# Approximate Bayesian Computation of diversification rates from molecular phylogenies

Janzen, Thijs; Hoehna, Sebastian; Etienne, Rampal S.

*Published in:*
Methods in ecology and evolution

*DOI:*
10.1111/2041-210X.12350



*Publication date:*
2015

Methods in Ecology and Evolution

British Ecological Society

*Methods in Ecology and Evolution* 2015, **6,** 566–575 doi: 10.1111/2041-210X.12350

# Approximate Bayesian Computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT

**Thijs Janzen[1]*, Sebastian Höhna[2] and Rampal S. Etienne[1]**

[1]*Groningen Institute for Evolutionary Life Sciences, University of Groningen, 9700 CC Groningen, the Netherlands; and* [2]*Department of Evolution and Ecology, University of California, Davis, Storer Hall, One Shields Avenue, Davis, CA 95616, USA*

## Summary

**1.** Molecular phylogenies form a potential source of information on rates of diversification, and the mechanisms that underlie diversification patterns. Diversification models have become increasingly complex over the past decade, and we have reached a point where the computation of the analytical likelihood of the model given a phylogeny is either unavailable or intractable. For such models, a likelihood-free approach such as Approximate Bayesian Computation (ABC) offers a solution. ABC is a Bayesian framework that uses one or more summary statistics instead of the likelihood function. Crucial to the performance of an ABC algorithm is the choice of summary statistics.

**2.** Here, we analyse the applicability of three traditional and often-used summary statistics (Gamma statistic, Phylogenetic Diversity and tree size) within an ABC framework and propose a new summary statistic: the normalized Lineages-Through-Time (nLTT) statistic.

**3.** We find that the traditional summary statistics perform poorly and should not be used as a substitute of the likelihood. By contrast, we find that the nLTT statistic performs on par with the likelihood.

**4.** We suggest to include the nLTT statistic in future ABC applications within phylogenetics. We argue that the use of ABC in diversification rate analysis is a promising new approach, but that care should be taken which summary statistics are chosen.

**Key-words:** Approximate Bayesian Computation, Lineages-Through-Time, phylogenetics, summary statistic

## Introduction

To understand present-day biodiversity, it is crucial to study which processes contribute to biodiversity, and how these processes have changed over time. Historically, inferences on the macro-evolutionary history of species, and more specifically, inferences of speciation rates and extinction rates, were drawn using the fossil record and morphological similarity (Gould & Eldredge 1977; Raup & Sepkoski 1982; Peters & Foote 2002; Coyne & Orr 2004). Molecular phylogenies have provided a new source of information on evolutionary histories. A large array of models that attempt to infer past speciation and extinction rates from phylogenies has been developed in the last decade (Morlon 2014), including, but not limited to, the constant-rate birth–death process (Nee, May & Harvey 1994), character-state-dependent diversification rates (Maddison, Midford & Otto 2007; FitzJohn, Maddison & Otto 2009), time-dependent diversification rates (Morlon, Parsons & Plotkin 2011; Stadler 2011; Höhna 2014), diversity-dependent diversification rates (Rabosky & Lovette 2008a; Etienne *et al.* 2012) and protracted speciation (Etienne & Rosindell 2012). All these models are crude simplifications of the underlying mechanics causing diversification. This allows these models to remain mathematically tractable, and exact likelihoods of the models given the phylogeny can be formulated to infer parameter estimates. At the same time, the mathematical tractability poses a restriction on the complexity of these models; simple extensions often result in intractable models. Consider for example, a model where speciation and extinction rates depend on geographical change (Pigot *et al.* 2010; Goldberg, Lancaster & Ree 2011), a model including interactions between population size and speciation rate (Jabot & Chave 2009; Davies *et al.* 2011) or a model where extinction rates that are heritable across species (Rabosky 2009). Fortunately, it is often fairly straightforward to simulate these models. This property can be exploited in Approximate Bayesian Computation (Tavaré *et al.* 1997; Beaumont, Zhang & Balding 2002; Sunnåker *et al.* 2013). Instead of using the likelihood as an indicator of model fit, Approximate Bayesian Computation (ABC) relies on summary statistics to obtain a Bayesian posterior distribution for the model parameters. Specifically, the 'fit' of a set of model parameters is determined by simulating the model for this set

*Correspondence author. E-mail: thijsjanzen@gmail.com

of parameters and then comparing the summary statistics for the simulated data to the summary statistics of the real data. Parameter sets which result in a smaller distance between observed and simulated summary statistics are given higher probability. So, apart from summary statistics, the method formally also requires a distance metric. The most common metrics are the absolute or squared difference.

Approximate Bayesian Computation methods are already widely used in population genetics to infer demographic parameters (Tavaré *et al*. 1997; Bazin, Dawson & Beaumont 2010; Csilléry *et al*. 2010) and are also used in fields beyond molecular genetics, including epidemiology (Blum & François 2009; Blum & Tran 2010) and systems biology (Barnes, Silk & Stumpf 2011). Within ecology, applications of ABC have been limited to the estimation of Lotka-Volterra dynamics (Toni *et al*. 2009), host–parasite dynamics (Drovandi & Pettitt 2011) and neutral community ecology (Jabot & Chave 2009). Surprisingly, ABC has only scarcely been applied to diversification rate analysis. Rabosky (2009) used ABC to estimate the heritability of extinction rates and Bokma (2010) used ABC to obtain estimates of anagenetic and cladogenetic evolution. Slater and colleagues used a hybrid approach of ABC to estimate trait values and traditional likelihoods to obtain diversification estimates (Slater *et al*. 2012). Kutsukake & Innan (2013) used ABC to obtain estimates of phenotypic evolution. These studies show great potential for future application of ABC within macro-evolutionary studies.

The choice of summary statistics is crucial to the performance of ABC. For diversification rate analysis, no systematic analysis of available summary statistics has been conducted yet. Efficient summary statistics for diversification rate analysis should be able to recover parameter values for a range of diversification models. Here, we analyse three summary statistics traditionally used in diversification rate analysis: tree size, the Gamma statistic (Pybus & Harvey 2000) and Phylogenetic Diversity (Clarke & Warwick 2001). We test the performance of these summary statistics within an ABC framework by comparing inferences made with ABC for trees simulated with the constant-rate birth–death model (Nee, May & Harvey 1994), a time-dependent speciation model (Rabosky & Lovette 2008b) and the diversity-dependent speciation model (Etienne *et al*. 2012) with inferences using the exact likelihood. We show that under these different scenarios, none of the established summary statistics is able to reliably recover parameter estimates. Furthermore, we introduce a novel summary statistic: the normalized Lineage-Through-Time (nLTT) difference. We show that this novel summary statistic can be used as a substitute for the likelihood.

## Methods

We simulated data using three different models of increasing complexity: the birth–death model (Nee, May & Harvey 1994), the time-dependent speciation model (Rabosky & Lovette 2008b; Höhna 2014) and the diversity-dependent speciation model (Rabosky & Lovette 2008a; Etienne *et al*. 2012). We inferred the parameters using ABC with three different summary statistics and with standard Bayesian Computation using the analytical likelihood. We then compared our obtained estimates with the true parameter values and the likelihood estimates.

*Tree size*, the *Gamma statistic* (Pybus & Harvey 2000) and *Phylogenetic Diversity* (Faith 1992; Clarke & Warwick 2001) are commonly used as summary statistics to capture properties of phylogenies. Additionally, we introduce a new statistic, the *normalized Lineage-Through-Time statistic* (*nLTT*).

Tree size is simply the number of tips of the phylogeny. The Gamma statistic (Pybus & Harvey 2000) is given by:

$$\gamma = \frac{\left(\frac{1}{n-2}\sum_{i=2}^{n-1}\left(\sum_{k=2}^{i} k g_k\right)\right) - \left(\frac{T}{2}\right)}{T\sqrt{\frac{1}{12(n-2)}}},\ T = \left(\sum_{j=2}^{n} j g_j\right)$$

where $g_2, g_3, \ldots g_n$ are the internode distances of a reconstructed phylogeny with $n$ taxa. A value of $\gamma > 0$ indicates that the phylogeny's internal nodes are closer to its tips than expected under the pure-birth model, whilst $\gamma < 0$ indicates internal nodes closer to the root than expected under the pure-birth model. We calculated the gamma statistic using the function gammaStat from the R-package ape (Paradis, Claude & Strimmer 2004). As Phylogenetic Diversity metric we chose the average Phylogenetic Diversity, which is independent of the number of taxa in the tree [in contrast to the traditional Phylogenetic Diversity metric (Faith 1992)]. Phylogenetic Diversity is defined as the sum of all branch lengths in the tree, divided by the number of branches (Clarke & Warwick 2001), that is

$$PD = \frac{\sum_{i=1}^{2(n-1)} b_i}{2n-2}$$

where $n$ is the number of taxa in the tree and $b_i$ is the length of branch $i$ in a tree, thus $\sum_{i=1}^{2(n-1)}$ represents the sum of all branches, and PD is thus the average branch length.

When analysing a phylogeny, or comparing multiple phylogenies generated by the same model, often the Lineages-Through-Time (LTT) curve of these phylogenies proves to be helpful (Paradis 2011). The LTT curve shows the number of lineages in the phylogeny over time. The shape of the LTT curve can reveal periods with high speciation and can identify signatures of extinction (Nee, May & Harvey 1994). A potentially powerful way to compare phylogenies lies therefore in the comparison of their LTT curves. Problems arise however when the two compared trees are not of the same size or same height. Differences in number of lineages or differences in the time to the most common recent ancestor might overshadow the differences in shape that we are trying to detect. Furthermore, differences in scale between the two compared trees can make it difficult to predict the range of expected differences between the LTT curves. We therefore normalize the LTT curves both in number of lineages and in time, that is we divide the number of lineages by the number of extant tips and the time by the time to the most recent common ancestor. This yields a function on relative time span [0,1], and the number of lineages also spans the interval [0,1] The difference in normalized LTT curves falls consistently within the same range and facilitates the use of a consistent threshold acceptance value in ABC analysis. Then, as a distance metric in ABC, we use the absolute distance between the normalized LTTs, which is given by:

$$\Delta nLTT = \int_0^1 |nLTT_1(t) - nLTT_2(t)|\mathrm{d}t \qquad \text{(eqn 1)}$$

where $nLTT_1(t)$ is the normalized number of lineages at normalized time $t$ for phylogeny 1, and $nLTT_2(t)$ is the normalized number of lineages at normalized time $t$ for phylogeny 2. Equation (1) thus presents a natural weighing of the contribution of all $nLTT(t)$-differences (hence, we are effectively using a vector of summary statistics) and captures the surface enclosed between the two normalized LTT curves; this

surface carries information about similarity between the curves (see Fig. 1). One could alternatively use the area under the LTT curve, but the difference between two areas under the LTT curve could be equal to 0 even if the LTT curves are very different, whereas our distance metric is equal to 0 if and only if the two LTT curves compared are identical.

## Simulation models

In this study, we focus on three different two-parameter diversification models; the birth–death model, the time-dependent speciation model (without extinction) and the diversity-dependent speciation model (without extinction). A full description of these diversification models and their associated likelihood formulas can be found in the Supplementary Information. The two parameters are speciation and extinction rate for the birth–death model, initial speciation rate and the rate of speciation decay in the time-dependent speciation model, and initial speciation rate and the diversity limit in the diversity–dependent speciation model. We did not include extinction in the time-dependent and diversity-dependent models to facilitate a fair comparison between the three different models. In our analysis, we chose a representative set of parameter values to generate our simulation data (see also Table 1). We kept the time to the most recent common ancestor constant, and adjusted the speciation rate such that under the chosen parameterization, the *expected* number of tips was fixed (e.g. with increasing extinction, we increased the speciation rate), to guarantee sufficiently large trees and thereby prevent poor estimation just because of little information in the data. For the diversity-dependent model, we did not keep the expected number of tips fixed, but fixed maximum diversity, thus ensuring that with different speciation rates we capture different degrees of diversity dependence. The alternative approach, keeping the number of tips fixed and letting the time to the most recent common ancestor vary, is not possible except for the birth–death model (Lambert & Stadler 2013; Stadler 2013). Keeping the expected number of tips fixed may seem to disqualify the tree size statistic as an efficient statistic from the start, but we note that tree size still varies in simulations with the same parameter values, so poor performance of the tree size statistic will be due to the expected lack of information in this statistic. One could view the tree size statistic as a reference point because it is a worst-case scenario, a statistic with hardly any information. Inferences made using the tree size statistic therefore show the results of ABC in the near-absence of information.

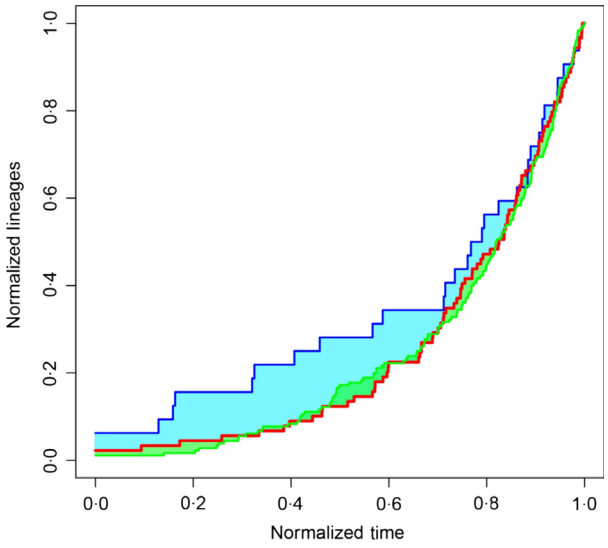


**Fig. 1.** Normalized LTT curves for three phylogenetic trees generated using the pure-birth model with parameters $\lambda = 0{\cdot}4$ (red and green lines) and $\lambda = 0{\cdot}25$ (blue line). Time to the most recent common ancestor was identical for all three trees and set at 10 million years. Coloured surfaces indicate the nLTT difference between the respective trees, with the light blue surface the difference between the blue and red line, and the lightgreen surface the difference between the green and the red line. Clearly, the surface between the two trees generated with the same $\lambda$ ($\lambda = 0{\cdot}4$, red and green lines) is smaller than the surface between the two trees generated with different $\lambda$ ($\lambda = 0{\cdot}25$ and $\lambda = 0{\cdot}4$, blue and red lines).

## Birth–death model

We set the ratio of the extinction and speciation rates to 0·0, 0·1, 0·3 and 0·9, ensuring trees with both low and very high extinction. Speciation rates were adjusted such that the expected number of species was always 100. This resulted in corresponding speciation rates of 0·39, 0·42, 0·51 and 1·77. For every parameter combination, 30 trees were simulated using the package TESS (Höhna 2013) in R 3.0.1. (R Core Team 2014).

## Time-dependent model

Extinction was set to zero to keep the number of free parameters at two. The speciation rate was chosen to decay in an exponential fashion: $\lambda(t) = \lambda_0 \exp[-\alpha(t - t_0)]$, where $\lambda_0$ is the initial speciation rate at time $t_0$ (here: the crown age) and $\alpha$ governs the time-dependent decay. For $\alpha = 0$, we recover the standard birth–death model. Values for $\alpha$ were chosen to be either [0·1, 0·5, 0·9], and $\lambda_0$ values were adjusted to keep the expected number of tips at 200. This is a larger expected tree size than in the birth–death model because variation in tree size is relatively large for the time-dependent speciation model, and hence, we avoided trees with an extremely low number of tips. Resulting $\lambda_0$ values were [0·73, 2·32, 4·15], respectively. We simulated 30 trees using the R-package TESS (Höhna 2013) in R 3.0.1. (R Core Team 2014).

## Diversity-dependent model

In the diversity-dependent model, we assumed that extinction is zero and the speciation rate decreases linearly with increasing diversity: $\lambda(N) = \lambda_0(1 - N/K)$, where $\lambda_0$ is the speciation rate at low diversity, $N$ is the current diversity and $K$ is the maximum diversity. As $N$ approaches $K$, the effective speciation rate approaches 0. We chose to adjust $\lambda_0$ whilst keeping $K$ constant, to vary the degree of limitation due to diversity dependence. $K$ was chosen to be 200, with $\lambda_0$ values being [0·5, 0·75, 1·0]. Expected tree sizes ranged between 100 and 200, depending on $\lambda_0$. We simulated 30 trees using the R-package DDD (Etienne *et al.* 2012) in R 3.0.1. (R Core Team 2014).

## Likelihood

To obtain our reference estimates using the analytical likelihood, we performed a Markov chain Monte Carlo (MCMC) analysis using a Metropolis–Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970). The MCMC chain was run for 1 000 000 iterations, with a discarded burn-in phase of 10 000 at the beginning and the chain was thinned out by sampling every 10th iteration. Likelihoods for the birth–death model and the time-dependent model were calculated using the package TESS (Höhna 2013), likelihoods of the diversity-depen-

**Table 1.** Parameter values used to simulate the three different models, for speciation rate (λ), extinction rate (μ), time-dependent decay (α) or maximum diversity (*K*) values. The last three columns show the mean (standard deviation between brackets) statistics of the 30 trees used in our analysis

| | Treatment | λ | μ/α/*K* | Number Tips | γ | PD |
|---|---|---|---|---|---|---|
| Birth–Death | 0 | 0·39 | 0·00 | 98 (72) | 0·07 (0·95) | 1·38 (0·29) |
| | 0·1 | 0·42 | 0·04 | 85 (67) | 0·22 (0·83) | 1·37 (0·26) |
| | 0·3 | 0·51 | 0·15 | 85 (65) | 0·86 (1·00) | 1·29 (0·37) |
| | 0·9 | 1·77 | 1·60 | 84 (60) | 5·15 (1·39) | 0·75 (0·23) |
| Time-dependent | 0·1 | 0·73 | 0·10 | 204 (129) | −2·42 (0·42) | 1·45 (0·10) |
| | 0·5 | 2·32 | 0·50 | 202 (150) | −13·54 (5·75) | 3·07 (0·13) |
| | 0·9 | 4·15 | 0·90 | 190 (140) | −17·85 (6·53) | 3·87 (0·07) |
| Diversity-dependent | 0·5 | 0·50 | 200 | 105 (35) | −2·40 (2·27) | 1·50 (0·20) |
| | 0·75 | 0·75 | 200 | 187 (11) | −9·57 (2·11) | 2·02 (0·31) |
| | 1 | 1·00 | 200 | 198 (2) | −14·83 (2·26) | 2·60 (0·49) |

dent model using the package DDD (Etienne *et al*. 2012). All calculations were performed using R 3.0.1. (R Core Team 2014).

## ABC

We used an Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC) approach (Toni *et al.* 2009). Within the SMC approach, particles are first generated from the prior distribution. Particles are then resampled from the obtained sample, and slightly perturbed. From these resampled particles, a new sample is formed, from which again particles are resampled, etc. The threshold value ε for the summary statistic – below which new particles are accepted – is lowered with every newly obtained sample. As a result, the acceptance rate decreases and the acceptance threshold ε approaches zero with an increase in the number of iterations (resamplings). This is in contrast to the more standard ABC-MCMC approach, where particles are continuously propagated using a fixed threshold ε. To get good results within an ABC-MCMC framework, manual adjustment of ε is often necessary, and the chain must often be started multiple times before obtaining the desired balance between a low threshold value and high acceptance rate. The ABC-SMC approach keeps these manual adjustments to a minimum. In a pilot study, we considered adaptive SMC or adaptive MCMC (Beaumont 2010; Drovandi & Pettitt 2011; Lenormand, Jabot & Deffuant 2013), where ε-values and the jumping distribution are adapted during runtime, but these algorithms proved to be relatively slow compared to the ABC-SMC algorithm used here (Toni *et al*. 2009).

The ABC-SMC scheme proceeds as follows:

**1.** Initialize $\varepsilon_{1 \ldots} \varepsilon_T$

Set the population indicator $t = 0$

**2.** Set the particle indicator $i = 1$.

**3.** If $t = 0$, sample θ** from the prior π

Otherwise sample θ* from the previous population $\theta^i_{t-1}$ with weights $w_{t-1}$ and perturb the particle to obtain θ** such that θ** = θ* + $N(0, 0{\cdot}05)$

**4.** If π(θ**) = 0, return to 3.

**5.** Simulate a tree $T^*$ ~ θ**

**6.** Calculate summary statistic $S$ for tree $T^*$

**7.** If $|SS(T^*) - SS(T_0)| > \varepsilon_t$ return to 3.

**8.** Set $\theta^i_t$ = θ** and calculate the weight for particle $\theta^i_t$:

If $t = 0$; $w^i_t = 1$

If $t > 0$; $w^i_t = \frac{\pi(\theta^i_t)}{\sum_{j=1}^{N} w^j_{t-1} N(\theta^j_t, \theta^i_t)}$

**9.** If $i < N$, set $i = i + 1$, go to 3.

**10.** Normalize the weights

**11.** If $t < T$, set $t = t + 1$, go to 2.

where $\varepsilon_{1 \ldots} \varepsilon_T$ are the threshold values for iterations $1 \ldots T$, π is the prior function, θ is a particle with associated parameter values, $N(\mu,\sigma)$ is a normal distribution with mean = μ and standard deviation = σ. $N$ is the total number of particles per iteration $t$, and $w^i_t$ is the weight of particle $i$, of iteration $t$.

Initial ε-values were estimated as follows. We generated 100 trees with the same age using the package TESS (Höhna 2013). For these 100 trees, we calculated the four summary statistics and used the standard deviation of this distribution as the initial epsilon value. This resulted in initial $\varepsilon_0$ values of [50, 1, 1, 0·2] for tree size, Gamma statistic, Phylogenetic Diversity and the normalized LTT statistic, respectively The ε-value was decreased in an exponential fashion as the sequential ABC scheme progresses, such that the epsilon value was $\varepsilon_t = \varepsilon_0 \exp(-0{\cdot}5t)$, where $t$ is the current sequential step. This exponential decrease follows the progression of ε in the adaptive-SMC scheme of Del Moral *et al*. (2012). We found that using 5000 particles per iteration was sufficient. Convergence was assumed when the acceptance rate of newly proposed particles had dropped below 1 in 100, and visual inspection of the interquartile ranges showed no change in the posterior distribution of the past 3 iterations. We provide code used to perform the ABC-SMC analysis in the nLTT package for R. As prior distributions, we chose to use a uniform prior on [0,5] for both parameters of the birth–death model, assuming that both speciation and extinction cannot become negative and limiting them from becoming extremely large values, and thus producing unrealistic trees. For the time-dependent model, we chose a log-normal prior with mean 0 and standard deviation of 1 for both parameters, such that both the initial speciation and the time-decay parameter were always positive. For the diversity-dependent model, we chose a log-normal prior with mean 0 and standard deviation of 1 for $\lambda_0$ and a log-normal prior with mean 4·6 and standard deviation of 1 for $K$ for the diversity-dependent model. Both these priors are always positive, and the mean for the $K$ is chosen to be centred around the expected tree size [exp(4·6) = 100]. These priors were used both in the Metropolis–Hastings MCMC approach and in the ABC-SMC approach.

## Convergence analysis

In the limit of $\varepsilon \to 0$ and $N \to \infty$, where $N$ is the number of ABC-samples, the Approximate posterior distribution converges to the true posterior distribution (Marjoram *et al*. 2003), if the summary statistic is sufficient. In our SMC approach, the threshold ε asymptotically moves towards 0 with an increasing number of iterations of the resampling algorithm. If the summary statistic used is indeed sufficient, then the mean of the approximate posterior distribution should asymptoti-

cally move towards the true mean, with decreasing ε. To test this, we plotted the natural logarithm of the threshold against the mean of the obtained posterior distribution using ABC. Furthermore, we plotted the 95% confidence interval of the means across the 30 different trees.

## Results

### TREES

The trees generated with the different models show different patterns, depending on the model and the parameterization of the model. With increasing extinction, we observed a slight decrease in the average number of tips (despite correcting the speciation rate) (Table 1 and Fig. 2), and only for very high levels of extinction ($\mu = 0{\cdot}9\lambda$), patterns in the LTT plot differ considerably from those having no extinction at all (Fig. 2). Increasing levels of extinction leads to an increase in the pull-of-the-present (Nee, May & Harvey 1994), which is reflected by an increase in the average Gamma statistic. Higher extinction also implies lower Phylogenetic Diversity. For the time-dependent model, LTT plots reveal that trees with an $\alpha$ of 0·1 are still fairly similar to the birth–death model (Fig. 2), and trees with higher $\alpha$ values show a slowdown in lineage accumulation. This slowdown is reflected in highly negative Gamma values (Table 1). Phylogenetic Diversity on the other hand, increases with an increase in $\alpha$, possibly because the initial speciation rate was higher. Diversity-dependent LTT plots show less speciation-limitation than the time-dependent plots, and higher $\lambda_0$ values seem to induce more limitation than lower values; $\lambda_0 = 0{\cdot}5$ seems fairly similar to the birth–death model (Fig. 2). Higher $\lambda_0$ values generate larger trees and trees with higher Phylogenetic Diversity.

### LIKELIHOOD

Crucial to our comparison between ABC and the likelihood-based estimates is the accuracy of the likelihood estimates, and whether or not patterns in our obtained likelihood estimates resemble patterns reported in the literature. Comparing our estimates for the three models with the parameter values, the trees were generated with, we find that for the constant-rate birth-death model, both the speciation and extinction parameter are overestimated (Table 2, Figs 3 and 4), a well-known result (Nee, May & Harvey 1994; Höhna 2014). Estimates of $\alpha$ for the time-dependent speciation model are accurate, whilst the initial speciation rate tends to be slightly underestimated (Table 2, Figs 3 and 4), especially when $\alpha$ is high. For the diversity-dependent model, maximum speciation rates are estimated accurately, and *K*-values are consistently overestimated (Table 2, Figs 3 and 4).

### TREE SIZE

The tree size statistic performs equally poorly for all three models (Table 2). Speciation and extinction rates are largely overestimated for the constant-rate birth–death model, and estimates do not differ much between different rates of extinction, suggesting that the tree size statistic does not detect any signature of extinction in the trees.

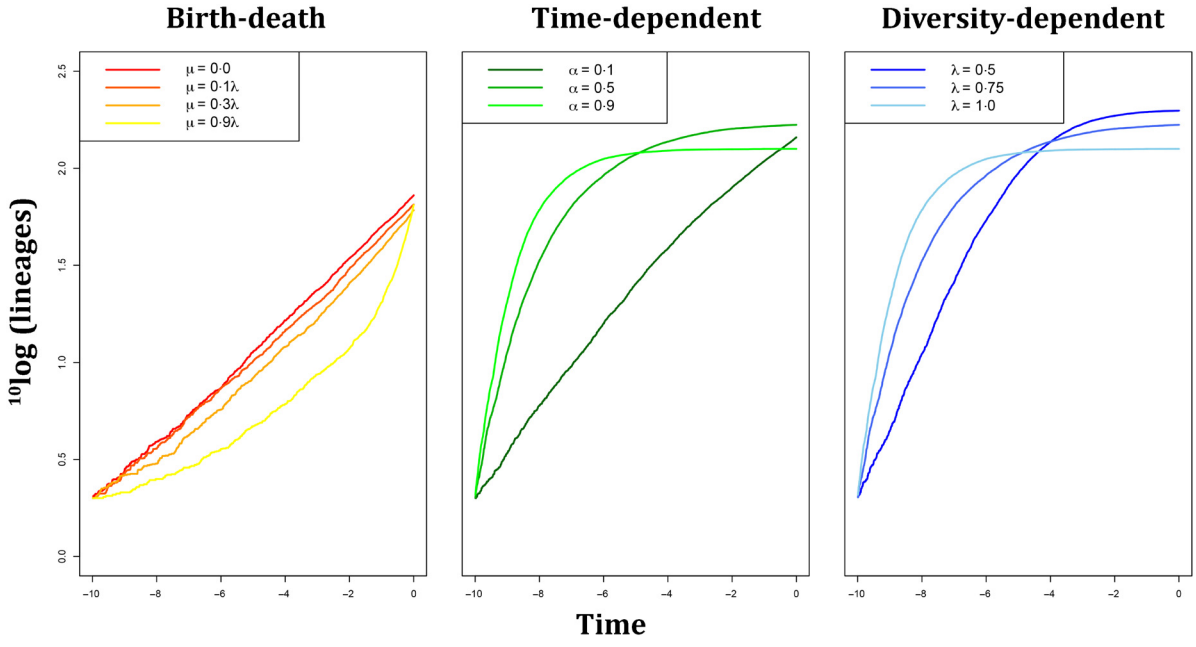


**Fig. 2.** LTT curves of the simulated phylogenetic trees used in our analysis. Shown are the mean LTT curves of the birth–death model, the time-dependent speciation and the diversity-dependent speciation model. Colours indicate different parameterizations. The number of replicates per parameterization was 30.

For the time-dependent speciation model, $\lambda_0$ and $\alpha$ are underestimated. Again, the tree size statistic shows little variation in estimates across various degrees of time dependence, indicating that the tree size statistic is unable to detect time dependence. For the diversity-dependent model, $\lambda_0$ is overestimated whilst *K*-values are underestimated. Considering that the net speciation rate becomes zero once diversity has reached *K*, it is surprising to see that estimates of *K* are smaller than the value with which the trees were generated, which suggests that the size of the generated trees was generally much smaller than the true *K* value, which is especially true for smaller values of $\lambda_0$. There is some variation in the estimates across the various degrees of diversity dependence. However, variation around the mean estimates is large and for any single tree, the tree size statistic will most likely be unable to detect patterns of diversity-dependent speciation.

### GAMMA

The Gamma statistic performs comparably well for all four parameterizations of the constant-rate birth–death model, which we expected considering that the Gamma statistic was devised as a way to detect deviations in the branching pattern from the pure-birth process (Pybus & Harvey 2000). Bias in the estimate seems to be comparable to the bias obtained using a likelihood-based approach (Table 2). Indeed, further inspection reveals that only when extinction is high (0·9 times the speciation rate), the Gamma statistic overestimates the speciation rate more than the likelihood-based approach (Fig. 3). As with the likelihood-based approach, the Gamma statistic overestimates the extinction rate, but the bias is comparable to the likelihood-based approach. The performance of the Gamma statistic for the time-dependent model is poor: with increasing time dependence, the Gamma statistic increasingly underestimates $\lambda_0$. Although variance in the $\alpha$ estimate is low, estimates are inaccurate. For trees generated with the diversity-dependent model, the Gamma statistic also fails to perform on par with the likelihood-based approach. $\lambda_0$ is overestimated, and *K* is underestimated. Estimates for *K* appear not to differ between parameterizations, suggesting that the Gamma statistic is unable to detect signals of diversity dependence.

### PHYLOGENETIC DIVERSITY

The performance of Phylogenetic Diversity for the constant-rate birth–death model is comparable to the tree size statistic (Table 2). Both speciation and extinction are overestimated, and variance in the estimate is high. Estimates across varying degrees of extinction are similar, indicating that Phylogenetic Diversity is unable to detect patterns of extinction in phylogenies. As with the Gamma statistic, $\lambda_0$ in the time-dependent model is increasingly underestimated with increasing time dependence, together with an underestimation of $\alpha$. $\lambda_0$ estimates for the diversity-dependent model are inaccurate and similar across degrees of diversity dependence, indicating that

**Table 2.** Mean parameter estimates for the four different summary statistics and the likelihood. Numbers between brackets represent the standard deviation

| Model used to generate tree | | | Tree size | | γ | | PD | | nLTT | | Likelihood | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Birth–Death | λ | μ | λ | μ | λ | μ | λ | μ | λ | μ | λ | μ |
| 0 | 0·39 | 0 | 3·65 (1·70) | 3·72 (1·89) | 0·41 (0·27) | 0·37 (0·48) | 1·72 (1·09) | 1·96 (1·40) | 0·47 (0·15) | 0·17 (0·22) | 0·5 (0·18) | 0·22 (0·27) |
| 0·1 | 0·42 | 0·04 | 3·77 (1·68) | 4·04 (1·92) | 0·45 (0·26) | 0·46 (0·49) | 1·78 (1·13) | 2·05 (1·48) | 0·48 (0·13) | 0·17 (0·17) | 0·5 (0·17) | 0·23 (0·23) |
| 0·3 | 0·51 | 0·15 | 3·65 (1·70) | 3·77 (1·92) | 0·59 (0·43) | 0·69 (0·71) | 1·81 (1·19) | 2·03 (1·53) | 0·59 (0·27) | 0·30 (0·38) | 0·62 (0·29) | 0·36 (0·39) |
| 0·9 | 1·77 | 1·6 | 3·84 (1·71) | 4·18 (1·94) | 2·70 (1·13) | 2·97 (1·31) | 1·81 (1·16) | 1·76 (1·49) | 2·02 (0·72) | 1·86 (0·84) | 2·18 (1·00) | 2·14 (1·21) |
| Time-dependent | λ | α | λ | α | λ | α | λ | α | λ | α | λ | α |
| 0·1 | 0·73 | 0·1 | 1·17 (0·72) | 0·23 (0·18) | 0·62 (0·40) | 0·43 (0·42) | 0·99 (0·33) | 0·13 (0·05) | 0·82 (0·22) | 0·11 (0·04) | 0·80 (0·21) | 0·11 (0·04) |
| 0·5 | 2·32 | 0·5 | 1·15 (0·71) | 0·23 (0·18) | 2·00 (0·90) | 0·44 (0·26) | 1·43 (1·08) | 0·36 (0·16) | 2·37 (0·59) | 0·50 (0·06) | 2·16 (0·59) | 0·49 (0·06) |
| 0·9 | 4·14 | 0·9 | 1·15 (0·70) | 0·23 (0·18) | 2·28 (1·00) | 0·50 (0·53) | 1·69 (1·68) | 0·52 (0·27) | 4·13 (1·01) | 0·89 (0·09) | 3·75 (0·95) | 0·87 (0·09) |
| Diversity Dependent | λ | K | λ | K | λ | K | λ | K | λ | K | λ | K |
| 0·5 | 0·5 | 200 | 1·92 (1·27) | 104·46 (34·11) | 0·91 (1·35) | 42·90 (68·45) | 0·46 (0·10) | 229·53 (154·07) | 0·51 (0·09) | 263·27 (150·13) | 0·48 (0·10) | 298·54 (243·67) |
| 0·75 | 0·75 | 200 | 2·11 (1·29) | 186·42 (11·00) | 3·11 (2·75) | 47·56 (23·93) | 0·41 (0·15) | 107·88 (110·51) | 0·72 (0·09) | 209·37 (123·34) | 0·72 (0·08) | 231·84 (167·93) |
| 1 | 1 | 200 | 2·13 (1·31) | 197·95 (2·16) | 3·95 (3·05) | 97·59 (35·25) | 0·48 (0·27) | 65·98 (57·66) | 0·98 (0·14) | 211·42 (139·13) | 0·98 (0·12) | 341·89 (366·81) |

 

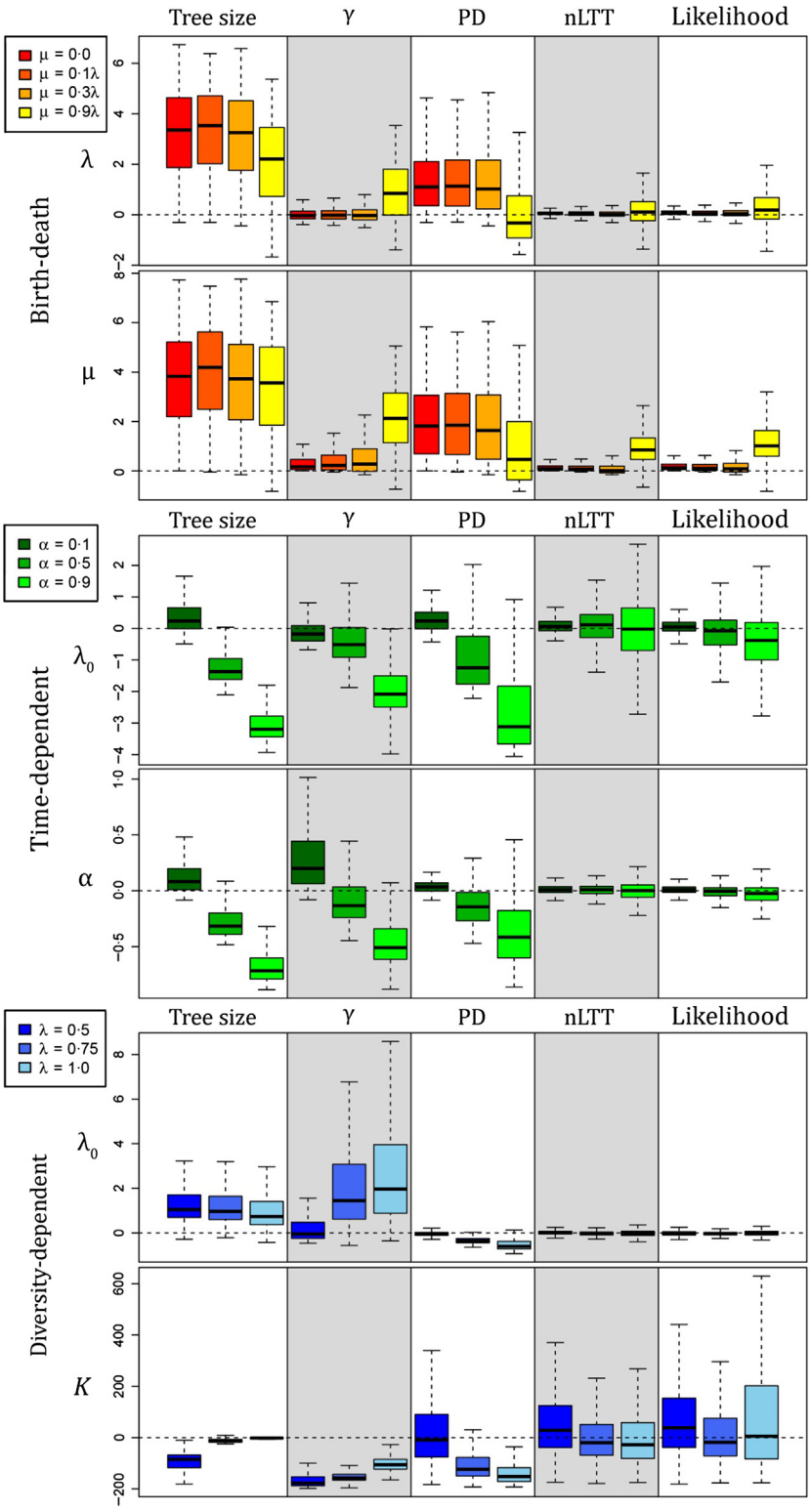


**Fig. 3.** Results of parameter inference using the likelihood within a Metropolis–Hastings MCMC method and using either of the four summary statistics tree size, Gamma, Phylogenetic Diversity and nLTT within an ABC-SMC framework (white background). Plotted are the residual estimates after subtracting the parameter value used to generate the tree.

Phylogenetic Diversity cannot detect patterns of diversity dependence.

### NORMALIZED LTT

Estimates of the speciation and extinction rate for the constant-rate birth–death model are similar to estimates obtained using the likelihood (Table 2). Estimates for $\lambda_0$ and for $\alpha$ are close to the values used to generate the trees and are similar to estimates obtained using the likelihood, both in mean and in variance. Estimates and variance for both parameters of the diversity-dependent model are generally similar to estimates obtained using a likelihood-based approach (Fig. 3). Variance in $K$ estimate appears to be a

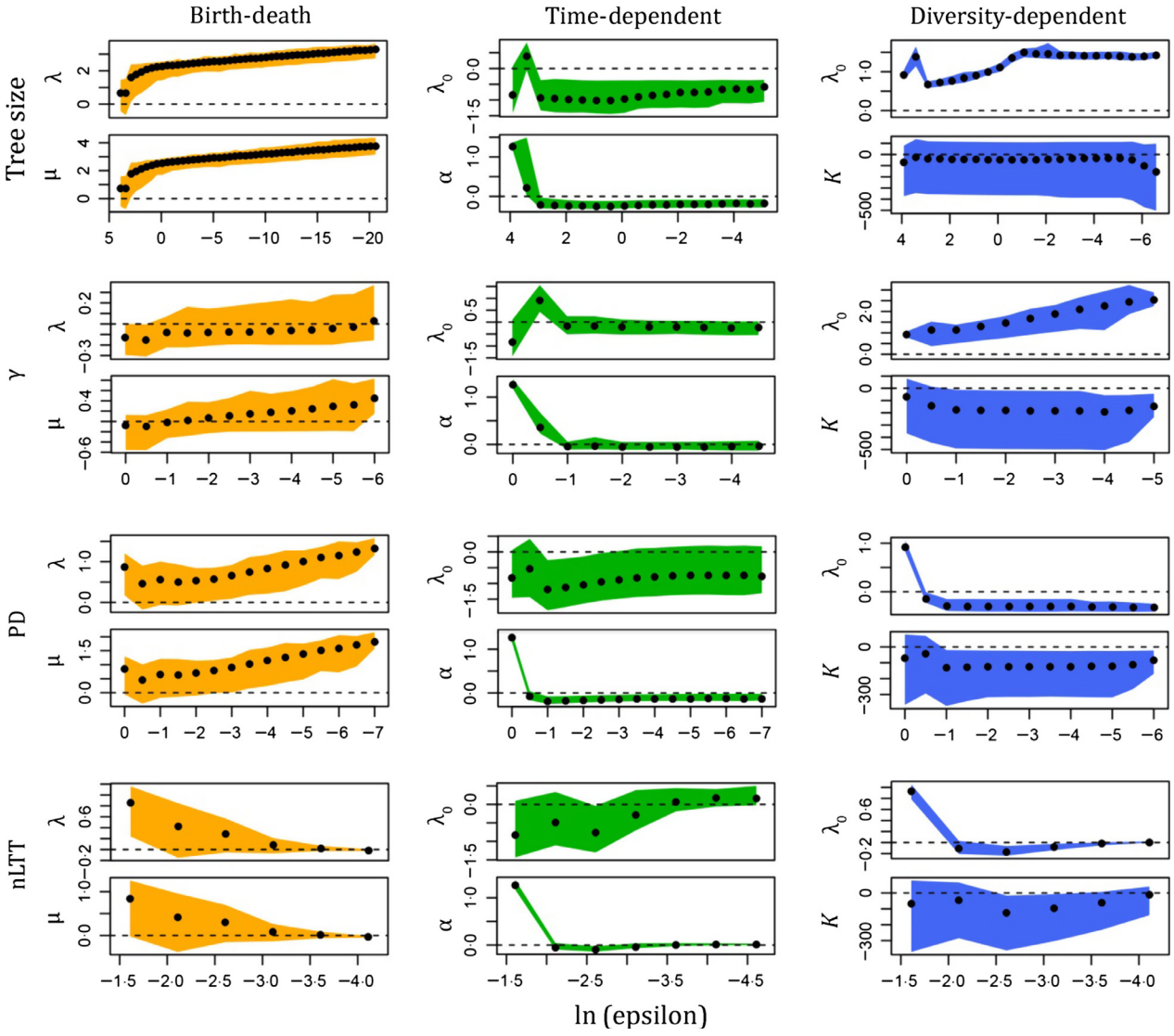


**Fig. 4.** Convergence plots of all four summary statistics for one representative parameterization of each model. On the *x*-axis the ln(epsilon) of the ABC-SMC algorithm is plotted, with on the *y*-axis the residual estimate after correcting for the likelihood-based estimate. Shaded areas indicate the 95% confidence interval across 30 replicates. Chosen parameterizations for the models were for the birth–death model, $\mu = 0{\cdot}3\lambda$, for the time-dependent model $\alpha = 0{\cdot}5$ and for the diversity-dependent model $\lambda_0 = 0{\cdot}75$. Results for the other parameterizations were similar and can be found in the supplementary material.

bit smaller than the variance using the likelihood-based approach.

### CONVERGENCE ANALYSIS

Convergence analysis reveals that the normalized LTT statistic behaves sufficiently, that is as the threshold approaches zero, the mean of the obtained posterior distribution approaches the mean of the likelihood posterior distribution (Figs 4, S1, S2 & S3). For the diversity-dependent model, uncertainty in the estimate of the carrying capacity increases when approaching a zero threshold, but the mean remains accurate. The other three summary statistics tend to either not converge at all, or only converge for a specific model [i.e. the Gamma statistic converges adequately for the constant-rate birth–death model (Fig. S1) and one parameterization of the time-dependent model (Fig. S2), but not for the diversity-dependent model (Fig. S3)].

## Discussion

We have compared the ability of three established summary statistics and one novel summary statistic to substitute the likelihood within an approximate Bayesian framework. Performances of the summary statistics were evaluated across three different diversification models: the constant-rate birth–death model, time-dependent speciation and diversity-dependent speciation. Across all the scenarios that we evaluated here, the three established summary statistics performed much worse than the likelihood. The

newly introduced normalized LTT statistic, however, was able to perform on par with the likelihood across all three modelling situations. We therefore suggest the nLTT as a suitable summary statistic for use within an Approximate Bayesian Computation framework.

In this paper, we have focused on models for which the likelihood was available (so we could make a comparison) and on inferring two parameters for each model. Although we find that the three established summary statistics carry little to no information about the underlying diversification model, these findings might not apply to other, novel diversification models. We urge that in future applications, the effectiveness of the summary statistics is tested before they are applied to empirical data. Novel applications include the use of a different diversification model, inference of more than two parameters or using a combination of summary statistics. We advocate that for any ABC application of phylogenetic summary statistics, validation of the summary statistics for use in parameter inference or model selection is crucial (Robert *et al*. 2011). Validation should be fairly straightforward by confirming the correct inference of simulation data, generated using a set of parameter combinations characteristic for the model at hand.

Whereas the tree size statistic and the Gamma statistic are connected to a direct interpretation [e.g. observed diversity and deviation from constant-rate diversification (but see Fordyce 2010)], the nLTT statistic is not directly associated with a clear interpretation. The nLTT statistic is primarily suited for comparing trees, rather than informing about a single tree. A possible application of the normalized LTT statistic could lie in the comparison of a normalized LTT curve with a standardized curve, such as a standard exponential increase (e.g. a pure-birth or Yule tree). The Lineage Diversity Index (LDI) measures something fairly similar: the difference between the LTT curve of a phylogeny and the LTT curve of a pure-birth model (Harmon *et al*. 2003). However, the LDI index compares the log of the number of lineages and as a result overemphasizes differences between LTT curves at low diversity, which the nLTT statistic does not.

The three models we used in our analysis generate only trees with balance equal to the equal-rates Markov expectation, and the associated likelihood functions are independent of the tree topology (Lambert & Stadler 2013). Therefore, we have not tested any metrics for expected balance, such as the Sackin, Colless or Beta statistic (Sackin 1972; Colless 1982; Blum & François 2006). When performing ABC for a model with properties that will leave patterns in balance, it will evidently be beneficial to include one of these statistics during analysis. Here, we focused on models for which a likelihood formula was available, in order to make the direct comparison between the summary statistics and the traditional likelihood approach. For models that introduce imbalance, unfortunately no likelihoods have been formulated yet (Lambert & Stadler 2013). The performance of the normalized LTT statistic for models that create unbalanced trees remains to be investigated, but we expect that as long as changes in parameters in such models are not only associated with changes in topology, but also with changes in the sequence of branching times, the normalized LTT statistic will be able to distinguish between different parameterizations.

Phylogenetics has not yet picked up on the possibilities of ABC. With our systematic analysis of summary statistics, we are confident that we have provided a step forward towards the future implementation of more complex models in diversification rate analysis. With the introduction of the normalized LTT statistic, we have provided a valuable alternative to the traditional likelihood that will prove to be an important addition to the Approximate Bayesian's and phylogeneticists' toolbox.

## Acknowledgements

We thank Luke Harmon and one anonymous reviewer for helpful comments. Thijs Janzen and Rampal S. Etienne thank the Netherlands Organisation for Scientific Research (NWO) for financial support. We thank the Center for Information Technology of the University of Groningen for their support and providing access to the Millipede high-performance computing cluster.

## Data accessibility

Computer code for this work has been made available as the R-package 'nLTT', available at http://cran.r-project.org/web/packages/nLTT/index.html.

## References

Barnes, C.P., Silk, D. & Stumpf, M.P.H. (2011) Bayesian design strategies for synthetic biology. *Interface Focus*, **1**, 895–908.

Bazin, E., Dawson, K.J. & Beaumont, M.A. (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 587–602.

Beaumont, M.A. (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–35.

Blum, M. & François, O. (2006) Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, **55**, 685–691.

Blum, M.G.B. & François, O. (2009) Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, **20**, 63–73.

Blum, M.G.B. & Tran, V.C. (2010) HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, **11**, 644–60.

Bokma, F. (2010) Time, species, and separating their effects on trait variance in clades. *Systematic Biology*, **59**, 602–7.

Clarke, K. & Warwick, R. (2001) A further biodiversity index applicable to species lists: variation in taxonomic distinctness. *Marine Ecology Progress Series*, **216**, 265–278.

Colless, D. (1982) Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, **31**, 100–104.

Coyne, J. & Orr, H. (2004) *Speciation*. Sinauer Associates, Sunderland, Massachusetts.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E. & François, O. (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–8.

Davies, T.J., Allen, A.P., Borda-de-Água, L., Regetz, J. & Melián, C.J. (2011) Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. *Evolution*, **65**, 1841–1850.

Del Moral, P., Doucet, A., Jasra, A. & Moral, P. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, **22**, 1009–1020.

Drovandi, C. & Pettitt, A. (2011) Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, **67**, 225–33.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204–13.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.
Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.
FitzJohn, R.G., Maddison, W.P. & Otto, S.P. (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, **58**, 595–611.
Fordyce, J.A. (2010) Interpreting the γ statistic in phylogenetic diversification rate studies: a rate decrease does not necessarily indicate an early burst. *PLoS ONE*, **5**, e11781.
Goldberg, E.E., Lancaster, L.T. & Ree, R.H. (2011) Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, **60**, 451–65.
Gould, S. & Eldredge, N. (1977) Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, **3**, 115–151.
Harmon, L.J., Schulte, J.A., Larson, A. & Losos, J.B. (2003) Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, **301**, 961–4.
Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Höhna, S. (2013) Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. *Bioinformatics*, **29**, 1367–74.
Höhna, S. (2014) Likelihood inference of non-constant diversification rates with incomplete taxon sampling. *PLoS ONE*, **9**, 17–20.
Jabot, F. & Chave, J. (2009) Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology letters*, **12**, 239–48.
Kutsukake, N. & Innan, H. (2013) Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. *Evolution; International Journal of Organic Evolution*, **67**, 355–67.
Lambert, A. & Stadler, T. (2013) Birth-death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theoretical Population Biology*, **90**, 113–28.
Lenormand, M., Jabot, F. & Deffuant, G. (2013) Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, **28**, 2777–2796.
Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–10.
Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **100**, 15324–8.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087.
Morlon, H. (2014) Phylogenetic approaches for studying diversification. *Ecology Letters*, **17**, 508–25.
Morlon, H., Parsons, T.L. & Plotkin, J.B. (2011) Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences, USA*, **108**, 16327–32.
Nee, S., May, R.M. & Harvey, P.H. (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **344**, 305–11.
Paradis, E. (2011) Time-dependent speciation and extinction from phylogenies: a least squares approach. *Evolution*, **65**, 661–72.
Paradis, E., Claude, J. & Strimmer, K. (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
Peters, S. & Foote, M. (2002) Determinants of extinction in the fossil record. *Nature*, **416**, 420–424.
Pigot, A.L., Phillimore, A.B., Owens, I.P.F. & Orme, C.D.L. (2010) The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Systematic Biology*, **59**, 660–73.
Pybus, O. & Harvey, P. (2000) Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 2267–72.
R Core Team (2014) R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.R-project.org
Rabosky, D.L. (2009) Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Systematic Biology*, **58**, 629–40.
Rabosky, D.L. & Lovette, I.J. (2008a) Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 2363–71.
Rabosky, D.L. & Lovette, I.J. (2008b) Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, **62**, 1866–1875.
Raup, D. & Sepkoski, J.J. Jr (1982) Mass extinctions in the marine fossil record. *Science*, **215**, 1501–1503.
Robert, C.P., Cornuet, J.-M., Marin, J.-M. & Pillai, N.S. (2011) Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15112–7.
Sackin, M. (1972) "Good" and "bad" phenograms. *Systematic Biology*, **21**, 225–226.
Slater, G.J., Harmon, L.J., Wegmann, D., Joyce, P., Revell, L.J. & Alfaro, M.E. (2012) Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. *Evolution; International Journal of Organic Evolution*, **66**, 752–62.
Stadler, T. (2011) Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, **108**, 6187–92.
Stadler, T. (2013) Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, **26**, 1203–19.
Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M. & Dessimoz, C. (2013) Approximate Bayesian computation. *PLoS Computational Biology*, **9**, e1002803.
Tavaré, S., Balding, D., Griffiths, R. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M.P.H. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187–202.

*Received 13 October 2014; accepted 14 January 2015*
*Handling Editor: Emmanuel Paradis*

## Supporting Information

Additional Supporting Information may be found in the online version of this article.

**Fig. S1.** Convergence plots of all four summary statistics for all parameterizations of the birth-death model. On the *x*-axis the ln(epsilon) of the ABC-SMC algorithm is plotted, and on the *y*-axis the residual estimate after correcting for the likelihood-based estimate.

**Fig. S2.** Convergence plots of all four summary statistics for all parameterizations of the time-dependent model. On the *x*-axis the ln(epsilon) of the ABC-SMC algorithm is plotted, and on the *y*-axis the residual estimate after correcting for the Maximum Likelihood estimate.

**Fig. S3.** Convergence plots of all four summary statistics for all parameterizations of the diversity-dependent model. On the *x*-axis the ln(epsilon) of the ABC-SMC algorithm is plotted, and on the *y*-axis the residual estimate after correcting for the likelihood-based estimate.