the number of grid points increases, as expected; furthermore, it approaches from above. The tendency to overestimate likelihood is also indirectly indicated under the gamma model: with more grid points, expected to yield a more accurate evaluation, the maximum likelihood value tends to be lower (Supplementary Fig. S6b).

We can explain this bias by considering the properties of the scale function.  $W(x)$  is an increasing function of  $x$  (Lambert 2010) and asymptotically proportional to  $e^{\eta x}$  (Surya 2008). Ignoring errors in rounding branching points to the nearest grid point, the second-order central difference approximation of the derivative yields an overestimate of  $W'(x)$  whenever  $W$  is concave up, which is the case at least for sufficiently large  $x$ . This in turn yields an overestimate of the likelihood as calculated in Equation (4) or (5) of the main text. Rounding branching points results in errors in both  $W(x)$  and  $W'(x)$ , with a tendency to increase some terms of the likelihood and decrease others. The overall effect is thus unclear from examining the likelihood formulae, but based on the aforementioned numerical results, it would appear that the overall effect is to increase the likelihood in the cases examined. Analytically, we expect the magnitude of both kinds of errors to be reduced as the grid spacing is reduced.

To summarize the approach developed, since the use of more than 500 grid points becomes computationally prohibitive, and since there appears to be little effect on the MLEs with increasing numbers of grid points, we chose to use 500 grid points for all further inference. At this point, there remains a tendency to slightly overestimate the actual likelihood value at the peak. For this reason, we use the numerical evaluation for inference under both the gamma and the exponential models, despite the availability of an analytical likelihood formula in the exponential case, in order to give a fair basis of comparison under the likelihood ratio test. The results of the simulation study suggest that this approach is appropriate: when the lifetime distribution is indeed exponential,



the proportion of simulated trees for which the exponential model is rejected is close to the expected Type I error rate of 0.05 (Table 2 in the main text). The numerical over-estimation of likelihood under our model could still potentially be problematic when comparing to other models. However, with 500 grid points, we see that the median error remaining in the likelihood value (by comparison to the analytical result for the exponential model) is around 0.1 log likelihood units, which is small compared to the likelihood differences relevant for model comparison.

Finally, we note that although we have for simplicity used an evenly spaced grid of points, this is not a requirement: the scale function  $W$  can in fact be numerically evaluated (i.e. the inverse Laplace transform computed) at arbitrary points. To evaluate the likelihood, at each node depth  $x$  in the data we minimally require values of  $W$  at  $x$  and at one neighbouring point  $x + \Delta x$  (for numerical approximation of the derivative). Given an evenly spaced grid of points, we rounded each node depth to the nearest grid point and took the nearest neighbouring points to evaluate the derivative. Purely in the interests of accuracy, it would obviously be preferable not to round off, but rather to evaluate  $W$  exactly at the node depths,  $\{x_i\}$ , and at a near neighbour of each,  $\{x_i + \Delta x\}$ . For a tree on  $n$  tips, we would then have  $2(n - 1)$  points to evaluate (assuming crown age conditioning). If  $n$  is small, this approach is more efficient than our present method (500 evenly spaced grid points), as well as more accurate. But for larger  $n$ , including the size of trees examined here, this approach becomes more computationally intensive. Nonetheless, given a fixed number of points at which we wish to evaluate  $W$  (say 500), it should be possible to distribute these points over  $[0, T]$  to achieve better accuracy than with an evenly spaced grid. For instance, since deep nodes are sparse, one could evaluate  $W$  at the  $m$  deepest nodes plus  $m$  near neighbours, and then create an evenly spaced grid of  $500 - 2m$  points covering the dense nodes closer to the present. A full investigation of this approach is beyond the scope of this paper, but could be considered for future computational implementations aiming to improve accuracy and/or efficiency.

## INITIAL POINTS FOR LIKELIHOOD OPTIMIZATION

Since the likelihood under our model does not have a simple closed-form expression, but rather is computed through numerical inverse Laplace transforms, the inference framework is complex and has potential for numerical problems. To avoid spurious results, we suggest that likelihood optimization should always be run from multiple initial points, reducing the chance of accepting a local but not global peak. Furthermore, variation around the same peak in the final value found by multiple runs gives an indication of the flatness of the likelihood surface and thus the size of confidence intervals. A rather flat likelihood surface suggests that the tree contains an inherently weak signal. However, if numerical error is suspected to be a problem, one can adjust the grid size used for numerical evaluation (as explored above) and/or the parameters of the inverse Laplace transform itself (Surya 2008).

For the simulation study, we selected six initial points for likelihood optimization under the gamma lifetime model, as follows:

1. Use two points each with  $k = 0.5, 1$ , or  $10$  (representing an exponential or a higher- or lower-variance distribution).
2. Draw  $\gamma := 1/(\lambda k \theta)$  (representing the reciprocal of the average number of branching events in a lifetime), uniformly at random on  $[0.1, 0.9]$ .

3. Draw the net diversification rate,  $\eta$ , uniformly at random on  $[\eta_{\min}, \eta_{\max}]$ ,

Considering  $\eta$  as a function of  $(\lambda, k, \theta)$ , we take  $\eta_{\min} = 0.05$  and

$\eta_{\max} = \min(2, \eta(5, k, \theta))$ . These upper limits are set to avoid numerical issues that can arise when parameters are too far from the true values used for simulations.

4. Calculate  $\lambda$  and  $\theta$  from the above values of  $k$ ,  $\gamma$ , and  $\eta$ .

For inference under an exponentially-distributed lifetime, three initial points for optimization were chosen in a similar fashion (with  $k$  now fixed to one). In practice, the above algorithm can easily be modified or initial points can be chosen by trial and error for data analysis. The goal is simply to obtain several initial points, not too close together, to check that the final result is not an artifact of a single starting point.

For the Aves trees, we used the following initial points for likelihood optimization. Under the gamma model:  $(\lambda, k, \theta) = (0.115, 1, 36.6), (0.115, 0.5, 100), (0.115, 10, 1.8), (0.080, 1, 100), (0.150, 1, 15), (0.100, 10, 10)$ . Under the exponential model:  $(\lambda, \theta) = (0.080, 100), (0.100, 50), (0.150, 15)$ . The MLEs inferred for each tree are largely consistent across runs from different initial points. Under the exponential model, relative differences between the best MLE and the MLE obtained from any other initial point for a given tree never exceeded  $\mathcal{O}(10^{-4})$  for either  $\lambda$  or  $\theta$ . Under the gamma model, the median relative difference was  $\mathcal{O}(10^{-5})$  for  $\lambda$  and  $\mathcal{O}(10^{-2})$  for  $k$  and  $\theta$ . A few trees showed much larger relative differences, examined in detail in the ‘Hfull’ set. In 4/100 trees, a large relative difference arose from only one initial point where the optimization appeared to fail (returning extreme parameter values with substantially lower likelihood). In another 11/100 trees, the six initial points yielded variable MLEs that nonetheless all had very similar likelihood, suggesting that the likelihood surface was rather flat for these trees. The remaining 85/100 trees showed much smaller deviations ( $< 0.001\%$  in  $\lambda$  and a few percent in  $k$  and  $\theta$ ).

For the self-incompatible nightshades, we used the following initial points for likelihood optimization. Under the gamma model:  $(\lambda, k, \theta) = (3, 1, 0.35), (3, 0.5, 0.7), (3, 10, 0.035), (1, 1, 1.1), (5, 1, 0.2), (5, 5, 0.05)$ . All six runs returned similar results. Under the exponential model:  $(\lambda, \theta) = (3, 0.5), (1, 2), (5, 0.3)$ . Two out of three runs returned very similar results, while the third appeared to find a lower local peak.

For the self-compatible nightshades, under the gamma model we began with the initial points  $(\lambda, k, \theta) = (4, 1, 0.25), (4, 0.5, 0.5), (4, 5, 0.05), (2, 1, 0.5), (6, 1, 0.15), (6, 5, 0.03)$ . Under stem age inference, four out of these six initial points converged to very similar optima, and under crown age inference, five out of six, with the remainder appearing to find lower local peaks. Despite this consistency, further exploration indicated that the likelihood actually continued to gradually increase in some direction, as explained in the main text. Under the exponential model, we used the initial points  $(\lambda, \theta) = (4, 0.25), (2, 0.5), (6, 0.15)$ ; two out of three runs under stem age inference and all three under crown age inference returned very similar optima, and there was no indication that the likelihood could be further increased.

# SIMULATING TREES OF FIXED SIZE

The simulation package TreeSimGM (Hagen and Stadler 2013) allows two options for obtaining trees with a fixed number of tips: either (i) stop the process once the tree attains  $n$  tips for the first time (setting the option ‘gsa=FALSE’) or (ii) simulate until the tree reaches a size substantially larger than  $n$ , such that it is unlikely to return to size  $n$  again, and then choose a time point uniformly at random among the time periods when the tree had  $n$  coexisting lineages (setting the option ‘gsa=TRUE’). Scenario (ii) is equivalent to assuming a uniform prior on  $(0, \infty)$  for the stem age and then conditioning on  $n$  coexisting lineages at present (Hartmann et al. 2010).

Trees simulated under (i) obviously tend to be shorter than those under (ii), and it has previously been observed that resulting parameter estimates can be slightly different, especially for small  $n$  and (in a constant-rates model) high extinction rate (Hartmann et al. 2010). Note that the population, which is growing in expectation, will spend less time fluctuating around  $n$  coexisting species when  $n$  is large; thus stopping the first time it reaches the desired number becomes less problematic. We chose to use method (i) for the simulation study presented in the main text, due to its faster computational time.

Our simulation study also reveals a tree size-dependent bias in the parameter estimates obtained under the gamma model. Among 1000-tip simulated trees (Supplementary Tables S1-S2),  $\lambda$  and  $\epsilon$  are slightly overestimated (i.e. median MLE higher than the true value) in all but one parameter set, while  $k$  and  $\ell$  are usually slightly underestimated, with these tendencies especially apparent for smaller true  $k$ . These biases are consistent, and much more striking, among 100-tip simulated trees (Fig. 2 in the main text and Supplementary Tables S7-S8). We tested our hypothesis that these biases arise from stopping trees the first time they reach  $n$  species, thus biasing toward shorter trees, by simulating an additional set of 100-tip trees with method (ii) described above (specifically, stopping the simulations at 150 tips and then selecting a time when the tree had 100 lineages), using the same model parameters as in the main simulation study ( $\eta = 0.5$  and varying  $k$ ). The results (Supplementary Tables

S7-S8) confirmed that the bias in parameter estimates was reduced by using this alternative simulation method. We thus conclude that the biases described above can be attributed to biases in the simulated trees used for the study, rather than a problem with the inference method itself.

## CONDITIONING THE LIKELIHOOD

Throughout this study, we base our inference on the likelihood formulae given in Equations (4) and (5) of the main text, which condition on the tree age (i.e. stem or crown age, respectively). Formulae conditioning on the observed number of tips in addition to tree age can be derived (Lambert and Stadler 2013), but it has been recommended not to condition on both observed characteristics for inference, because this eliminates the information about the model parameters that is actually contained in the number of tips obtained after a given time has passed (Stadler 2013). Parameter combinations that are unlikely to give rise to the particular tree age and number of tips become likely when conditioning on both these characteristics. One might instead use likelihood formulae retaining conditioning on the observed number of tips while supposing the tree age is drawn from some distribution, and integrate the likelihood formula given both tree age and number of tips over the distribution of tree age. An appropriate choice of distribution can in turn be debated. It has been found for the constant-rate birth-death model that likelihood formulae conditioning on different features of a given tree generally result in similar parameter estimates for sufficiently large trees (Stadler 2013), and we expect this result would extend to other birth-death-type models. For simplicity, we therefore chose to base our inference on the likelihood conditioning on fixed tree age alone.

## CORRELATION IN ESTIMATED LIFETIME DISTRIBUTION PARAMETERS

In the simulation study, the maximum likelihood estimates of the lifetime distribution parameters,  $\hat{k}$  and  $\hat{\theta}$ , are observed to be negatively correlated, such that the

estimated mean lifetime  $\hat{\ell} = \hat{k}\hat{\theta}$  is much more precise than either individual parameter. An example of the estimated  $(\hat{k}, \hat{\theta})$  pairs across 100 trees simulated under identical parameters is illustrated in Supplementary Fig. S7. We furthermore observe that with true  $\ell = k\theta$  fixed, the higher the true  $k$  value, the stronger the correlation between  $\hat{k}$  and  $\hat{\theta}$ . This observation can be explained analytically. The model parameters appear in the Laplace exponent  $\psi$  (see “Mathematical Likelihood Formulae” in the main text), which determines the scale function  $W$  and in turn the likelihood. Under the gamma lifetime distribution model,  $\psi$  takes the form:

$$\psi(y) = y - \lambda \left( 1 - (1 + \theta\lambda)^{-k} \right)$$

Taking a Taylor series expansion, which converges when  $|\theta y| < 1$ , we have:

$$\psi(y) = y - \lambda \left( 1 - \exp \left( -k(\theta y - (\theta y)^2/2 + \dots) \right) \right)$$

Thus, for sufficiently small  $\theta$ ,  $\psi(y) \approx y - \lambda(1 - \exp(-k\theta))$ , such that  $k$  and  $\theta$  become almost indistinguishable in the likelihood expression. This statement is exact in the limit as  $k \rightarrow \infty$  and  $\theta \rightarrow 0$  such that  $k\theta \equiv \ell$  is fixed (a Dirac delta distribution). On the other hand, if  $\theta$  is not small, the higher order terms of the Taylor expansion are non-negligible, and  $k$  and  $\theta$  make distinct contributions. Expressed more intuitively, the gamma distribution has mean  $k\theta$ , variance  $k\theta^2$ , etc. So  $k$  and  $\theta$  do not generally play symmetric roles, but the information to distinguish these two parameters disappears as variance in the lifetime distribution decreases.

# REFERENCES

\*

## References

- Hagen, O. and T. Stadler. 2013. TreeSimGM: simulating phylogenetic trees under a general model. Available online at <http://cran.r-project.org/web/packages/TreeSimGM/index.html>.
- Hartmann, K., D. Wong, and T. Stadler. 2010. Sampling trees from evolutionary models. *Syst. Biol.* 59:465–476.
- Lambert, A. 2010. The contour of splitting trees is a Lévy process. *Ann. Probab.* 38:348–395.
- Lambert, A. and T. Stadler. 2013. Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.* 90:113–128.
- 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. 2013. How can we improve accuracy of macroevolutionary rate estimates? *Syst. Biol.* 62:321–329.
- Surya, B. A. 2008. Evaluating scale functions of spectrally negative Lévy processes. *J. Appl. Prob.* 45:135–149.