A New Method for Handling Missing Species in Diversification Analysis Applicable to Randomly or Nonrandomly Sampled Phylogenies

NATALIE CUSIMANO, TANJA STADLER, AND SUSANNE S. RENNER

Correspondence to be sent to: Systematic Botany and Mycology, University of Munich (LMU), Menzinger Str. 67, 80638 Munich, Germany; E-mail: renner@lrz.uni-muenchen.de.

Abstract.—Chronograms from molecular dating are increasingly being used to infer rates of diversification and their change over time. A major limitation in such analyses is incomplete species sampling that moreover is usually nonrandom. While the widely used γ statistic with the Monte Carlo constant-rates test or the birth–death likelihood analysis with the ΔAICrc test statistic are appropriate for comparing the fit of different diversification models in phylogenies with random species sampling, no objective automated method has been developed for fitting diversification models to nonrandomly sampled phylogenies. Here, we introduce a novel approach, CorSiM, which involves simulating missing splits under a constant rate birth–death model and allows the user to specify whether species sampling in the phylogeny being analyzed is random or nonrandom. The completed trees can be used in subsequent model-fitting analyses. This is fundamentally different from previous diversification rate estimation methods, which were based on null distributions derived from the incomplete trees. CorSiM is automated in an R package and can easily be applied to large data sets. We illustrate the approach in two Araceae clades, one with a random species sampling of 52% and one with a nonrandom sampling of 55%. In the latter clade, the CorSiM approach detects and quantifies an increase in diversification rate, whereas classic approaches prefer a constant rate model; in the former clade, results do not differ among methods (as indeed expected since the classic approaches are valid only for randomly sampled phylogenies). The CorSiM method greatly reduces the type I error in diversification analysis, but type II error remains a methodological problem. [Birth–death likelihood analysis; diversification rates; missing-species-problem; model fitting; nonrandom species sampling; γ statistic.]

Large time-calibrated phylogenies are now readily obtained and are increasingly being used to infer diversification patterns (Hey 1992; Nee et al. 1992; Sanderson and Bharathan 1993; Harvey et al. 1994; Sanderson and Donoghue 1994; Paradis 1997, 1998; Baldwin and Sanderson 1998; Magallón and Sanderson 2001; Nee 2006; Rabosky 2006b; Rabosky et al. 2007; McPeek 2008; Phillimore and Price 2008; Stadler 2011a). However, inferring rates of diversification is statistically challenging, and the sensitivity of methods when their underlying assumptions are not met is poorly understood. A major problem in diversification analysis is incomplete species sampling (Pybus and Harvey 2000; Cusimano and Renner 2010; Brock et al. 2011; Höhna et al. 2011). This is a common problem when clades are species rich and access to samples is problematic and costly. As a result, phylogenies for large clades are often highly incompletely sampled. Several methods have been proposed that attempt to correct for biases introduced by incomplete sampling. Some of them attempt the correction before the analysis; others attempt correction after the analysis (Nakagawa and Freckleton 2008, for a review of methods for handling missing data). Of the methods that try to correct for missing (not sequenced) species before the analysis, survival analysis (SA; Paradis 1997) adds them as censored events. Alternatively, missing species have been added halfway along the branch where they are thought to belong (Barraclough and Vogler 2002) or to the stem of their clade (Purvis et al. 1995). Another approach is to add missing species to random locations within their clade, using a Markov chain Monte Carlo (MCMC) tree chain (Day et al. 2008, the legend of fig. S1 in this study is misleading in stating that species were added at specific nodes; T. Barraclough, Imperial College, personal communication, 18 August 2009). All these a priori corrections require knowledge about the phylogenetic relationships of the missing species; censoring moreover requires knowing the missing species' minimum ages.

Approaches that correct for missing species after the analysis, that is, after diversification models have been fit to the topology/branching times, involve the creation of a null distribution. For this, one carries out numerous simulations of trees under a null model, with the number of tips corresponding to the complete number of species in the focal clade. Trees are then randomly pruned to the sample size (the number of species actually sequenced), and the pruned data sets are tested for rate constancy, using either the Monte Carlo constant-rates (MCCR) test for the γ statistic (Pybus and Harvey 2000) or the ΔAICrc test statistic for birth–death likelihood (BDL) analyses (Rabosky 2006a). An assumption underlying this approach is that species sampling is random. Nonrandom species sampling introduces strong biases (Cusimano and Renner 2010; Brock et al. 2011; Höhna et al. 2011). Brock et al. (2011) recently presented a method that generates more appropriate null distributions in the MCCR test by introducing a scaling parameter alpha, which allows the degree of nonrandom sampling to be controlled. Determining the scaling parameter, however, is problematic.

Here, we introduce an objective and automated method for handling missing species, which involves simulating missing splits under a constant rate birth–death model, essentially using model-based data augmentation and multiple imputation (Nakagawa and Freckleton 2008). The new method, which we call CorSiM for "Correction by Simulating Missing splits," makes use of information that the user may have about species sampling being random or nonrandom but does not require knowledge about precise species relationships or ages. Simulating the missing species onto an empirical phylogeny results in numerous completed phylogenies that can be used in further diversification analysis and allows calculating confidence intervals around estimates. We apply our new approaches in two plant clades with similarly incomplete species sampling (52% and 55%), one of them randomly incompletely sampled, the other nonrandomly. The investigated clades belong to the Araceae family and occur in the Mediterranean basin and Southeast Asia, regions with different geological histories and present day climates, which sets up an expectation of different diversification patterns during the past 5 million years. Having a nonrandomly and a randomly sampled clade allows us to compare the CorSiM approach with previous methods for inferring diversification rates from incompletely sampled trees, which presupposed random species sampling.

Materials and Methods

Study Systems, Taxon Sampling, and Sequencing

The Areae comprise 153 species in nine genera (Cusimano et al. 2010) and are a tribe of the monocot family Araceae (Cusimano et al. 2011). All Areae are geophytes with a seasonal life cycle. Within Areae, our focal groups are the Typhonium clade with 58 species and the Arum clade with 62 species in five genera (Arum, Biarum, Dracunculus, Eminium, and Helicodiceros). The Arum clade is centered in the Mediterranean basin and the Near East; a few species also occur in cold temperate regions of the Himalayas and in Northern Europe. We henceforth refer to it as the Mediterranean clade. Our phylogeny includes 32 of the 62 species (52%) and is nonrandomly sampled because we included a few species from each of the five genera; we lack 11 species of Arum, 12 of Biarum, and 7 of Eminium. The Typhonium clade occurs in the Southeast Asian mainland tropics and subtropics; we sequenced 32 of its 58 species (55%) and sampling is random. Tree rooting and outgroup sampling is based on Renner and Zhang (2004) and is influenced by the need to include taxa with a fossil record for calibration of genetic distances. Table 1 lists the 16 outgroup taxa with voucher information and GenBank numbers; information about the sequenced ingroup species is provided in Cusimano et al. (2010). The sequenced plastid loci were the rpl20-rps12 intergenic spacer and the tRNALys (UUU) gene (trnK) including its group II intron with the maturase K (matK) gene. For some species, we also sequenced the nuclear phytochrome C gene (PhyC), using the primers of Cusimano et al. (2010).

Divergence Time Estimation

The divergence time estimation relied on Bayesian relaxed clock approach implemented in BEAST version 1.6.1 (Drummond et al. 2006; Drummond and Rambaut 2007). The data matrix included 112 species and 4352 aligned nucleotides (TreeBASE S12261). Analyses used a speciation model that followed a Yule tree prior, with rate variation across branches uncorrelated and lognormally distributed; the substitution model was GTR + Γ + I. Three groups were constrained to be monophyletic; the Pistia clade, the Areae clade (Renner and Zhang 2004), and the Alocasia/Colocasia clade (which is problematic; Cusimano et al. 2011). MCMC chains were run for 10 million generations, with parameters sampled every 1000th generation. The appropriate burn-in fraction was assessed using Tracer version 1.4.1 (http://beast.bio.ed.ac.uk/Tracer) and AWTY (Nylander et al. 2008). We carried out

Table 1. The 16 outgroup taxa used in this study with their herbarium vouchers or accession numbers of living plants and GenBank numbers for the sequenced DNA regions

Species	Herbarium voucher or botanical garden living accession	trnK	rpl20-rps12	PhyC
Alocasia cucullata (Lour.) G. Don	MO living acc. 751658	EU886579	AY248908	—
Alocasia gageana Engl. & K. Krause	MO living acc. 78364	EU886580	AY248909	JQ238980
Alocasia navicularis (Blume) Hook.	T. Croat & V. D. Nguyen 78014 (MO)	EU886581	AY248925	JQ238981
Ariopsis protanthera N.E.Br.	H. Hara leg. 1960 (TI), Nepal	EU886587	AY248910	JQ083567
Arisarum vulgare Targ.Toz.	Bot. Garden Bonn living acc. 11472	EU886582	EU886630	—
Caladium bicolor (Aiton) Vent.	T. Croat 60868 (MO)	EU886501	AY248943	—
Colocasia esculenta (L.) Schott	J. Bogner 2958 (M)	JQ238890	JQ238972	JQ083569
Colocasia gigantea (Bl.) Hook.f.	J. Bogner 427 (M)	JQ238893	JQ238975	JQ083571
Peltandra virginica Raf.	J. Bogner 2119 (M)	EU886583	AY248942	JQ235756
Pinellia ternata (Thunb.) Breit.	J. McClements s.n., 30 Jul 2001	EU886503	AY248931	JQ083574
Pistia stratiotes L.	J. Bogner, Bot. Garden Munich	EU886585	AY248932	JQ083575
Protarum sechellarum Engl.	J. Bogner 2545 (M)	EU886588	AY248933	JQ083576
Remusatia vivipara (Lodd.) Schott	MO living acc. 69705b	EU886584	AY248934	—
Steudnera discolor Bull	J. Bogner 1582 (M)	EU886586	EF517221	JQ083580
Typhonodorum lindleyanum Schott	J. Bogner s.n. (M)	EU886578	EU886627	—
Xanthosoma sagittifolium (L.) Schott & Endl.	MO living acc. 850652b, Kemper Code C752	EU886500	AY248944	—

Notes: Information about the sequenced ingroup species is provided in Cusimano et al. (2010). Not all PhyC sequences were used.

two independent BEAST runs and then combined the log output files using LogCombiner (part of the BEAST package). We used Fig Tree version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) to chose the "maximum clade credibility tree," which is the tree in the posterior sample that has the maximum sum of posterior probabilities on its $n - 2$ internal nodes. This tree is not necessarily the majority rule consensus tree.

Diversification Analyses When Trees are Incomplete

Diversification analyses relied on an ultrametric tree obtained under the Bayesian relaxed clock model (above) and were repeated on 100 random trees from the MCMC chain. We (i) carried out the two most widely used methods for diversification analysis, the γ statistic and the BDL analysis, with the standard ways of handling missing species, the MCCR test and the ΔAICrc test statistic, both of which rely on tree simulation and pruning. We then (ii) used our newly developed method, CorSiM, which also uses the γ statistic and the BDL analysis but relies on simulating missing splits before any further analysis, that is, it augments the data under a model. Our method additionally allows the user to chose whether to use random or nonrandom species simulation (see Appendix). All analyses were carried out in R (R Developmental Core Team 2006), and CorSiM is available as an R package available on the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/TreeSim/TreeSim.pdf). We now briefly introduce the γ statistic and the BDL analysis and then explain analyses (i) and (ii).

γ statistic.—The γ statistic (Pybus and Harvey 2000) tests for departure from a constant-rate pure birth model. For completely sampled phylogenies, Pybus and Harvey (2000) found that $\gamma = -1.645$ (+1.645) represents the critical value of the constant-rates test. Values below this cutoff reject the pure birth model ($\gamma = 0$). We relied on the implementation of this statistic (gamStat) in Laser version 2.2 (Rabosky 2006a).

BDL analysis.—We also compared the fit of the likelihood models implemented in the fitdAICrc function in Laser, namely two constant rate models of diversification (a pure birth model and a birth–death model) and three variable rate models (logistic density dependence, exponential density dependence, and a two-rates variant of the pure birth model with a rate shift at a certain time point). Additionally, we fitted four constant rate models available in the R package TreePar (Stadler 2011a), the pure birth (Yule) model, the birth–death model, and the Yule and birth–death two-rates models with bd.shifts.optim. We allowed rate shifts over the whole range of branching times, with a grid dividing the range into 100 parts.

Correcting for missing species by tree simulation and pruning (the traditional approach).—Using the TreeSim package (Stadler 2011b), we simulated 1000 trees with the number of tips corresponding to the total number of species in the focal clades, here 62 and 58. Speciation and extinction rates used for tree simulation were obtained by fitting the constant rate birth–death model to the empirical data. Simulated trees were then randomly pruned to the sample sizes, here 32 and 32. This yields a null distribution of γ values against which the empirical γ value is compared using the MCCR test statistic (Pybus and Harvey 2000). For BDL analysis, the five diversification models in Laser were fit to the 1000 pruned trees, and the resulting ΔAIC values compared with the ΔAIC value of the empirical tree using the ΔAICrc test statistic (Rabosky 2006a). Additionally, we checked for type I errors in the inference of rate upswings using the criteria proposed by Rabosky (2006b). His fig. 3 shows the distribution of ΔAIC scores as a function of the number of taxa or number of model parameters in simulated phylogenies: As these numbers increase, a greater difference in AIC scores between the best rate constant and rate variable models is required to maintain $\alpha = 0.05$.

Correcting for missing species by simulating missing splits (the CorSiM approach).—For simulating the missing species, we used the sim.missing function (R package CorSiM; for details, see Appendix), which requires as input data the empirical branching times, a speciation and an extinction rate, the number of missing species and, optionally, a time interval during which the missing speciation events may have happened. Missing speciation events are simulated under the assumption that evolution followed a constant rate birth–death model.

For the nonrandomly sampled Mediterranean clade, we calculated the input rates with the TreePar function bd.groups.optim (Stadler and Bokma, in review), which estimates the maximum likelihood speciation and extinction rates (under a constant rate birth–death model) by taking into account information about sampling density (here 52%) and the time of the missing speciation events, here set to 16 to 0 myr, because the genera are older than 16 myr (as seen in the relaxed clock chronogram, Fig. S1 available at http://datadryad.org, doi: 10.5061/dryad.r8f04fk2). For the randomly sampled Typhonium clade, we calculated the input rates with the bd.shifts.optim function, which also estimates the maximum likelihood speciation and extinction rates (under a constant rate birth–death model), by taking into account sampling incompleteness (here 55%) but assumes random species sampling. Missing branching times were simulated 1000 times for each clade; the simulated times were then added to the empirical branching times yielding 1000 completed data sets.

For each focal clade (the Typhonium clade and the Mediterranean clade), we then applied the γ statistic and the BDL analysis (using Laser and TreePar) to the 1000 completed data sets, which yielded means and standard deviations (SDs) for the γ statistic, the AIC values, and the inferred rate parameters from the BDL

analyses. We also calculated the percentage of the 1000 completed data sets for which a particular model fit best.

RESULTS

Analyses Using the Traditional Approaches

Speciation and extinction rates used for tree simulations and results obtained with the traditional approaches versus the CorSiM approach are shown in Tables 2 and 3. The γ value for the Mediterranean clade is −0.3, not significantly different from zero, implying that the constant rate pure birth (Yule) model is not rejected. The best-fitting model as inferred from the BDL analysis also is the Yule model. For the Typhonium clade, the γ value is −2.96, which according to the MCCR test is significantly different from zero (P = 0.01), implying that diversification occurred mostly near the root and may be slowing down. The BDL analysis preferred the logistic density dependence model, hence also inferred a slowdown.

Analyses Using CorSiM

Speciation and extinction rates used for tree simulations with CorSiM were λ = 0.94 and μ = 0.03 for the Mediterranean clade and λ = 0.09 and μ = 0 for the Typhonium clade. The resulting completed data sets are visualized as lineage-through-time plots in Fig. 1 and results are shown in Tables 2–4. For the Mediterranean clade, the mean γ value is 4.26 ± 0.35, rejecting a constant rate diversification with high confidence (P=1). With BDL analysis, Laser prefers the Birth/Death model in 55.6% of the data sets and the Yule two-rates model in 44.4% of the data sets. Both models have nearly the same mean AIC values (−70.3 and −70.46). TreePar preferred the latter model in 82.2% of the data sets. The inferred rate change in both analyses is an increase at 1.95 Ma with an SD of 0.98 Ma (Table 2).

For the Typhonium clade, the mean γ value is −1.56 ± 0.65, and the γ statistic rejected a constant rate model in 50% of the cases; all inferred γ values are negative. With BDL analysis, Laser prefers the Yule two-rates model (based on the mean AIC) but only in 37% of the 1000 data sets. In 28% of the data sets, the logistic density dependence model was the best fit. TreePar preferred the Yule two-rates model in 96% of the data sets. The rate change is a decrease at 9.73–12.94 Ma (Table 2, with large SDs).

DISCUSSION

In the present study, we propose a new approach, CorSiM, for the problem of inferring diversification rates from incompletely sampled phylogenies. CorSiM provides an objective and robust way of estimating diversification from randomly or nonrandomly sampled phylogenies. Previous methods created null models and inferred diversification rates from incomplete data and could validly only be applied to randomly sampled

TABLE 2. Results from BDL analyses of two empirical (incomplete) data sets compared with the respective model-completed CorSiM data sets (mcc: maximum clade credibility tree), the latter yielding mean values with SDs from a 1000 completed data sets

			Yule	BD		DDL			DDX		Yule-2r			BD-2r				
			r	r	a	r	k	r	x	r1	r2	st	r1	a1	r2	a2	st
Mediterranean clade	Empirical phylogeny		0.09	0.09	0.00	0.12	94.07	0.12	0.09	0.10	0.06	2.74	0.08	0.52	0.00	1.00	2.74
	CorSiM data sets	Mean	0.17	0.04	0.90	0.17	1,051,301	0.02	−0.68	0.11	0.41	1.95	−0.06	0.74	−0.05	0.94	3.73
		SD	0.01	0.01	0.03	0.00	18,017	0.01	0.08	0.01	0.10	0.98	0.55	0.26	1.04	1.44	5.98
Typhonium clade	Empirical phylogeny		0.06	0.06	0.00	0.17	35.05	0.41	0.67	0.11	0.03	11.03	0.11	0.00	0.03	0.00	10.16
	CorSiM data sets from mcc tree	Mean	0.09	0.09	0	0.13	1252	0.24	0.30	0.13	0.07	9.73	0.04	0.38	0.01	0.82	10.01
		SD	0.00	0.00	0.01	0.01	35,676	0.05	0.07	0.06	1.78	5.66	0.64	1.34	0.45	2.26	5.88
	CorSiM data set from MCMC chain sample	Mean	0.09	0.09	0.01	0.13	14,520	0.28	0.32	0.17	0.08	12.94	−0.1	0.66	−0.09	0.86	13.9
		SD	0.01	0.01	0.04	0.02	121,219.93	0.12	0.13	0.09	0.09	6.79	1.3	0.61	2.06	2.32	7.43

Notes: Models fitted to the data are four constant rate models: the pure birth model and the birth–death model once with stable rate(s) (Yule, BD) and once with changing rate(s) at a breakpoint time st (Yule-2r, BD-2r); and two variable rate models: the logistic density-dependent model (DDL), and the exponential density-dependent model (DDX); r = diversification rate (speciation–extinction in species/myr), a = extinction fraction (extinction/speciation), st = time of rate shift (in myr) of the two-rate models, r1/a1 and r2/a2 = parameters before and after st; k = carrying capacity parameter of the DDL model, x = rate change parameter of the DDX model, BD = Birth/Death. Bold parameters indicate the best fitting model(s) (see Tables 3 and 4).

TABLE 3. Results of diversification rate analyses using the γ statistic and the MCCR test

	γ statistic			BDL analysis using Laser (ΔAIC test statistic)				BDL analysis using TreePar (AIC)			
	γ	Critical value	P	Best-fitting rate constant	Best-fitting rate variable	Observed ΔAIC	P	Yule	BD	Yule-2r	BD-2r
Mediterranean clade	−0.30	−2.44	0.73	Yule	DDL	−1.62	0.86	204.24	205.60	204.38	207.18
Typhonium clade	−2.96	−2.37	0.01	Yule	DDL	9.14	0.03	227.60	229.60	219.01	222.84

Notes: For the BDL analysis with Laser, the best fitting constant-rate model and the best fitting variable-rate model and the differences in their AIC values (ΔAIC) are shown, as well as the probability P with which the constant rate model is rejected by the ΔAIC test statistic. For the BDL analysis with TreePar, AIC values are shown for the four models fitted to the data. Boldface indicates the preferred model. For model descriptions, see Table 2.

data. To improve diversification estimation from nonrandomly sampled phylogenies, Brock et al. (2011) proposed an approach that creates null distributions for nonrandomly sampled data, but the approach requires the specification of a scaling parameter alpha for which there is unclear justification. By contrast, CorSiM is applicable to both random and nonrandomly sampled phylogenies and infers all rates with confidence intervals. In the current implementation of CorSiM, the nonrandomly sampled phylogenies are completed by simulation of the missing branching times within a user-specified time interval under a constant rate birth–death model. If the user requires another sampling scenario, CorSiM can be extended to simulate under that other scenario.

We apply the new approach to two similarly incomplete phylogenies that differ in the randomness of their species sampling (one has the deeper nodes oversampled, the other is randomly sampled). Results from the new approach were compared with those obtained with the two most widely used traditional approaches. Table 5 summarizes results from the different methods. For the randomly sampled clade, the traditional methods and our CorSiM approach as expected led to the same results. In the specific case of our focal clade (Typhonium), this was a decrease in diversification rates. For the nonrandomly sampled clade, traditional methods and the CorSiM approach yield different results. In our case (i.e., for the Mediterranean clade), both traditional methods prefer the Yule model, whereas CorSiM reveals an increase in diversification. Since our method has a low type I error (below), we have confidence in this results.

Strengths and Weaknesses of Different Methods for Handling Missing Species

The new approach presented here is based on the assumption that analyzing complete data sets is the best way for inferring diversification rate changes and

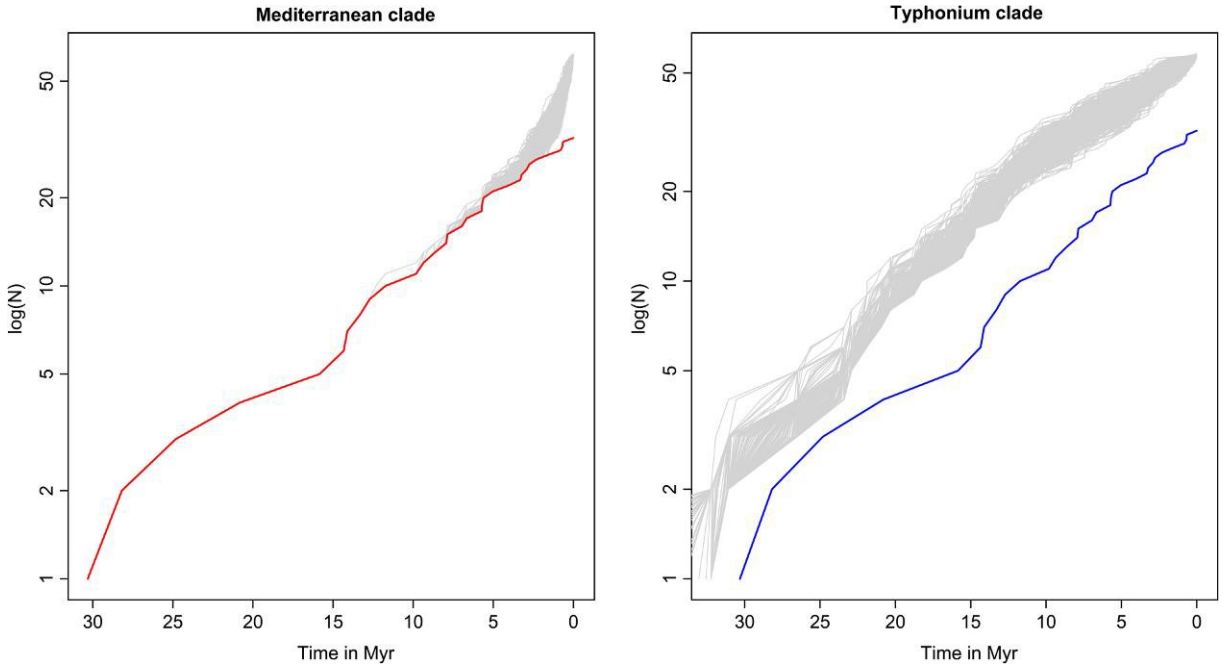

FIGURE 1. Lineage-through-time (LTT) plots obtained with the empirical incomplete data for the Mediterranean clade (left panel, dark line) and the Typhonium clade (right panel, dark line) along with the LTT plots obtained from a 1000 completed data sets using the CorSiM approach (pale lines).

TABLE 4. Results of the CorSiM analyses of diversification rate changes in two empirical (incomplete) phylogenies, showing the γ statistic, the mean γ value with SD, and the mean probability P (with SD) with which the null hypothesis (constant rate diversification) is rejected

		γ statistic		BDL analyses (AIC values)								
				Laser				TreePar				
		γ	p	Yule	BD	DDX	DDL	Yule-2r	Yule	BD	YULE-2r	BD-2r
Mediterranean clade	Mean	4.26	1	−52.5	−70.33	−65.53	−50.53	−70.46	333	315.2	312.4	314.2
	SD	0.4	0	1.9	5.5	4.6	1.9	7.2	2.7	6.1	7.5	7.5
	% best fitting	—	—	0	55.6	0	0	44.4	0	7.5	82.2	10.3
Typhonium clade CorSiM data sets from mcc tree	Mean	−1.56	0.1	33.44	35.44	32.12	31.86	31.68	386.2	388.2	382.9	386.1
	SD	0.7	0.1	5.2	5.2	4.0	3.7	3.8	5.2	5.2	3.9	4.2
	% best fitting	—	—	16.7	0	18.4	27.9	37	2.8	0	96	1.2
Typhonium clade CorSiM data set from MCMC chain sample	Mean	−1.46	0.13	30.028	32.02	28.105	28.64	26.782	390.9	392.9	386.1	388.21
	SD	0.8	0.1	14.73	14.73	14.13	14.3	14.21	14.7	14.7	14.1	14.6
	% best fitting			16	0	15	9	59	1.1	0	86.2	12.7

Notes: For BDL analyses, the table lists the mean AICs (with SD) inferred with Laser or TreePar, and the percentage of the 1000 completed data sets for which a particular model fit best. Boldface indicates the preferred model; for the Mediterranean clade, the AIC values of the two best fitting models are statistically undistinguishable. For model descriptions, see Table 2.

on the wide agreement among statisticians that model-based data augmentation and multiple imputation is the best way of dealing with missing data (Nakagawa and Freckleton 2008). If instead of augmenting the data based on a model, one adds missing species "by hand" (e.g., Purvis et al. 1995; Barraclough and Vogler 2002), this has three undesirable effects: It is subjective; one risks adding bias to the data if species sampling is extremely low and many nonsequenced species have to be added; the approach only works with sufficient knowledge of species relationships. An alternative approach, SA (Paradis 1997), requires minimum ages for adding missing splits, which often will be unavailable. The new method of Brock et al. (2011) suffers from the need to subjectively specify a scaling parameter.

The CorSiM approach overcomes these problems. Missing data are added beforehand to create completed data sets, and this is done under a constant rate birth–death model and repeated 1000 times, yielding objective model-based data augmentation. The completed batches of data sets (consisting of the empirical splits, plus the simulated ones) can then be analyzed with any of the available methods for diversification estimation to obtain mean values and SDs. The approach importantly also allows specifying whether species sampling likely is random or nonrandom. In this paper, we assumed that the species sampling procedure was the same across all subtrees of the empirical phylogenies (either random sampling or oversampling of deep [old] nodes). However, if in a large phylogeny one had reason to think that species sampling was random in some subclades, but nonrandom in others, simulating the missing splits could be done separately for subtrees, using the appropriate assumptions. The subsequent γ statistic and BDL analysis would then be done based on the completed phylogeny, that is, the completed subtrees would be combined.

A caveat about the CorSiM approach is that the branching times in the completed data sets will be biased towards constant rate diversification since they were simulated under this process. Thus, analyses of the completed data sets testing for the constant rate model must have low type I error, but a high type II error.

CONCLUSIONS

The growing field of evolutionary diversification studies requires objective methods that can be applied to empirical data sets with different properties, including random or nonrandom species sampling. CorSiM is the first method achieving this. It has a low type I error rate, but failing to reject the null hypothesis (type II error) remains a problem of the method. We suspect that reducing type II error rates would require having full likelihood approaches for inferring diversification rates from incomplete phylogenies. A recent study (Höhna et al. 2011) provides such an approach for nonrandom sampling, however, only under a constant rate birth–death model of diversification. In the current study, we were instead concerned with testing if a constant rate model is appropriate or if more complex models are

TABLE 5. Summary of the contrasting inferences about diversification rates in two clades of Araceae obtained with different approaches for handling missing species

	Mediterranean clade		Typhonium clade	
	Tree simulation	CorSiM-corrected data	Tree simulation	CorSiM-corrected data
γ statistic	Constant diversification	Increasing diversification	Decreasing diversification	Constant diversification (decreasing diversification)
Model fitting	Constant diversification	Yule two-rates model with rate increase at 1–2 myr (BD model)	Decreasing diversification (DDL)	Yule two-rates, with rate decrease at 10 myr

Note: Shown in brackets, nearly equally likely models.

required instead (such as logistic density dependence, exponential density dependence, and a two-rates variant of the pure birth model with a rate shift at a certain time point). More work is needed to develop a full likelihood framework for general models of diversification, without requiring to simulate the nonsampled branching times under the conservative constant rate birth–death model as done in CorSiM.

The comparison of our new approach with traditional methods as expected yielded consistent results in a randomly sampled clade. In a nonrandomly sampled clade, however, results differed strongly, and CorSiM rejected the constant rate birth–death model, which the traditional methods supported. Since incomplete nonrandomly sampled phylogenies are pervasive, diversification rate estimation from such phylogenies should switch to the robust method proposed here.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository (DOI:10.5061/dryad.r8f04fk2).

FUNDING

Supported by the German Research Council (grant RE 603/7-1).

ACKNOWLEDGMENTS

We thank R. E. Ricklefs for helpful discussions of an early version of the CorSiM approach.

REFERENCES

Baldwin B.G., Sanderson M.J. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc. Natl. Acad. Sci. U.S.A. 95:9402–9406.

Barraclough T.G., Vogler A.P. 2002. Recent diversification rates in North American tiger beetles estimated from a dated mtDNA phylogenetic tree. Mol. Biol. Evol. 19:1706–1716.

Brock C.D., Harmon L.J., Alfaro M.E. 2011. Testing for temporal variation in diversification rates when sampling is incomplete and nonrandom. Syst. Biol. 60:410–419.

Cusimano N., Barrett M., Hetterscheid W.L.A., Renner S.S. 2010. A phylogeny of the Areae (Araceae) implies that Typhonium, Sauromatum and Lazarum are distinct clades. Taxon. 59:439–447.

Cusimano N., Bogner J., Mayo S.J., Boyce P.C., Wong S.Y., Hesse M., Hetterscheidt W.L.A., Keating R.C., French J.C. 2011. Relationships within the Araceae: comparisons of morphological patterns with molecular phylogenies. Am. J. Bot. 98:654–668.

Cusimano N., Renner S.S. 2010. Slowdowns from real phylogenies may not be real. Syst. Biol. 59:458–464.

Day J.J., Cotton J.A., Barraclough T.G. 2008. Tempo and mode of diversification of Lake Tanganyika Cichlid Fishes. PLoS One. 3:e1730.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Rambaut A. 2007. Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Hartmann K., Wong D., Stadler T. 2010. Sampling trees from evolutionary models. Syst. Biol. 59:465–476.

Harvey P.H., May R.M., Nee S. 1994. Phylogenies without fossils Evolution. 48:523–529.

Hey J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution. 46:627–640.

Hö hna S., Stadler T., Ronquist F., Britton T. 2011. Inferring speciation and extinction rates under different sampling schemes. Mol. Biol. Evol. 28 :2577–2589.

Magalló n S., Sanderson M.J. 2001. Absolute diversification rates in angiosperm clades. Evolution. 55:1762–1780.

McPeek M.A. 2008. The ecological dynamics of clade diversification and community assembly. Am. Nat. 172:E270–E284.

Nakagawa S., Freckleton R.P. 2008. Missing inaction: the danger of ignoring missing data. Trends Ecol. Evol. 23:592–596.

Nee S. 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Syst. 37:1–17.

Nee S., Mooers A.O., Harvey P.H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. U.S.A. 89:8322–8326.

Nylander J.A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 24:581–583.

Paradis E. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. Proc. R. Soc. Lond. B Biol. Sci. 264:1141–1147.

Paradis E. 1998. Detecting shifts in diversification rates without fossils. Am. Nat. 152:176–187.

Phillimore A.B., Price T.D. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6:483–489.

Purvis A., Nee S., Harvey P.H. 1995. Macroevolutionary inferences from primate phylogeny. Proc. R. Soc. Lond. B Biol. Sci. 260: 329–333.

Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B Biol. Sci. 267:2267–2272.

Rabosky D. L. 2006a. LASER: a maximum likelihood toolkit for inferring temporal shifts in diversification rates. Evol. Bioinform. Online. 2:257–260.

Rabosky D.L. 2006b. Likelihood methods for detecting temporal shifts in diversification rates. Evolution. 60:1152–1164.

Rabosky D.L., Donnellan S.C. Talaba A.L., Lovette I.J. 2007. Exceptional among-lineage variation in diversification rates during the

radiation of Australia's most diverse vertebrate clade. Proc. R. Soc. Lond. B Biol. Sci. 274:2915–2923.
R Developmental Core Team. 2006. A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Renner S.S., Zhang L.-B. 2004. Biogeography of the Pistia clade (Araceae): based on chloroplast and mitochondrial DNA sequences and Bayesian divergence time inference. Syst. Biol. 53:422–432.
Sanderson M.J., Bharathan G. 1993. Does cladistic information affect inferences about branching rates? Syst. Biol. 42:1–17.
Sanderson M.J., Donoghue M.J. 1994. Shifts in diversification rate with the origin of angiosperms. Science. 264:1590–1593.
Stadler T. 2008. Lineage-through-time plots of neutral models for speciation. Math. Biosci. 216:163–171.
Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.
Stadler T. 2011a. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. U.S.A. 108:6187–6192.
Stadler T. 2011b. Simulating trees on a fixed number of extant species. Syst. Biol. 60:676–684.
Stadler T., Bokma F. In review. Estimating speciation and extinction rates for phylogenies of higher taxa.

APPENDIX

Mathematical Derivation of the CorSiM Approach

The CorSiM approach takes as input parameters x, λ, μ, m, and optional the two parameters t_{min} and t_{max}, and then simulates m missing speciation times based on the vector x of empirically known speciation times from an incomplete phylogeny. The underlying model for simulations is the constant rate birth–death model under which a speciation event happens with rate λ and an extinction event happens with rate μ. The missing speciation events are added within the time interval t_{min} and t_{max}. If t_{min} is not specified, t_{min} is set to the present. If t_{max} is not specified, t_{max} is sampled as explained in Step (1) below.

We define t0 to be the present and increasing going into the past. We further define n to be the number of species in the empirical tree.

Our simulations employ the inverse transform method. This method samples a continuous random variable X with cumulative distribution function (c.d.f.) F(x) by sampling a value r uniformly at random from [0, 1] and plugging it into the inverse of the c.d.f. $F^{-1}(x)$ to obtain a sampled value. It holds $X = F^{-1}(x)$, that is, the sampled values $F^{-1}(x)$ are drawn from X, because F(x) is distributed uniformly at random on [0, 1].

The Algorithm CorSiM Performs the Following Steps

(1) Determining t_{max}: If t_{max} is not specified, t_{max} is sampled. The inverse of the c.d.f. of the time of the first individual of a birth–death tree, which has N individuals is given in Hartmann et al. (2010),

$$\frac{1}{\lambda-\mu}\ln\frac{-1\lambda\ \mu r^{1/N}}{1-r^{1/N}}\quad .$$

Our tree has n species, and the sampling fraction is $\rho = n/(m+n)$. An incompletely sampled tree can be interpreted as a completely sampled tree by using $\frac{\mu-\lambda(1-\rho)}{\lambda\rho}$ instead of $\frac{\mu}{\lambda}$ (Stadler 2009). Thus, we use,

$$Q^{-1}(r|\lambda,\ \mu, n) = \frac{1}{\lambda-\mu}\ln\left(\frac{\lambda\rho + (\lambda(1-\rho)-\mu)r^{1/n}}{\lambda\rho(1-r^{1/n})}\right) \tag{A.1}$$

to draw a sample for t_{max}, where r is drawn from the uniform distribution on [0, 1].

(2) Adding the missing speciation times sequentially: We define the vector $z = t_{max} = y_0 > y_1 > \cdots > y_{l-1} > y_l = t_{min}$, where $y_1, \ldots, y_{k-1}$ are the speciation times from vector x, which fall in the interval $[t_{max}, t_{max}]$. We now successively add to the vector z the m missing speciation times.

(2.1) Let the number of species descending the speciation event y_i be k_i. The probability that a deleted speciation event occurred in the time interval during which the tree (without this speciation event) has k species is $p_k \propto k$ (Stadler 2008). Based on the probabilities p_k, we sample the interval into which we insert the missing speciation event. Let the sampled interval be $[y_i, y_{i+1}]$.

(2.2) We will now sample the exact time of the missing speciation event. The c.d.f. for the time of a missing speciation event in a tree of age t is (Gernhard 2008),

$$H(r|\lambda,\ \mu, t) = \frac{1-e^{-(\lambda-\mu)r}}{\lambda-\mu e^{-(\lambda-\mu)r}}\frac{\lambda-\mu e^{-(\lambda-\mu)t}}{1-e^{-(\lambda-\mu)t}},$$

and thus, conditioned on the speciation event being between y_i and y_{i+1}, the c.d.f. is,

$$F(r|\lambda,\ \mu, y_i, y_{i+1}) = \frac{\frac{1-e^{-(\lambda-\mu)r}}{\lambda-\mu e^{-(\lambda-\mu)r}} - \frac{1-e^{-(\lambda-\mu)y_{i+1}}}{\lambda-\mu e^{-(\lambda-\mu)y_{i+1}}}}{\frac{1-e^{-(\lambda-\mu)y_i}}{\lambda-\mu e^{-(\lambda-\mu)y_i}} - \frac{1-e^{-(\lambda-\mu)y_{i+1}}}{\lambda-\mu e^{-(\lambda-\mu)y_{i+1}}}}$$

The time of the simulated speciation event is again obtained from the inverse function,

$$F^{-1}(r|\lambda,\ \mu, y_i, y_{i+1}) = \frac{1}{\mu-\lambda}\ln\frac{1-(r+c_1)c_2\lambda}{1-(r+c_1)c_2\mu},$$

$$c_1 = \frac{\frac{1-e^{-(\lambda-\mu)y_{i+1}}}{\lambda-\mu e^{-(\lambda-\mu)y_{i+1}}}}{c_2},$$

$$c_2 = \frac{1-e^{-(\lambda-\mu)y_i}}{\lambda-\mu e^{-(\lambda-\mu)y_i}} - \frac{1-e^{-(\lambda-\mu)y_{i+1}}}{\lambda-\mu e^{-(\lambda-\mu)y_{i+1}}}, \tag{A.2}$$

where r is again drawn from the uniform distribution on [0, 1].

The sampled speciation time is added to the vector z. Once m speciation times were added to z, the algorithm terminates, otherwise the algorithm continues with Step 2.1).