BMC Evolutionary Biology

METHODOLOGY ARTICLE Open Access

A Bayesian framework to estimate diversification rates and their variation through time and space

Daniele Silvestro[1,2,3*†], Jan Schnitzler[1,2†] and Georg Zizka[1,2,3]

Abstract

Background: Patterns of species diversity are the result of speciation and extinction processes, and molecular phylogenetic data can provide valuable information to derive their variability through time and across clades. Bayesian Markov chain Monte Carlo methods offer a promising framework to incorporate phylogenetic uncertainty when estimating rates of diversification.

Results: We introduce a new approach to estimate diversification rates in a Bayesian framework over a distribution of trees under various constant and variable rate birth-death and pure-birth models, and test it on simulated phylogenies. Furthermore, speciation and extinction rates and their posterior credibility intervals can be estimated while accounting for non-random taxon sampling. The framework is particularly suitable for hypothesis testing using Bayes factors, as we demonstrate analyzing dated phylogenies of *Chondrostoma* (Cyprinidae) and *Lupinus* (Fabaceae). In addition, we develop a model that extends the rate estimation to a meta-analysis framework in which different data sets are combined in a single analysis to detect general temporal and spatial trends in diversification.

Conclusions: Our approach provides a flexible framework for the estimation of diversification parameters and hypothesis testing while simultaneously accounting for uncertainties in the divergence times and incomplete taxon sampling.

Background

Patterns of species diversity have been shaped by both speciation and extinction throughout the history of life, and one of the key questions in evolutionary biology is to understand the temporal and spatial dynamics of these processes [1-6]. In addition to the fossil record, molecular phylogenetic data of extant lineages can provide valuable information on the process of diversification in form of branch length and the distribution of divergence times throughout the evolutionary history of a clade. Despite the omission of extinct lineages, it has been shown that differential patterns of speciation and extinction can leave a discernible signature on phylogenetic trees of extant taxa [7,8]. Methodological advances [9-14] as well as the growing number of well sampled, dated molecular phylogenies have generated considerable interest in unraveling the temporal dynamics of species diversification. Indeed, diversification rates have been assessed for a wide range of taxa from the tree of life to address questions concerning rapid radiations [15-18], mass extinction events [19], and differences among lineages [20,21] and geographic regions [11,22,23]. In particular, the identification of potential correlates of speciation and/or extinction rates, either extrinsic [e.g. climate or ecology; [24]] and/or intrinsic [e.g. key innovations; [25-28]], has received increased attention by relating rate shifts to external conditions or the evolution of species' traits.

Nee at al. [9] first applied the generalized birth-death process [29] to molecular phylogenies of extant lineages to extract information on the evolutionary process and proposed a likelihood approach to estimate both speciation and extinction rates (λ and μ, respectively). Given the need to distinguish different modes of diversification (e.g. deviations from constant rates), further approaches have been developed to incorporate rate variation through time [8,12,13,30] and across clades [14,31]. In addition, the original birth-death process was modified to

* Correspondence: dsilvestro@senckenberg.de
† Contributed equally
[1]Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany
Full list of author information is available at the end of the article

correct rate estimates in case of incomplete taxon sampling [32-35].

While Bayesian Markov chain Monte Carlo (MCMC) methods are now commonly employed in phylogenetics to accommodate for uncertainties in model parameters, the temporal uncertainty of node age estimates is usually not taken into account when studying the dynamics of species diversification, resulting in an erroneous impression of precision. Here, we present a novel MCMC approach to estimate rates of speciation and extinction over the posterior distribution of trees generated in Bayesian molecular clock analyses. Several models of diversification have been implemented, including the constant rate birth-death and pure-birth processes modified to account for incomplete taxon sampling [33], a birth-death process with continuously varying rates [8], and a pure-birth process with rate shifts [12], for which the posterior distribution of λ and the temporal position of each rate shift are jointly estimated. Within the Bayesian framework, we describe a meta-analysis approach that aims at evaluating general patterns of species diversification across different taxonomic groups. In addition to the estimation of rate parameters, the approach presented here can also be used to distinguish between different modes of diversification, and test explicit hypotheses of rate variation through time and between clades using Bayes factors. We assess the power of the MCMC and of the Bayes factor test using simulated data sets, and demonstrate the application on empirical data sets: rate variation through time in the diversification of Mediterranean cyprinid genus *Chondrostoma*, geographic patterns in the radiation of the genus *Lupinus*, and a meta-analysis of four clades from the Cape flora of South Africa.

Results

Bayesian rate estimation across phylogenies

The birth-death process was implemented in a Markov chain Monte Carlo framework to estimate the parameters of species diversification (speciation and extinction rate) while accounting for phylogenetic uncertainty. Several modifications of the birth-death process originally described by Nee et al. [9] were implemented in the MCMC algorithm to describe different patterns of diversification and allow model selection and hypothesis testing. We included a modification of the birth-death process that accounts for incomplete taxon sampling based on Yang and Rannala [33] and Stadler [34] in which the fraction of the sampled species out of the total diversity (ρ) is used to correct the estimate of the diversification parameters. Although the missing species are assumed to be randomly distributed within the phylogeny, unlike in other models [e.g. [35]], we incorporate an option to assign different sampling fractions to predefined clades (see partitioned models below). In addition, Rabosky and Lovette's [8] SPVAR model was implemented to analyze the commonly observed pattern of "explosive-early" radiations, in which clades show an initial burst of diversification followed by a gradually declining speciation rate. Another model that accounts for rate variation through time is a pure-birth process in which a fixed number of shifts in diversification rate is assumed [12]. In contrast to the continuously varying birth-death process, the rate is assumed to vary only at specific times and otherwise remain constant. For a given number of rate shifts, the MCMC estimates their temporal position and the respective rates. Finally, our approach can be used to assign independent rates to predefined clades and is especially intended for hypothesis testing, complementing other approaches in which the rate constancy across-clades is relaxed [31], or in which the number and position of the rate shifts on the tree are estimated [14]. The data set is partitioned a priori by defining clades of interest, based for example on morphology or biogeography, and independent rates, models, and sampling proportions can be assigned to each clade.

F-model: A meta-analysis approach

We develop a new method to investigate the strength and significance of general patterns of species diversification across different taxonomic groups through time or between clades in a meta-analysis framework. Within a collection of N data sets $d_1, d_2, ..., d_N$ (e.g. phylogenies of different taxonomic groups), each data set d_i is partitioned *a priori* into two time frames or clades $d_i^{(p)}$ and $d_i^{(q)}$. The definition of these partitions can be based on criteria that are applicable to all phylogenies analyzed e.g. geologic events or geographic distribution. Their respective speciation rates $\lambda_i^{(p)}$ and $\lambda_i^{(q)}$ are described as a function of a multiplier m_i and a parameter F so that:

$$\lambda_i^{(p)} = \frac{2m_i F}{1+F} \tag{1a}$$

$$\lambda_i^{(q)} = \frac{2m_i}{1+F} \tag{1b}$$

where $m_i = 1/2(\lambda_i^{(p)} + \lambda_i^{(q)})$ represents a taxon specific mean rate that is assumed independent for each data set d_i, and $F = \lambda^{(p)}/\lambda^{(q)}$ is constrained to be equal for all data sets and quantifies the overall magnitude of the rate difference between the two partitions. Based on these definitions, the rates are equal when $F = 1$, whereas $\lambda^{(p)} > \lambda^{(q)}$ with F greater than 1, and $\lambda^{(p)} < \lambda^{(q)}$ with F smaller than 1. We use MCMC sampling to obtain posterior estimates of the parameters $m_1, m_2, ..., m_N$ and F from the joint likelihood of all data sets L_D

$$L_D = \prod_{i=1}^{N} L\left[d_i^{(p)}; \lambda_i^{(p)}\right] L\left[d_i^{(q)}; \lambda_i^{(q)}\right] \tag{2}$$

Proposals for the parameters m are sampled from normal distributions centered on their current values, whereas new values of the F parameter are obtained from a log-normal distribution to achieve a symmetric proposal distribution in log(F). Reflection at the boundary was used to avoid proposals outside of the valid range (e.g. $m \leq 0$). Uniform priors are assigned to m in range [0, 10] and to log(F) in range [-2.3, 2.3], which corresponds to an F value in range [0.1, 10]. The clade-specific F-model can be extended to a birth-death process by assigning the parameter m to the mean net diversification (r) and introducing a second parameter n for the mean extinction fraction (a). Consequently, two parameters F_r and F_a are defined to measure the overall variation of r and a between clades of each data set. The significance of an overall rate difference across partitions is assessed via Bayes factor between a model in which F is allowed to vary and a model with constrained $F = 1$ (i.e. equal rates across partitions).

Model selection using Bayes factor

Our analyses on simulated data sets show that the power of the Bayes factor test (BF) in finding the correct model is generally very high and not particularly affected by the model settings. Bayes factors were calculated between the model used to simulate each data set and a range of possible alternative models (Table 1) based on their respective marginal likelihoods (Additional file 1) obtained through thermodynamic integration [36-38]. Positive Bayes factors (Table 1) allow to correctly distinguish between diversification models in the majority of the simulations even in data sets with very low taxon sampling. Only when the extinction fraction is low (10%), the pure birth model obtains a slightly higher marginal likelihood than the birth-death. With variable rate pure-birth models, the number of rate shifts is correctly estimated when the magnitude of rate variation is moderate (five-fold) or higher. The effect of extinction and an increase in speciation rate in absence of extinction, both resulting in a similar pattern of increasing net diversification through time, can be distinguished with intermediate to high extinction fraction (> 50%) or a moderate (> two-fold) rate increase. Furthermore, the power of Bayes factors in model selection improves with the size of the phylogeny (Table 1). For instance in case of a small (two-fold) increase in the diversification rate, the correct model is found only on larger phylogenies (100 taxa).

Rate estimation on simulated phylogenies

Analyses on simulated data sets indicate that the MCMC has a rather short burn-in phase and achieves a good chain mixing (measured as Effective Sample Size) with 110, 000 generations and a sampling frequency of 100. The posterior estimates of the speciation rate under the different models of diversification are found to be accurate with a relative error generally below 10% (Tables 2, 3, and 4). The relative error drops below 5% in data sets with 100 or more taxa, indicating that the size of the phylogeny has an impact on the accuracy of the parameter estimation. In addition the width of the rates' credibility intervals decreases with increasing size of the phylogeny: The HPDs are on average 25% narrower with 100 taxa compared to 50, and further reduce by another 50% with 400 tips (Additional file 2). In addition, about one third of the width of the 95% HPD is due to accounting for phylogenetic uncertainty (Figure 1).

The constant rate birth-death model yields accurate posterior estimates of the speciation rate λ (Table 2) and efficient measures of the extinction rate are obtained when the extinction fraction is high ($a = 0.9$). The accuracy of the estimate, however, decreases substantially when the extinction is low ($a = 0.1$). This is likely due to the MCMC sampling, which is constrained by the fact that μ cannot become negative [32]. This results in a strongly skewed posterior distribution for which the mean is a poor estimator; a more accurate estimate is in this case provided by the mode. The 95% credibility interval of the posterior rates is always wide for μ (0 - 0.47 with $\mu = 0.05$, $a = 0.1$, and 0.19 - 0.71 with $\mu = 0.45$, $a = 0.9$; Additional file 2). While the sampling proportion does not significantly affect the accuracy in the estimation of λ and μ (Table 2), it has a strong impact on the width of the credibility intervals. The size of the 95% HPD increases by 20, 30, and 50 percent with $\rho = 0.75$, 0.5, and 0.25, respectively, compared to the complete data set (Figure 2).

The model with continuously decreasing diversification rates (SPVAR) yields accurate estimates of the parameter k (which determines the magnitude of the temporal decrease of the speciation rate; Table 3). On the other hand, the estimated initial speciation rate λ_0 tends to be overestimated when k is small (with relative errors between 0.2 and 0.3), and underestimated for higher k values.

Estimates of the speciation rates in data sets with rate-shifts were found to be accurate with a relative error on average lower than 0.1 (Table 4). The marginal rates estimated within 1 Myr time frames, reflect the rate variation through time (Figure 3). For time frames in which a rate shift occurs, the marginal rate is often represented by a bimodal distribution, which reflects the uncertainty on the temporal placement of the shift and results in an intermediate rate estimate with a wider 95% credibility interval (Figure 3A). This uncertainty is reflected in the frequency distribution of the rate shift in the posterior sample (Figure 3B). A highly accurate estimate of the time of rate shift is provided by the modal value of its sampling frequency, with relative errors lower than 0.05 (Table 4).

Table 1 Bayes factors (BF) tests to distinguish different modes of diversification

Simulation settings			Bayes Factors (TDI)				
no. of tips (ρ)	λ	μ	BD	PB	PB2	PB3	PB4
50	0.5	0.05	**0**	-0.71	0.17		
100	0.5	0.05	**0**	-2.40	-1.77		
50	0.5	0.25	**0**	2.59	1.51		
100	0.5	0.25	**0**	3.66	-0.21		
50	0.5	0.45	**0**	24.98	3.57		
100	0.5	0.45	**0**	36.68	5.06		
50	0.5	0	1.05	**0**	1.18		
100	0.5	0	2.92	**0**	0.69		
100 (25%)	1	0	2.21	**0**			
200 (50%)	1	0	4.10	**0**			
300 (75%)	1	0	5.33	**0**			
400	1	0	6.37	**0**			
100 (25%)	1	0.9	**0**	36.65			
200 (50%)	1	0.9	**0**	68.30			
300 (75%)	1	0.9	**0**	98.18			
400	1	0.9	**0**	131.57			
50	0.1, 0.2	0	-0.59	1.74	**0**	1.76	
100	0.1, 0.2	0	0.44	6.78	**0**	0.88	
50	0.05, 0.25	0	4.64	23.73	**0**	1.22	
100	0.05, 0.25	0	5.79		**0**	0.73	
50	0.02, 0.16	0	13.94	40.60	**0**	0.69	
100	0.02, 0.16	0	29.53	90.60	**0**	0.93	
50	0.2, 0.1	0	4.02	0.54	**0**	1.47	
100	0.2, 0.1	0	11.57	5.57	**0**	1.53	
50	0.5, 0.1	0	23.22	18.52	**0**	1.49	
100	0.5, 0.1	0	58.15	51.49	**0**	0.81	
50	0.16, 0.02	0	23.92	19.51	**0**	0.85	
100	0.16, 0.02	0	56.32	49.53	**0**	0.60	
50	0.1, 0.2, 0.1	0	-0.55	-2.26	-1.28	**0**	2.04
100	0.1, 0.2, 0.1	0	4.81	0.35	0.69	**0**	1.52
50	0.1, 0.5, 0.1	0	12.93	12.56	8.26	**0**	0.67
100	0.1, 0.5, 0.1	0	45.38	40.22	21.47	**0**	-0.10
50	0.02, 0.16, 0.02	0	22.34	21.49	18.18	**0**	-1.11
100	0.02, 0.16, 0.02	0	63.95	64.01	54.67	**0**	-1.31

Bayes factors are calculated under birth-death (BD), pure-birth (PB), and pure-birth with rate shifts (PB2-PB4) based on thermodynamic integration. The BF values are calculated between the model applied in the simulation and the alternative models: Positive values support the true model.

Contrasting times of rate shift: Diversification of Mediterranean cyprinids

The use of a pure-birth process with rate shift and its implementation in hypothesis testing are illustrated in an analysis of the cyprinid genus *Chondrostoma* (Teleostei: Cyprinidae). A recently published molecular phylogeny of the genus [39] places the origin of the present lineages in the mid-Miocene around 15 Mya. Two alternative hypotheses on the diversification of *Chondrostoma* have been proposed, placing its radiation in the Mediterranean region either during the Messinian salinity crisis [40] or earlier in the Miocene [41]. The comparison of different models of diversification using Bayes factor tests led to the selection of a two-rate pure-birth process and estimated a fourfold decrease in speciation rates (dropping from an initial 0.441 to 0.108), and indicating a substantial slowdown during the Miocene. We used alternative two-rate pure-birth models to specifically test the fit of a rate shift during the Messinian [5.33 - 7.25 Mya; [40]] or earlier in the Miocene [7.25 - 23.03 Mya; [41]]. The rate shift was constrained in two separate analyses to lie within those periods, and the two models were compared by approximating a Bayes factor. Robalo et al. [42] favored the latter hypothesis based on their molecular clock analysis, although without specifically testing it in a statistical framework. Our analysis suggests that the

Table 2 Estimates of speciation (λ) and extinction (μ) rates from simulated phylogenies

Simulation settings			Estimates	
no. of tips (ρ)	λ	μ	λ_{MEAN} (rel. error)	μ_{MEAN} (rel. error)
50	0.5	0	0.47 (-0.07)	-
100	0.5	0	0.48 (-0.04)	-
50	0.5	0.05	0.57 (0.14)	0.23 (3.61)
100	0.5	0.05	0.55 (0.10)	0.18 (2.55)
50	0.5	0.25	0.49 (-0.03)	0.28 (0.11)
100	0.5	0.25	0.52 (0.03)	0.30 (0.20)
50	0.5	0.45	0.49 (-0.02)	0.43 (-0.04)
100	0.5	0.45	0.49 (-0.02)	0.44 (-0.03)
100 (25%)	1	0	1.09 (0.09)	-
200 (50%)	1	0	1.07 (0.07)	-
300 (75%)	1	0	1.06 (0.06)	-
400 (100%)	1	0	1.05 (0.05)	-
100 (25%)	1	0.9	0.99 (-0.02)	0.82 (-0.09)
200 (50%)	1	0.9	1.02 (0.02)	0.85 (-0.05)
300 (75%)	1	0.9	1.04 (0.03)	0.88 (-0.03)
400 (100%)	1	0.9	1.04 (0.04)	0.89 (-0.01)

Rates were inferred using the constant rate pure-birth or birth-death model, averaged over 100 phylogenies for each simulation. The taxon sampling (ρ) of incomplete phylogenies and relative errors are reported in parenthesis.

Messinian Lago Mare phase had no particular effect on the radiation of the genus *Chondrostoma* (BF = 3.14), as a significant decrease in the speciation rate has to be placed before that period, thus supporting Robalo et al.'s conclusion.

Clade-specific analysis: geographic patterns in the radiation of *Lupinus* (Fabaceae)

We demonstrate the application of models in which the rates can vary between predefined clades by analyzing the geographic patterns of diversification in *Lupinus* (Fabaceae) [16]. The phylogeny of the genus *Lupinus* shows a strong geographic structure which we used to define four partitions: I) an early diverging Old World clade, II) a group of eastern New World taxa, III) a clade occurring mainly in western North America and Central America, and IV) a clade including most of the Andean species (Figure 4). The latter displays a radiation from which around 80 species arose in the past 1.5 million years. While the model with equal rates among clades was strongly rejected in favor of variable rate models (Table 5), the highest marginal likelihood was assigned to a model with three different rates assigned to the clades I, II+III, and IV, respectively. The posterior distributions of the diversification rates are plotted as relative densities (Figure 4), showing a four-fold variation in speciation rate between the Old World *Lupinus* (λ_I = 0.191) and the non-Andean New World lineages (λ_{II+III} = 0.687). The Andean clade, as described by Hughes and Eastwood [16], represents an explosive radiation with a posterior rate estimate of λ_{IV} = 2.510.

A meta-analysis approach: Diversification of the Cape flora, South Africa

The high and unique plant diversity of the Cape Floristic Region (CFR) of South Africa is the result of an extraordinary contribution of lineages, that radiated extensively in the CFR [so called 'Cape floral clades'; [43]]. The high diversity and endemism suggest that Cape clades may have diversified at a faster rate within the CFR than elsewhere. Valente et al. [23] however recently showed that in the genus *Protea*, diversification rates in the Cape were, if anything, lower than in neighboring regions. To test for a general rate difference between Cape and non-Cape clades, we analyzed four data sets [23,44], all containing clades distributed within and outside the CFR, representing in total 537 plant species. The posterior rates estimated for the Cape/non-Cape clades of each individual data set are: *Babiana* 0.500/0.556, *Moraea* 0.259/0.288, Podalyrieae 0.150/0.167, and *Protea* 0.195/0.218 (Figure 5B). The *F* parameter is estimated as 1.118 (95% HPD 0.86 - 1.39), indicating that diversification rates are overall slightly higher outside of the Cape (Figure 5A) region. However a Bayes factor of 4.78 between the constrained *F*-model with *F* = 1 (i.e. no rate difference between clades) and the unconstrained model suggests that this difference is not significant and that equal rates should be preferred. These results indicate an overall rate uniformity between Cape and non-Cape clades based on the four data sets analyzed, suggesting that the great diversity in the CFR might not be the result of a faster diversification process. It should be noted however that, as pointed out by Valente et al. [23], species ranges in

Table 3 Parameter estimation for the continuously varying birth-death process

Simulation settings				Estimates		
no. of tips	λ	μ	*k*	λ_{MEAN} (rel. error)	μ_{MEAN} (rel. error)	k_{MEAN} (rel. error)
50	1	0.1	0.25	1.91 (0.91)	0.50 (3.99)	0.23 (-0.08)
100	1	0.1	0.25	1.68 (0.68)	0.34 (2.38)	0.19 (-0.25)
50	5	0	0.95	3.72 (-0.26)	0.31 (-)	0.75 (-0.21)
100	5	0	0.95	5.22 (0.04)	0.33 (-)	0.97 (0.02)

Estimates of speciation (λ) and extinction (μ) rates and the *k* parameter are inferred using the variable rate birth-death model (SPVAR), averaged over 100 phylogenies. Relative errors are given in parenthesis.

Table 4 Rate estimates for the pure-birth process with rate shifts

Simulation settings			Estimates	
no. of tips	λ	*s*	λ_{MEAN} (rel. error)	s_{MODE}
50	0.1, 0.2	5	0.11 (0.12), 0.24 (0.22)	4.86
100	0.1, 0.2	5	0.10, 0.22 (0.12)	4.81
50	0.05, 0.25	3.5	0.05, 0.26	3.59
100	0.05, 0.25	5	0.05 (0.07), 0.26	4.97
50	0.02, 0.16	5	0.02 (0.07), 0.17 (0.07)	4.95
100	0.02, 0.16	5	0.02, 0.16	5.05
50	0.2, 0.1	5	0.19, 0.12 (0.17)	5.09
100	0.2, 0.1	7	0.21, 0.11	6.97
50	0.5, 0.1	3	0.50, 0.11 (0.12)	3.06
100	0.5, 0.1	5	0.51, 0.10	5.04
50	0.16, 0.02	5	0.16, 0.03 (0.31)	5.05
100	0.16, 0.02	5	0.16, 0.02 (0.19)	5.01
50	0.1, 0.2, 0.1	3, 7	0.12 (0.22), 0.19 (-0.06), 0.12 (0.22)	3.01, 6.96
100	0.1, 0.2, 0.1	5, 10	0.10, 0.19 (-0.06), 0.11 (0.08)	4.88, 10.04
50	0.1, 0.5, 0.1	2, 4	0.11 (0.07), 0.32 (-0.36), 0.12 (0.21)	2.05, 3.95
100	0.1, 0.5, 0.1	2, 6	0.12 (0.18), 0.50, 0.11 (0.10)	1.94, 6.05
50	0.02, 0.16, 0.02	15, 20	0.03, 0.15, 0.03 (0.41)	15.49, 19.77
100	0.02, 0.16, 0.02	15, 20	0.02, 0.16, 0.02 (0.10)	15.48, 20.58

Speciation rates (λ) and temporal position of rate shifts (*s*) are inferred under the variable rate pure-birth model, averaged over 100 phylogenies for each simulation. The marginal rates are estimated as the mean of their posterior distributions, positions of rate shifts are estimated as the modal values of their posterior distributions. Relative errors are given in parenthesis if higher than 0.05.

these regions are vastly different, indicating that the key to understanding the Cape biodiversity hotspot instead lies in understanding why so many lineages have speciated and persisted in such a small area.

Implementation

The method described has been implemented in a computer program called "BayesRate" (available at http://sourceforge.net/projects/bayesrate/ or from the authors) written in Python [45] (based on the Numpy [46] and Scipy [47] libraries) and R [48], integrating codes using the Python module rpy2 [49]. The thermodynamic integration supports multi-core computation by simultaneously running individual Markov chains on different processors. The log files can be examined with the

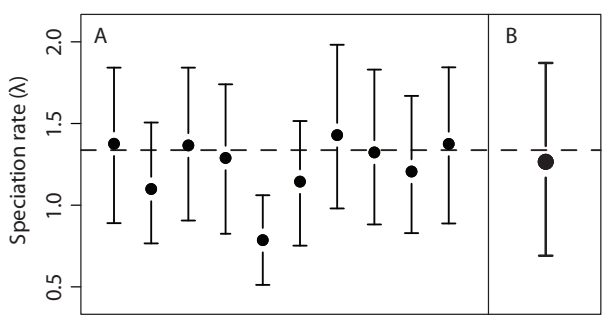

Figure 1 Effect of accounting for phylogenetic uncertainty on rate estimates. Comparison of the speciation rate estimated on 10 trees randomly sampled from the posterior distribution (A), and averaged over 100 trees (B). The dashed line indicates the maximum likelihood estimate based on the consensus tree. Error bars represent the 95% highest posterior density (HPD) interval of the rate estimates. Accounting for phylogenetic uncertainty results in an average increase of the width of the 95% HPD by 30% (B).

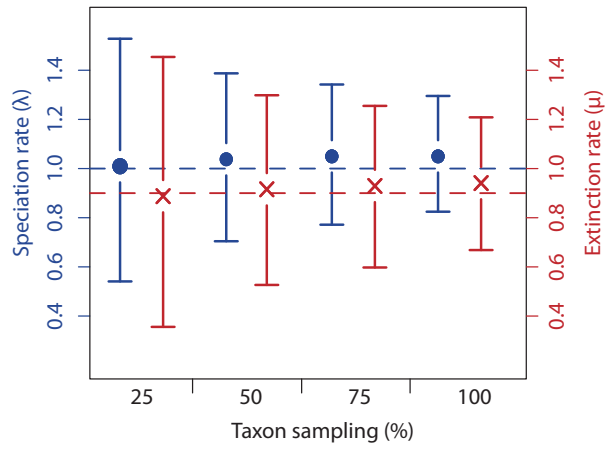

Figure 2 Effect of taxon sampling on rate estimates. Speciation (blue circles) and extinction (red crosses) rates estimated on simulated data sets of 400 taxa with a taxon sampling of 25%, 50%, 75%, and 100%, respectively. Phylogenetic trees were simulated with a speciation and extinction rate of 1 and 0.9, respectively (indicated by the dashed lines). Decreasing taxon sampling is correlated with an increase in the 95% highest posterior density (HPD) interval, whereas the accuracy of the rate estimate largely remains unaffected.

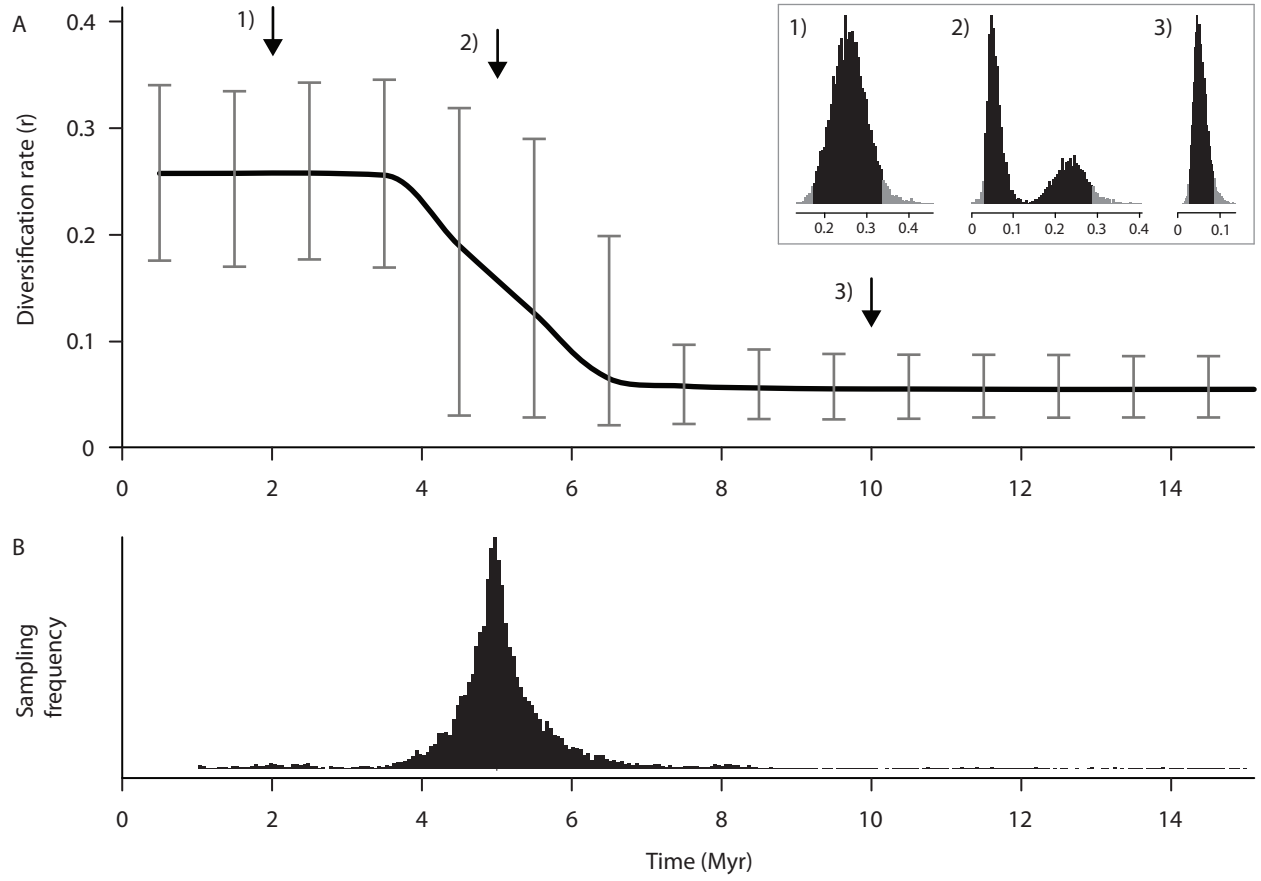

Figure 3 Rates-through-time plot. Diversification rates through time resulting from the analysis of 100 phylogenies simulated under a fivefold increase in diversification rates. The upper plot (A) shows the marginal rates for 1 Myr time categories (line) and the 95% highest posterior density (error bars). The x-axis represents time (Myr), and the y-axis is the average per-lineage diversification rate (spp/Myr). The insert displays three examples of marginal distributions of the diversification rate for three points along the phylogenies (indicated by arrows on the rates through time plot): 1) close to the tips (2 Mya), 2) at the point of rate shift (5 Mya), and 3) towards to root of the trees (10 Mya). Note the bimodal distribution of rates when a rat-shift is found (both the lower and higher rates are sampled). In the lower plot (B), the frequencies of a rate shift are proportional to the probability of a rate shift in that time frame.

program Tracer [50] to check for efficiency of the sampling (ESS), and convergence between independent runs.

Discussion

We presented a new Bayesian approach, which provides a powerful tool to estimate rates of speciation and extinction on dated phylogenies based on the likelihood functions of the pure-birth, and birth-death processes while accounting for phylogenetic uncertainty and incomplete taxon sampling. On empirical data, our rate estimation requires a two steps analysis: 1) sampling a posterior distribution of dated phylogenies using a Bayesian molecular clock approach, and 2) estimating posterior diversification rates on these trees. Available programs such as BEAST [51] and mcmctree [52] apply in their relaxed molecular clock implementation different birth-death processes as priors on the node ages [53,54]. When applied to phylogenies obtained under these assumptions, our approach therefore requires that priors on the diversification parameters are specified twice, while ideally divergence times and diversification rates should be estimated jointly. The majority of the diversification processes considered here are however currently not implemented in these programs, and thus need to be estimated independently. Analyses on simulated data show that the default uniform priors on net diversification and extinction fraction in BEAST do not affect the subsequent rate estimates (relative rate variations lower than 3%; Additional file 3). Alternatively, researchers could choose to run molecular clock analyses in which the prior on the node ages is not based on a birth-death process, but modeled using uniform or Dirichlet distributions [55,56], as implemented in e.g. Multidivtime [57] and PhyloBayes [58].

The range of models implemented can be used to detect a number of different scenarios of rate variation, including specific events of rate increase or decrease,

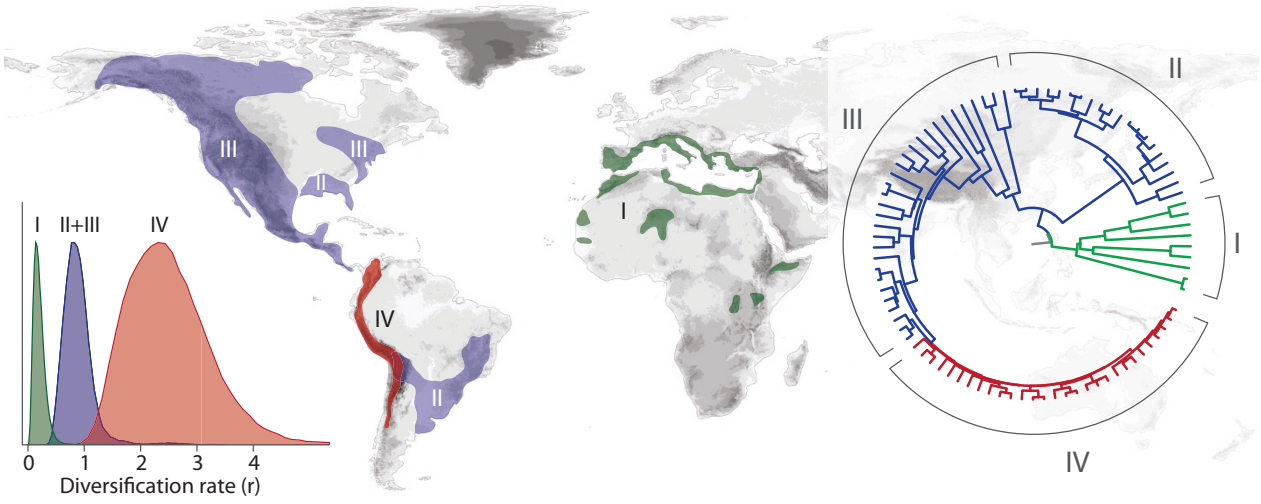

Figure 4 Clade-specific analysis of the radiation of the genus *Lupinus*. Clades in the phylogenetic tree are labeled according to their distribution: I - Old World (Europe and North Africa), II - Eastern New World, III - Western North America and Central America, IV - Andean. Colors indicate the mean rates of diversification, with posterior rate estimates (relative densities) shown in the plot on the left (I - 0.20, II+III - 0.89, IV - 2.56). Note that the best model identified using Bayes Factors is a three-rate model in which clades II and III evolve at the same rate.

continuous rate variation through time, and clade-specific diversification rates, while simultaneously accounting for taxon sampling. In addition, the Bayesian framework extends the use of the birth-death models beyond the simple rate estimation allowing the comparison of alternative scenarios of diversification for hypothesis testing.

The Bayes factor test computed via thermodynamic integration has shown to represent a reliable and powerful approach to choose among different models of diversification. BF can be applied to compare non-nested models and does not require to be explicitly corrected for the number of model parameters. For these properties it is particularly suitable not only to select the best model in a Bayesian framework, but also to compare specific hypotheses. The power of Bayes factors to detect the correct number of rates predominantly depends on the magnitude of the rate shifts. Similarly, different birth-death and pure-birth processes can generate diversification patterns that might be difficult to distinguish [4,8]. For instance, an increase in the net diversification rate can be the result of an increased speciation rate in the absence of extinction or a high extinction rate in a constant rate birth-death process. Nevertheless, we found that the Bayes factors test has the power to discern between most of such scenarios.

Table 5 Model comparison in *Lupinus*

no. rates	partition settings	L_M	BF
1	$\lambda_{I+II+III+IV}$	-184.46	77.01
2	$\lambda_{I+II+III}$, λ_{IV}	-157.46	23.02
3	λ_{I}, λ_{II+III}, λ_{IV}	-145.96	0
4	λ_{I}, λ_{II}, λ_{III}, λ_{IV}	-146.08	0.25

The marginal likelihoods (L_M) of models with different partition settings are estimated via thermodynamic integration. Log Bayes factors (BF) are calculated by comparing the best fitting model (with 3 parameters) against all the others.

Analyses on simulated data show that for both speciation and extinction rates the posterior estimates are accurate. However, the width of the 95% HPD intervals also highlights the sometimes considerable uncertainty in the parameter estimates, especially in case of small phylogenies. Our simulations have shown that this uncertainty is most pronounced with a low relative extinction rate, in which case extinction tends to be overestimated [see also [32]], and estimates have a wide 95% credibility interval. This corroborates previous studies that pointed out that the estimation of extinction rates from molecular phylogenies with reasonable degrees of confidence is very problematic [7,12,59,60]. It should be noted, however, that the wide credibility intervals reflect not only the uncertainties of the parameter estimation, but also the uncertainty of the data (i.e. the node ages). Estimates of the speciation rate on the other hand appear to be more robust. In contrast to Paradis [60], we find that the posterior estimate of λ has a small relative error, even under high relative extinction rates, suggesting that the accuracy in the estimates of λ might be decoupled from the relative rate of extinction. The pure-birth and birth-death models with taxon sampling are found to provide accurate rate estimates, although a poor sampling yields substantially wider 95% HPDs. Finally, the pure-birth process with rate-shifts tends to slightly underestimate the true variation of λ, as a consequence of accounting for the uncertainty of the time of rate shift.

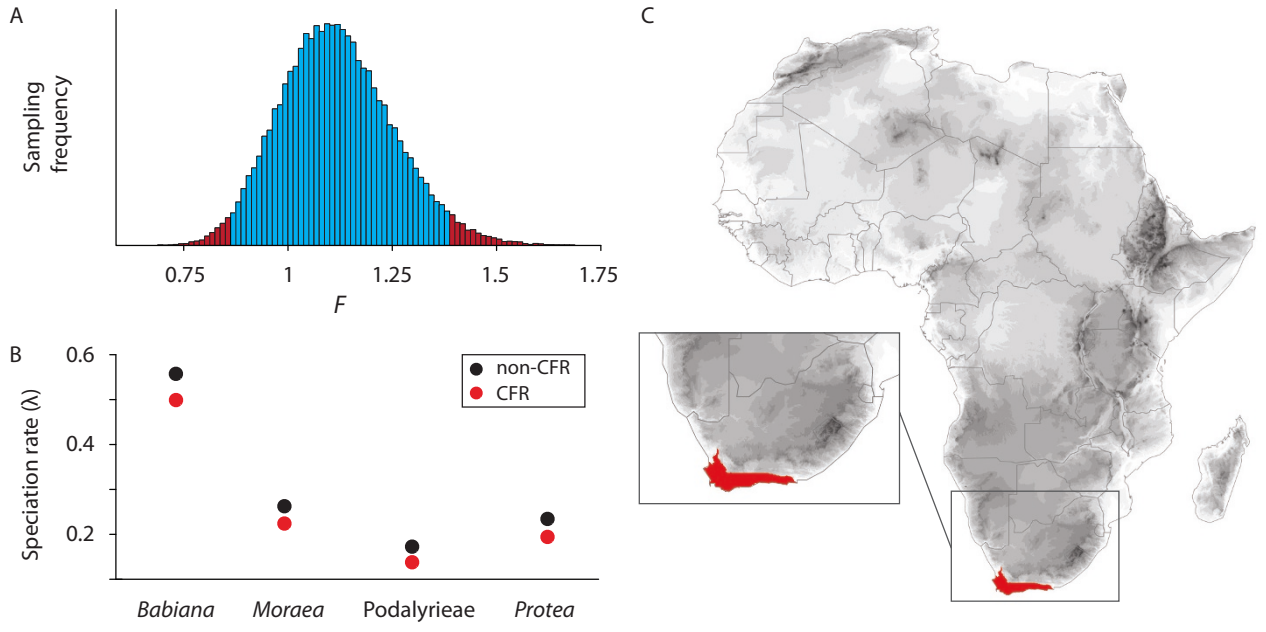

Figure 5 Meta-analysis on plant diversification in the Cape Floristic Region (CFR). The map (C) shows the location of the CFR (highlighted in red) at the southwestern tip of the African continent. Rates of speciation (B) within the CFR (red) and in the rest of Africa (black) for each data set are found almost identical within and outside the CFR (*Babiana* 0.5/0.556, *Moraea* 0.259/0.288, Podalyrieae 0.15/0.167, and *Protea* 0.195/0.218). The posterior density of the parameter $F = 1.118$ (95% HPD 0.86 - 1.39), (A) indicates overall similar rates in the two regions.

The approach provides a very flexible framework for customized analyses and hypothesis testing such as predefined times of rate shift or constrained parameter values. We have shown with the diversification of *Chondrostoma*, that the pure-birth model with variable rates can be easily adapted to test for specific hypotheses, running the analysis on fixed time frames defined for example on the basis of geological events or climate changes. The implementation of clade-specific rate estimation further extends the range of options for hypothesis testing, and its application on the radiation of *Lupinus* showed that it can be used to identify differential rates of diversification between clades. In particular, the option to account for clade-specific sampling biases provides an important feature, as complete taxon sampling is often difficult to achieve, especially for species-rich groups.

Finally, with the *F*-model, we introduce a new approach to test hypotheses in a meta-analysis framework, and extend the focus from the taxon-specific rate of diversification to a parameter that might be linked to the difference between e.g. geologic periods or geographic regions. The current implementation allows to compare hypotheses with two rates, assigned to either fixed time frames or clades, while accounting for clade-specific taxon sampling. Because the *F* parameter is constant across data sets, we assume that the magnitude of the rate variation in time or between clades is equal among all data sets. While this certainly represents a simplification of the diversification process, the *F*-model allows the analyses of potentially many data sets, limiting the number of parameters, and yielding an estimation of general trends across different taxonomic groups. Moreover, even relatively small rate variations can be detected if supported by a sufficient number of data sets.

Conclusions

In summary, the approach presented here shows that temporal dynamics of species diversification resulting from biologically relevant events such as key innovations or the impact of environmental change should best be studied in a Bayesian framework. The use of MCMC sampling provides an elegant way to estimate speciation and extinction rates while taking into account the often considerable uncertainty on divergence times. Furthermore, the model with taxon sampling represents an important step towards a more realistic estimation of the diversification parameters, where a non-random distribution of missing taxa can be incorporated with clade-specific sampling proportions. In addition to the models implemented in this study, recently developed modifications of the birth-death process [13,61] could also be integrated in the algorithm. With the possibility to run customized analyses specifically designed for hypothesis testing, this method provides a useful and flexible statistical framework to investigate diversification processes. A promising future development would be to relax the *F*-model to incorporate more than

two rates, by assigning a specific multiplier to each time frame or group of clades.

Methods

Bayesian estimation of the diversification parameters

The likelihood of a birth-death process (BD) describing the speciation and extinction events of a dated phylogeny can be written as a function of the branching times $\boldsymbol{x}$, the number of extant species s, and the speciation and extinction rates λ and μ, respectively [9]:

$$L(x;\lambda,\mu) = (s-1)!(\lambda-\mu)^{s-2}\exp\left((\lambda-\mu)\sum_{i=3}^{s}x_i\right) \times\left(1-\frac{\mu}{\lambda}\right)^{s}\prod_{i=2}^{s}\left((\exp(\lambda-\mu)x_i)-\frac{\mu}{\lambda}\right)^{-2} \quad (3)$$

The function reduces to a pure-birth process (PB) in the absence of extinction (μ = 0).

We implemented this likelihood function in a Markov Chain Monte Carlo framework and applied the Metropolis-Hastings algorithm [62,63] to sample the posterior distribution of the birth-death model parameters. The algorithm is structured as follows:

1. Assign initial values to the model parameters (e.g. λ, μ)
2. Sample new r, a values (from which new λ, μ are obtained)
3. Accept or reject the proposal based on the acceptance probability
4. Repeat steps 2 and 3 many times
5. Repeat steps 1 to 4 over different trees sampled from their posterior distribution
6. Summarize the MCMC over all sampled trees by calculating mean and credibility interval for each parameter of interest

The MCMC iteration starts with random parameter values and successive proposals for λ and μ (step 2) are based on the sampling strategy described by Bokma [32], randomly drawing values of $r = \lambda - \mu$ (net diversification) and $a = \mu/\lambda$ (extinction fraction) from normal distributions centered on their current values. To avoid proposals lying outside of the valid interval (e.g. negative values) we use reflection at the boundary. The acceptance probability is proportional to the likelihood ratio, and uniform distributions are applied as flat priors on the rates. The MCMC is run over a distribution of trees, sampling λ and μ on each tree individually after a burnin phase (step 5) and the parameters of interest are summarized over all trees to account for the uncertainty on the node ages (step 6). The means of the posterior distributions of λ and μ are used as rate estimates and the respective credibility intervals are calculated as the 95% highest posterior density (HPD) intervals.

Assuming that a number of species are missing in a phylogeny, the missing lineages can be modeled as the result of an extinction event that occurs exactly at the present time [33]. Thus, when only a subset s of the total species S is included in the phylogeny, the likelihood of a set of branching times becomes a function of the proportion of sampled species $\rho = s/S$. This model assumes that taxon sampling is random with respect to the phylogeny. However, in case of a non-random sampling bias, individual clades in the phylogeny are represented to different extents. Thus, pure-birth and birth-death models were implemented in the MCMC framework with the possibility to assign a different sampling proportion (ρ) to each clade.

An approach to measure the variation of speciation and extinction rates through time has been introduced by Rabosky and Lovette [8], to model high initial rates of diversification followed by gradually declining net diversification rates. Their maximum likelihood method uses an exponential transformation of λ and μ through time with the introduction of two additional parameters, namely k and z, which specify the magnitude of λ decrease and μ increase, respectively:

$$\lambda(t) = \lambda_0 \exp(-kt) \quad (4a)$$

$$\mu(t) = \mu_0 (1 - \exp(-zt)) \quad (4b)$$

where λ_0 is the initial speciation rate, and μ_0 the final extinction rate. A constant speciation rate is found with $k = 0$, whereas the extinction tends to be constant when z is very large. We implement Rabosky and Lovette's [8] SPVAR model (where speciation rates decrease through time while extinction rates remain constant) by applying a uniform prior in range [0, 10] for k and setting z to 10, 000. The parameters sampled by the algorithm are λ_0, μ, and k.

The assumption of a pure-birth process (μ = 0) simplifies equation (3) as described by Kendall [29] and Nee et al. [9], and a likelihood-based approach has been described to detect shifts in diversification rates through time [12]. We implement this variable rate pure-birth model in which, given a number of rate shifts n, the estimated parameters are the temporal position of the shifts $\boldsymbol{s} = s_1, s_2, \ldots, s_n$, and the corresponding rates $\boldsymbol{\lambda} = \lambda_1, \lambda_2, \ldots, \lambda_{n+1}$. Proposals for λ and s are sampled from normal distributions centered on their current values. The likelihood ratio is based on the product of the likelihoods of the branching times x_i within each time frame delimited by s_{i-1}, s_i under the rate λ_i:

$$L(x;\lambda) = \prod_{i=1}^{n} L(x_i;\lambda_i) \quad (5)$$

A uniform prior from 0 to the root age is assumed for the temporal position of the rate shift *s*. Because the temporal position of the rate-shift is not fixed but estimated through the MCMC sampling, we summarize the marginal rates as mean value and 95% credibility interval within predefined time frames (e.g. 1 Myr intervals) from the root age to the tips of the trees and use these estimates to draw rates-through-time plots (RTT). Thus, the marginal rates reflect the uncertainty on the time of rate shift. The sampling frequencies of the rate shifts through time are used to infer the temporal placement of the rate variation events. We identify the time of rate shift by finding the modal value of the frequency distribution of each parameter s_i, representing the most frequently sampled value and approximating the *maximum-a-posteriori* estimate (MAP).

When testing for rate differences across predefined clades, the joint likelihood of all clades *C* is used to estimate the posterior distribution of the speciation and extinction rates (λ_c, μ_c) of each individual clade c. The parameters λ_c and μ_c can be constrained to be equal among clades (linked model) or estimated independently (unlinked model). A model comparison between linked and unlinked parameters is performed using Bayes factors (see below) to assess the significance of the rate difference between clades.

Model selection: Bayes factors via thermodynamic integration

The fit of different models of diversification was assessed by comparing their respective marginal likelihoods, which are defined as the probability of the data *D* conditional on the model *M*, p(*D*|*M*). Alternative models can be compared using the Bayes factor test, which is defined as the ratio between their respective marginal likelihoods [64,65]. Calculating the marginal likelihood involves the integration of the probability of the data over the entire parameter space Θ:

$$L_M = p(D \mid M) = \int_{\theta_i} p(D \mid \theta_i, M)\, p(\theta_i \mid M) d\theta_i \tag{6}$$

Several approaches have been described to approximate L_M. One simple approximation of L_M is obtained as the harmonic mean of the likelihood values sampled via MCMC [65,66]. Although commonly used for model comparison in phylogenetics [e.g. [67,68]], the harmonic mean estimator has been found unstable, and thus often unreliable [37]. An alternative approach is thermodynamic integration (TDI) or path sampling [36-38], that has been shown to provide more accurate estimates of the marginal likelihood and has recently been applied in phylogenetics [37,69-71] and population genetics [72]. This method allows the exploration of regions of the parameters space with low likelihood by altering the acceptance ratio of the MCMC by a scaling factor β. The scaling factor ranges from 0 to 1 and is applied as an exponent to the likelihood function so that with $\beta = 0$ the MCMC samples from the prior distribution only, and with $\beta = 1$ the distribution of interest is sampled. The marginal likelihood L_M is then obtained by integrating the likelihood expectations E_β over all values of β:

$$L_M = \int_0^1 \mathrm{E}_\beta \ln p(D \mid \theta, M)\, d\beta \tag{7}$$

We discretize the integral by using a number *C* of scaling factors β_0, β_1, ... β_C evenly spaced from 0 to 1, and estimating the respective log-likelihood expectations $U_{\beta i}$ as the mean of the MCMC sample. The discrete integral is then calculated applying Simpson's trapezoidal rule:

$$L_M = \sum_{i=2}^{\mathrm{C}} \frac{1}{2} (\beta_i - \beta_{i-1}) (U_{\beta_i} + \mathrm{U}_{\beta_i - 1}) \tag{8}$$

The accuracy of this discrete approximation of (9) depends on the number of classes *C* that can represent a limiting factor since the computational time increases linearly with the number of categories. Beerli and Palczewski [72] showed that a highly accurate estimate of L_M can be obtained with a small number of scaling factors when integrating analytically over the first interval [β_0, β_1] using Bézier cubic spline. This approach uses control points P_1 and P_2 based on the likelihood expectations at the first three scaling factors U_0, U_1, U_2:

$$P_1 = \left(\beta_0, \frac{1}{5}U_0 + \frac{4}{5}U_1\right) \tag{9a}$$

$$P_2 = \left(\beta_0, \frac{\beta_1 U_2 - \beta_2 U_1}{\beta_1 - \beta_2}\right) \tag{9b}$$

The four control points are used to define the cubic Bézier curve $B_{P0,\ P1,\ P2,\ P3}$, with the first and the last points being

$$P_0 = (\beta_0, U_0), P_3 = (\beta_1, U_1) \tag{10}$$

The integral of the marginal likelihood over the interval [β_0, β_1] is then calculated as

$$\begin{aligned} L_{M(\beta_0,\beta_1)} &= \int_{\beta_0}^{\beta_1} B_{P_0,P_1,P_2,P_3} d\beta \\ &= \frac{1}{20}\left((\beta_1 - \beta_0)\left(U_0 + 3c_\gamma^{(0)} + 6c_\gamma^{(1)} + 10U_1\right)\right) \end{aligned} \tag{11}$$

We found that the shape of the Bézier spline described by Beerli and Palczewski [72] provided a good

approximation of the curve obtained by applying many scaling factors under different models of diversification, and therefore adopted it in our computation of the marginal likelihood. After testing various numbers of scaling classes to calculate the discrete thermodynamic estimate of the marginal likelihood (not shown), six classes were found to be a good compromise between accuracy of the result and computational time. Once the log marginal likelihoods L_M were obtained via TDI, the log Bayes factor (BF) between pairs of models M_0 and M_1 was computed as $BF_{01} = 2(M_1 - M_0)$ and its interpretation based on the values suggested by Kass and Raftery [65]. Thus BF_{01} greater than 2 represent positive evidence for model M_1, and greater than 6 provide strong evidence. To assess the power of Bayes factor in discerning between different modes of diversification, we analyzed several simulated data set under birth-death models and pure-birth assuming one to three rate shifts with a special focus on processes that generate similar patterns (e.g. birth-death and pure-birth with rate increase).

Statistical evaluation

To test the performance of our method, we analyzed simulated phylogenies generated under a range of models using the R-package TreeSim [34,73]. A total of 38 data sets of 100 phylogenies (with 50, 100, or 400 tips) were simulated under different models of diversification (Additional file 4). We simulated constant rate birth-death models with extinction fractions ranging from low to very high (0.1, 0.5, and 0.9), and different taxon sampling proportions (25%, 50%, 75%, and 100%). Pure-birth processes were simulated with either constant rates, or including shifts in diversification rates (one or two shifts) under small (twofold), moderate (fivefold), and large (eightfold) rate variations, respectively (Additional file 4). Trees (50 and 100 tips) with approximately continuously decreasing speciation rates were obtained by imposing nine equally spaced rate shifts, under two different diversification scenarios where speciation rates follow an exponential decrease ($\lambda_0 = 1$, $\mu = 0.1$, and $k = 0.25$; $\lambda_0 = 5$, $\mu = 0$, and $k = 0.95$). Because of the limitations of the SPVAR [8] model under variable or high extinction rates [74,75], we assumed absent or very low and constant extinction. As these simulations only approximate continuously decreasing rates, we report the parameter estimates under the SPVAR model, but do not perform model comparisons via Bayes factors.

To assess the accuracy of the rate estimates, we calculated the relative errors [cf. [12]] as $(r_{est} - r_{true})/r_{true}$, where r_{est} is the estimated rate of speciation or extinction and r_{true} is the true value. A positive relative error indicates overestimation of the parameter, whereas a negative value indicates its underestimation. For the pure-birth model with rate variation, the marginal diversification rates through time were calculated for time categories of 1 million years, and their relative errors were calculated in relation to the true values between shift points. The modal values of the posterior distribution of the shift points were compared against the true shift times and their relative error was calculated as $(t_{est} - t_{true})/T$, where t_{est} is the estimated time of rate shift, t_{true} is its true value, and T is the average root node age of the analyzed trees.

To address the impact of estimating rates on a single tree compared to analyzing a distribution of trees, we used a tree topology simulated in Phyl-o-Gen [76] under the birth-death process (100 tips; r = 1; $a = 0.9$) to simulate nucleotide sequences (3978 bp, HKY+I+Γ) using the program SeqGen [77]. Phylogenetic trees were then reconstructed in BEAST [v.1.6.1; [51]]. For comparison, we also inferred the maximum likelihood estimates of λ and μ on the consensus tree (Figure 3) through a birth-death optimization as implemented in LASER [78].

To empirically assess the potential impact of specifying priors on the birth-death parameters in both the molecular clock analysis and the subsequent rate estimation, additional simulations were performed. We generated trees in Phyl-o-Gen (50 tips; extinction fraction $a = 0$, 0.5, and 0.9) on which nucleotide sequences (5000 bp, HKY+I+Γ) were simulated using the program SeqGen [77]. Dated phylogenies were reconstructed in BEAST using the default uniform priors on the birth-death parameters and constraining the root node to the age of the initial tree. The posterior rates were then estimated using our MCMC approach on both the initial tree (used to simulate the alignment) and the distribution of trees obtained from BEAST. The rate estimates were compared by calculating their variation in terms of relative error (Additional file 3).

Analyses on the case studies

The phylogenetic relationships of the genus *Chondrostoma* were reconstructed using mitochondrial cytochrome *b* and nuclear ß-actin gene sequences for all currently recognized taxa [39]. Phylogenetic trees and divergence times were reconstructed using BEAST [v.1.6.1; [51]] and assuming the GTR+I+Γ model of sequence evolution. A speciation model following a Yule process was selected as the tree prior, with an uncorrelated lognormal (UCLN) model for the rate variation among branches. Secondary calibration points were used, following Gante et al. [39], constraining nodes to a normal prior: the crown node of *Chondrostoma* was constrained with a mean of 15.1 Mya (central 95% range 12.5 - 17.6 Mya). The split between *C. olisiponensis* and its sister clade was constrained with a mean of 10.1 Mya (central 95% range 7.7 - 12.4 Mya). The analysis was run for

15 million generations, sampling states every 2, 000 generations. The adequacy of the sampling was assessed with Tracer [50] using the Effective Sample Size diagnostic (Additional file 2). We evaluated the temporal patterns of diversification using all diversification models implemented in our approach applied on a random sample of 100 trees from the molecular clock analysis.

We reconstructed the phylogeny and divergence times of the genus *Lupinus* based on a combined alignment of ITS and LEGCYC1. Following Hughes and Eastwood [16], the two markers were partitioned and analyzed under GTR+Γ and GTR+I models, respectively. A relaxed molecular clock analysis was carried out using BEAST, assuming an uncorrelated lognormal clock model, running 30 million MCMC generations. A normal distribution with a mean of 16.01 Mya and a standard deviation of 2.6 for the stem node of *Lupinus* was set as calibration point. We carried out the diversification analyses on a random sample of 100 trees obtained from a relaxed molecular clock analysis, applying a pure-birth process after model selection. The estimation of the diversification rates was performed assuming clade-specific taxon sampling (ρ_I = 0.77, ρ_{II} = 0.55, ρ_{III} = 0.18, ρ_{IV} = 0.40) under different models in which the rates were linked or unlinked among clades.

The meta-analysis on the Cape Floristic Region was based on four dated phylogenies of different Cape clades [23,44]. The analyses using the *F*-model were performed on a random set of 100 trees sampled from the posterior distribution of each data set. All clades have on average only a small proportion of missing taxa, which was accounted for by means of clade-specific sampling fractions. Each data set was split into Cape and non-Cape clades based on their main geographic distribution, high degrees of endemism - particularly within the CFR - allowed a simple assignment of clades (Figure 5C). The meta-analysis was performed implementing pure-birth models, which were favored over a birth-death process with a Bayes factor value of 3.46.

Additional material

Additional file 1: Marginal Likelihoods for different models of diversification. Marginal Likelihoods for simulated data sets calculated under birth-death (BD), pure-birth (PB), and pure-birth with rate shift (PB2-PB4) models based on thermodynamic integration.

Additional file 2: Posterior rate estimates. Parameter estimates, 95% credibility intervals and ESS values for all data sets simulated. All simulations settings are provided in Additional file 4.

Additional file 3: Effect of sequential estimation of divergence times and diversification rates. The potential impact of specifying priors on the birth-death parameters in both the molecular clock analysis and the subsequent rate estimation is assessed through generating a starting tree, simulating a molecular alignment on it, and run BEAST analyses on the alignment. The rate are then estimated on both the starting tree and the BEAST posterior trees, and compared.

Additional file 4: List of the simulation settings. Simulations were obtained from birth-death (BD), pure-birth (PB), pure-birth with rate shift (PB2-PB4), and (approximately) continuously decreasing speciation rates (SPVAR). The parameters included are speciation rates (λ), extinction rate (μ), time of rate shift (*s*), sampling fraction (ρ), and the shape parameter of the exponential transformation of λ through time (*k*). The continuously decreasing rates (SPVAR model) were approximated by imposing nine equally spaced rate shifts where speciation rates follow an exponential decrease. The value of λ reported for the SPVAR model represents the initial speciation rate (λ_0).

Acknowledgements
We thank Nicolas Salamin and Bob O'Hara for valuable discussions, Hugo F. Gante and Colin Hughes for providing the alignments for *Chondrostoma* and *Lupinus*, respectively, Ziheng Yang for help implementing the birth-death model with taxon sampling, Fernando Fernandez-Mendoza and Peter Beerli for aid with the thermodynamic integration. We also thank David Posada, Folmer Bokma, and Jeffrey L. Thorne for helpful comments on an earlier version on the manuscript. All simulations were run on the Frankfurt Cloud Server, we thank Kai Bosch and Matthew Forrest for support. This study was funded by Hesse's Ministry of Higher Education, Research, and the Arts (funding program "LOEWE - Landes-Offensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz") and the German Research Foundation (DFG ZI 557/7-1, SCHU 2426/1-1).

Author details
[1]Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [2]Department of Botany and Molecular Evolution, Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [3]Diversity and Evolution of Higher Plants, Institute of Ecology, Evolution and Diversity, Goethe University, Senckenberganlage 31, 60325 Frankfurt am Main, Germany.

Authors' contributions
DS, JS, and GZ designed the study. DS and JS developed the Bayesian framework and the models. DS wrote the Python code, JS designed and performed the simulations and wrote the R scripts. DS and JS wrote the paper, all authors read and approved the final version of the manuscript.

Received: 23 April 2011 Accepted: 21 October 2011
Published: 21 October 2011

References

1. Nee S, Mooers AØ, Harvey PH: **Tempo and mode of evolution revealed from molecular phylogenies.** *Proc Natl Acad Sci USA* 1992, **89(17)**:8322-8326.
2. Sanderson MJ, Donoghue MJ: **Reconstructing shifts in diversification rates on phylogenetic trees.** *Trends Ecol Evol* 1996, **11(1)**:15-20.
3. Barraclough TG, Nee S: **Phylogenetics and speciation.** *Trends Ecol Evol* 2001, **16(7)**:391-399.
4. Ricklefs RE: **Estimating diversification rates from phylogenetic information.** *Trends Ecol Evol* 2007, **22(11)**:601-610.
5. Jablonski D, Roy K, Valentine JW, Price RM, Anderson PS: **The impact of the pull of the recent on the history of marine diversity.** *Science* 2003, **300(5622)**:1133-1135.
6. Jaramillo C, Rueda MJ, Mora G: **Cenozoic plant diversity in the Neotropics.** *Science* 2006, **311(5769)**:1893-1896.
7. Nee S, Holmes EC, May RM, Harvey PH: **Extinction rates can be estimated from molecular phylogenies.** *Phil Trans R Soc B* 1994, **344(1307)**:77-82.
8. Rabosky DL, Lovette IJ: **Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time?** *Evolution* 2008, **62(8)**:1866-1875.
9. Nee S, May RM, Harvey PH: **The reconstructed evolutionary process.** *Phil Trans R Soc B* 1994, **344**:305-311.
10. Magallón S, Sanderson MJ: **Absolute diversification rates in angiosperm clades.** *Evolution* 2001, **55(9)**:1762-1780.
11. Ricklefs RE: **Global variation in the diversification rate of passerine birds.** *Ecology* 2006, **87(10)**:2468-2478.

12. Rabosky DL: **Likelihood methods for detecting temporal shifts in diversification rates.** *Evolution* 2006, **60(6)**:1152-1164.
13. Morlon H, Potts MD, Plotkin JB: **Inferring the dynamics of diversification: a coalescent approach.** *PLoS Biol* 2010, **8(9)**:e1000493.
14. Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ: **Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates.** *Proc Natl Acad Sci USA* 2009, **106(32)**:13410-13414.
15. Baldwin BG, Sanderson MJ: **Age and rate of diversification of the Hawaiian silversword alliance (Compositae).** *Proc Natl Acad Sci USA* 1998, **95(16)**:9402-9406.
16. Hughes C, Eastwood R: **Island radiation on a continental scale: Exceptional rates of plant diversification after uplift of the Andes.** *Proc Natl Acad Sci USA* 2006, **103(27)**:10334-10339.
17. Valente LM, Savolainen V, Vargas P: **Unparalleled rates of species diversification in Europe.** *Proc R Soc Lond B* 2010, **277(1687)**:1489-1496.
18. Day JJ, Cotton JA, Barraclough TG: **Tempo and mode of diversification of lake Tanganyika cichlid fishes.** *PLoS ONE* 2008, **3(3)**:e1730.
19. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: **The delayed rise of present-day mammals.** *Nature* 2007, **446**:507-512.
20. Magallón S, Castillo A: **Angiosperm diversification through time.** *Am J Bot* 2009, **96(1)**:349-365.
21. Phillimore AB, Freckleton RP, Orme CDL, Owens IPF: **Ecology predicts large-scale patterns of phylogenetic diversification in birds.** *Am Nat* 2006, **168(2)**:220-229.
22. Linder HP: **Plant species radiations: where, when, why?** *Phil Trans R Soc B* 2008, **363**:3097-3105.
23. Valente L, Reeves G, Schnitzler J, Pizer Mazon I, Fay M, Rebelo T, Chase M, Barraclough T: **Diversification of the African genus *Protea* in the Cape biodiversity hotspot and beyond: equal rates but different spatial scales.** *Evolution* 2010, **64(3)**:745-760.
24. Rabosky DL: **Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions.** *Ecol Lett* 2009, **12(8)**:735-743.
25. Ree RH: **Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis.** *Evolution* 2005, **59(2)**:257-265.
26. Hodges SA, Arnold ML: **Spurring plant diversification: Are floral nectar spurs a key innovation?** *Proc R Soc Lond B* 1995, **262**:343-348.
27. Maddison WP, Midford PE, Otto SP: **Estimating a binary character's effect on speciation and extinction.** *Syst Biol* 2007, **56(5)**:701-710.
28. Moore BR, Donoghue MJ: **A Bayesian approach for evaluating the impact of historical events on rates of diversification.** *Proc Natl Acad Sci USA* 2009, **106(11)**:4307-4312.
29. Kendall DG: **On the generalized birth-and-death process.** *Ann Math Stat* 1948, **19(1)**:1-15.
30. Stadler T: **Mammalian phylogeny reveals recent diversification rate shifts.** *Proc Natl Acad Sci USA* 2011, **108(15)**:6187-6192.
31. Rabosky DL: **Extinction rates should not be estimated from molecular phylogenies.** *Evolution* 2010, **64(6)**:1816-1824.
32. Bokma F: **Bayesian estimation of speciation and extinction probabilities from (in)complete phylogenies.** *Evolution* 2008, **62(9)**:2441-2445.
33. Yang Z, Rannala B: **Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method.** *Mol Biol Evol* 1997, **14(7)**:717-724.
34. Stadler T: **On incomplete sampling under birth-death models and connections to the sampling-based coalescent.** *J Theor Biol* 2009, **261(1)**:58-66.
35. Höhna S, Stadler T, Ronquist F, Britton T: **Inferring speciation and extinction rates under different species sampling schemes.** *Mol Biol Evol* 2011, **28(9)**:2577-2589.
36. Gelman A, Meng X-L: **Simulating normalizing constants: from importance sampling to bridge sampling to path sampling.** *Statistical Science* 1998, **13(2)**:163-185.
37. Lartillot N, Philippe H: **Computing Bayes factors using thermodynamic integration.** *Syst Biol* 2006, **55(2)**:195-207.
38. Ogata Y: **A Monte-Carlo method for high dimensional integration.** *Numer Math* 1989, **55(2)**:137-157.
39. Gante HF, Santos CD, Alves MJ: **Phylogenetic relationships of the newly described species *Chondrostoma olisiponensis* (Teleostei: Cyprinidae).** *J Fish Biol* 2010, **76(4)**:965-974.
40. Bianco PG: **Potential role of the palaeohistory of the Mediterranean and Paratethis basins on the early dispersal of Euro-Mediterranean freshwater Wshes.** *Ichthyol Explor Freshwaters* 1990, **1**:167-184.
41. Bănărescu P: **Zoogeography of fresh waters. Volume 2: distribution and dispersal of freshwater animals in North America and Eurasia.** Wiesbaden: AULA-Verlag; 1991.
42. Robalo JI, Almada VC, Levy A, Doadrio I: **Re-examination and phylogeny of the genus *Chondrostoma* based on mitochondrial and nuclear data and the definition of 5 new genera.** *Mol Phylogenet Evol* 2007, **42(2)**:362-372.
43. Linder HP: **The radiation of the Cape flora, southern Africa.** *Biol Rev (Camb)* 2003, **78**:597-638.
44. Schnitzler J, Barraclough TG, Boatwright JS, Goldblatt P, Manning JC, Powell MP, Rebelo T, Savolainen V: **Causes of plant diversification in the Cape biodiversity hotspot of South Africa.** *Syst Biol* 2011, **60(3)**:343-357.
45. Python Core Development Team: **Python programming language v.2.6.4.** 2010 [http://www.python.org/].
46. Ascher D, Dubois PF, Hinsen K, Hugunin J, Oliphant T: **Numerical Python.** *UCRL-MA-128569* Livermore, CA 94566: Lawrence Livermore National Laboratory; 2001.
47. Jones E, Oliphant T, Peterson P: **SciPy: Open source scientific tools for Python.** 2001 [http://www.scipy.org].
48. R Development Core Team: **R: A language and environment for statistical computing v.2.13.1.** R Foundation for Statistical Computing; 2011 [http://www.R-project.org/].
49. Gautier L: **rpy2: A simple and efficient access to R from Python.** 2008 [http://rpy.sourceforge.net/].
50. Rambaut A, Drummond AJ: **Tracer v.1.5.** 2007 [http://beast.bio.ed.ac.uk/Tracer].
51. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7**:214.
52. Yang ZH: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24(8)**:1586-1591.
53. Gernhard T: **The conditioned reconstructed process.** *J Theor Biol* 2008, **253(4)**:769-778.
54. Yang ZH, Rannala B: **Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds.** *Mol Biol Evol* 2006, **23(1)**:212-226.
55. Thorne JL, Kishino H: **Divergence time and evolutionary rate estimation with multilocus data.** *Syst Biol* 2002, **51(5)**:689-702.
56. Lepage T, Bryant D, Philippe H, Lartillot N: **A general comparison of relaxed molecular clock models.** *Mol Biol Evol* 2007, **24(12)**:2669-2680.
57. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of molecular evolution.** *Mol Biol Evol* 1998, **15(12)**:1647-1657.
58. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21(6)**:1095-1109.
59. Kubo T, Iwasa Y: **Inferring the rates of branching and extinction from molecular phylogenies.** *Evolution* 1995, **49(4)**:694-704.
60. Paradis E: **Can extinction rates be estimated without fossils?** *J Theor Biol* 2004, **229(1)**:19-30.
61. Rabosky DL: **Primary controls on species richness in higher taxa.** *Syst Biol* 2010, **59(6)**:634-645.
62. Hastings WK: **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika* 1970, **57(1)**:97-109.
63. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AW, Teller E: **Equations of state calculations by fast computing machines.** *J Chem Phys* 1953, **21(6)**:1087-1091.
64. Jeffreys H: **Some tests of significance, treated by the theory of probability.** *P Camb Philos Soc* 1935, **31(2)**:203-222.
65. Kass RE, Raftery AE: **Bayes Factors.** *J Amer Stat Assoc* 1995, **90(430)**:773-795.
66. Newton MA, Raftery AE: **Approximate Bayesian inference with weighted likelihood bootstrap.** *J Roy Stat Soc B* 1994, **56(1)**:3-48.
67. Nylander JAA, Ronquist F, Huelsenbeck J, Nieves-Aldrey JL: **Bayesian phylogenetic analysis of combined data.** *Syst Biol* 2004, **53(1)**:47-67.
68. Pagel M, Meade A: **Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo.** *Am Nat* 2006, **167(6)**:808-825.
69. Rodrigue N, Aris-Brosou S: **Fast Bayesian choice of phylogenetic models: Prospecting data augmentation-based thermodynamic integration.** *Syst Biol* 2011.

70. Fan Y, Wu R, Chen M, Kuo L, Lewis P: **Choosing among partition models in Bayesian phylogenetics.** *Mol Biol Evol* 2011, **28(1)**:523-532.
71. Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H: **Improving marginal likelihood estimation for Bayesian phylogenetic model selection.** *Syst Biol* 2011, **60(2)**:150-160.
72. Beerli P, Palczewski M: **Unified framework to evaluate panmixia and migration direction among multiple sampling locations.** *Genetics* 2010, **185(1)**:313-326.
73. Hartmann K, Wong D, Stadler T: **Sampling trees from evolutionary models.** *Syst Biol* 2010, **59(4)**:465-476.
74. Bokma F: **Problems detecting density-dependent diversification on phylogenies.** *Proc R Soc Lond B* 2009, **276(1659)**:993-994.
75. Rabosky DL, Lovette IJ: **Problems detecting density-dependent diversification on phylogenies: reply to Bokma.** *Proc R Soc Lond B* 2009, **276(1659)**:995-997.
76. Rambaut A: **Phyl-O-Gen. Phylogenetic Tree Simulator Package v.1.1.** University of Oxford; 2002 [http://tree.bio.ed.ac.uk].
77. Rambaut A, Grassly NC: **Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees.** *Comput Appl Biosci* 1997, **13(3)**:235-238.
78. Rabosky DL: **LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies.** *Evol Bioinform Online* 2006, **2**:257-260.

doi:10.1186/1471-2148-11-311
Cite this article as: Silvestro *et al.*: A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evolutionary Biology* 2011 **11**:311.