Syst. Biol. 71(6):1440–1452, 2022

https://doi.org/10.1093/sysbio/syac037
Advance Access publication May 24, 2022

The Occurrence Birth–Death Process for Combined-Evidence Analysis in Macroevolution and Epidemiology

JÉRÉMY ANDRÉOLETTI[1,*], ANTOINE ZWAANS[1], RACHEL C. M. WARNOCK[2], GABRIEL AGUIRRE-FERNÁNDEZ[3], JOËLLE BARIDO-SOTTANI[4], ANKIT GUPTA[1], TANJA STADLER[1], AND MARC MANCEAU[1]

[1]*Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland;* [2]*GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany;* [3]*Paleontological Institute and Museum, University of Zürich, Zürich, Switzerland and* [4]*Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA, USA*

**Correspondence to be sent to: Computational Evolution group, Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland;*
E-mail: jeremy.andreoletti@ens.psl.eu.
Jérémy Andréoletti and Antoine Zwaans contributed equally to this article.

Received 26 November 2021; reviews returned 2 May 2022; accepted 6 May 2022
Associate Editor: Sebastian Höhna

Abstract.—Phylodynamic models generally aim at jointly inferring phylogenetic relationships, model parameters, and more recently, the number of lineages through time, based on molecular sequence data. In the fields of epidemiology and macroevolution, these models can be used to estimate, respectively, the past number of infected individuals (prevalence) or the past number of species (paleodiversity) through time. Recent years have seen the development of "total-evidence" analyses, which combine molecular and morphological data from extant and past sampled individuals in a unified Bayesian inference framework. Even sampled individuals characterized only by their sampling time, that is, lacking morphological and molecular data, which we call *occurrences*, provide invaluable information to estimate the past number of lineages. Here, we present new methodological developments around the fossilized birth–death process enabling us to (i) incorporate occurrence data in the likelihood function; (ii) consider piecewise-constant birth, death, and sampling rates; and (iii) estimate the past number of lineages, with or without knowledge of the underlying tree. We implement our method in the RevBayes software environment, enabling its use along with a large set of models of molecular and morphological evolution, and validate the inference workflow using simulations under a wide range of conditions. We finally illustrate our new implementation using two empirical data sets stemming from the fields of epidemiology and macroevolution. In epidemiology, we infer the prevalence of the coronavirus disease 2019 outbreak on the Diamond Princess ship, by taking into account jointly the case count record (occurrences) along with viral sequences for a fraction of infected individuals. In macroevolution, we infer the diversity trajectory of cetaceans using molecular and morphological data from extant taxa, morphological data from fossils, as well as numerous fossil occurrences. The joint modeling of occurrences and trees holds the promise to further bridge the gap between traditional epidemiology and pathogen genomics, as well as paleontology and molecular phylogenetics. [Birth–death model; epidemiology; fossils; macroevolution; occurrences; phylogenetics; skyline.]

INTRODUCTION

Birth–death processes are stochastic processes used to model population dynamics with two main parameters, the birth rate and the death rate, which are respectively the rate at which new lineages appear, and the rate at which lineages are removed from the process. In macroevolution, these two rates correspond to the speciation and extinction rates, while in epidemiology they correspond to the transmission and recovery rates. These processes already enjoy a long history of applications in evolutionary biology. In the first half of the 20th century, Yule (1925) introduces them in the field with macroevolutionary applications in mind, to model the number of species within genera. Kendall (1948) then derives analytically the transition probabilities for linear birth–death processes, and discusses their use in the context of evolutionary biology, with a special focus on epidemiology. Ground-breaking work by Nee et al. (1994) followed on the probability density of the *reconstructed tree* in a linear birth–death process, that is, the tree obtained by pruning all extinct lineages from the full genealogical history of the process (see Fig. 1b). The linear birth–death process was then later extended to allow rates to vary in different parts of the tree (Alfaro et al. 2009), over time (Morlon et al. 2011), or depending on some character of interest (Maddison et al. 2007).

Although diversification histories inferred from extant species sometimes agree with those inferred from the fossil record (Morlon et al. 2011; Xing et al. 2014; Silvestro et al. 2018), there remains a gap between these two approaches in macroevolution (Marshall 2017). On the one hand, extant species provide invaluable information regarding the dynamics of the diversification process, especially close to the present. On the other hand, the fossil record, albeit incomplete, could much better inform extinction estimates (Quental and Marshall 2010). An extension introduced by Stadler (2010) and dubbed the *Fossilized Birth–Death Process* (FBDP) (Heath et al. 2014) aimed at jointly modeling extant and extinct taxa along the same tree, and thus helped bridge the gap between paleontology and molecular phylogenetics. In this model, each species can be sampled and included in the reconstructed tree throughout its lifetime at a fixed rate (see Fig. 1c). The probability density of the resulting phylogeny is derived in closed-form and has been successfully used as a prior in Bayesian phylogenetic analyses to study the diversification history of hymenopterans (Zhang et al. 2015), as well as the penguins (Gavryushkina et al. 2016). The same model

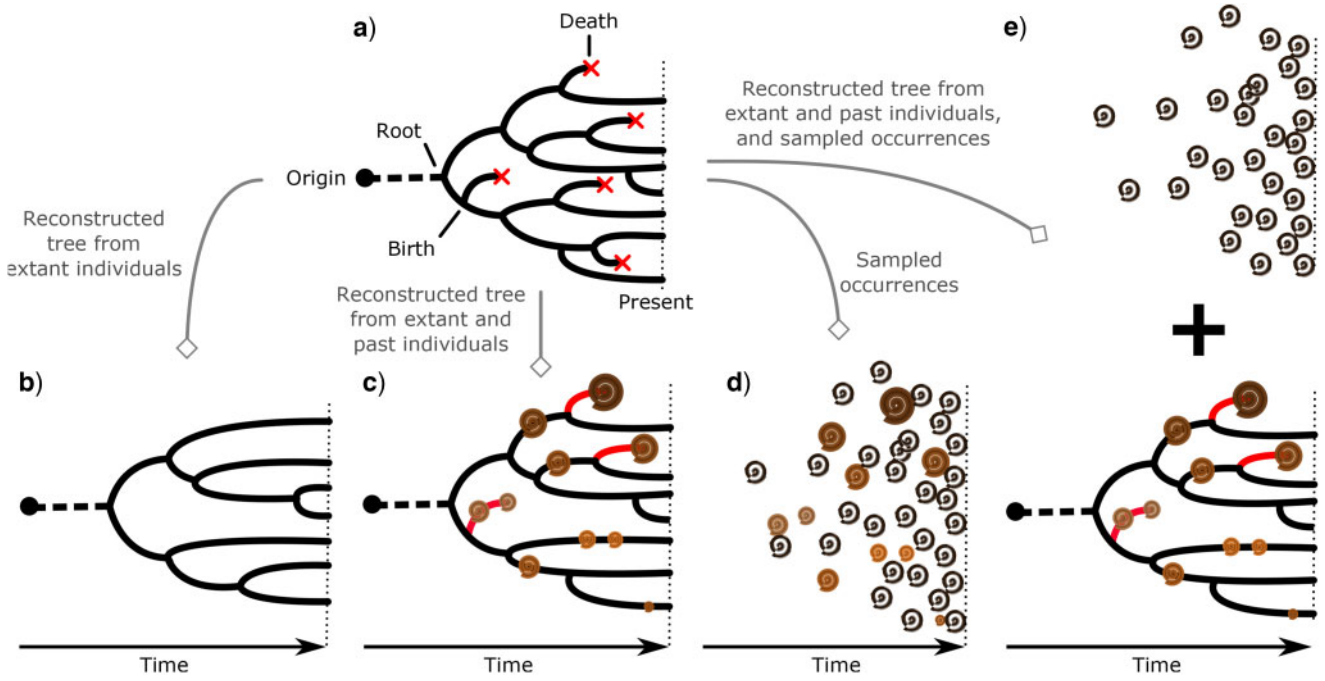

FIGURE 1. Different approaches to infer past history and number of lineages. a) The full unknown history of the population. Different types of data can be used in order to infer the past history of the population. b) Genetic sequencing data and character data for present-day individuals only. c) Enriched tree with past samples with documented character data (represented as diverse ammonite morphs). d) Occurrences alone (represented as all ammonite fossils indistinctly). Finally, e) a more comprehensive total-evidence method integrates extant genetic sequences, samples with character data, and occurrences in a unified framework.

has also been used in the context of epidemiology, where infected individuals can as well be sampled throughout the infectious period and be included in the reconstructed tree (Stadler et al. 2013). Finally, model extensions have been introduced to explicitly take into account stratigraphic ranges (i.e., the interval between the first and the last fossil appearance of a species) or, in the context of epidemiology, patient-specific durations of infection (Stadler et al. 2018). Although we focus here on the birth–death modeling framework, we note that an alternative line of the phylogenetic literature has evolved in parallel around the coalescent framework, following pioneering work in population genetics by Kingman (1982). Models and methods have been developed in coalescent models as well to incorporate sampling through time (Drummond et al. 2002; Parag et al. 2020).

An important feature of many standard paleontological data sets is that only a fraction of fossils have been thoroughly described and are associated with morphological data. Similarly, in standard epidemiological surveys, only a fraction of the recorded case count data is typically sequenced. In this article, we call *samples with character data* the subset of samples with either morphological data or molecular data, and *occurrences* the recorded samples without character data, which contain valuable information regarding the underlying number of lineages (see Fig. 1d). For this reason, they have long been used in paleontology to infer diversity trajectories (Raup 1972; Sepkoski et al. 1981), and even preservation, origination, and extinction rates in an alternative Bayesian setting (Silvestro et al. 2014; Silvestro et al. 2019). Some authors have analyzed occurrences in the standard FBD-based Bayesian framework, considering them as leaves in the tree with missing character data, and integrating over the unknown topology, in a paleontogogical (Heath et al. 2014; Gavryushkina et al. 2014; O'Reilly and Donoghue 2020) or epidemiological (Featherstone et al. 2021) context. Applying the standard FBD model in this case implicitly means that both samples with character data and occurrence data are assumed to have been generated under the same process, with the same rates. A second step towards integrating these occurrences was performed by Vaughan et al. (2019), who explicitly modeled an additional sampling process for occurrences, allowing for the joint analysis of the observation of a phylogeny and a record of occurrences (see Fig. 1e). Vaughan et al. (2019) additionally proposed an inference framework based on the use of a particle filter to compute the likelihood. Rasmussen et al. (2011) present another method based on a particle filtering algorithm to consider occurrences and trees in tandem, although in a coalescent framework instead of a birth-death framework. Gupta et al. (2020) then built on previous work by Vaughan et al. (2019) and described soon after a fast algorithm to compute the likelihood of the data, focusing on a special case of the model where all lineages sampled through time are removed from the process upon sampling. Finally, Manceau et al. (2021) presented a method to compute the distribution of number of lineages conditioned on a reconstructed tree and a record of occurrences.

In this article, we extend these last two methods to include piecewise-constant parameters, allowing us to explicitly incorporate known variation in birth, death, and sampling rates through time. We implement our work as a new distribution, coined the occurrence birth–death process (OBDP), available in the Bayesian phylogenetic software RevBayes (Höhna et al. 2016) to compute the joint probability density of a tree and a record of occurrences. This can readily be used to sample the posterior of trees and the number of lineages through time, given an observed record of occurrences and a list of samples with character data attached. We illustrate the versatility of the method on two empirical data sets coming from the fields of epidemiology and macroevolution. In epidemiology, we infer the prevalence through time for the coronavirus disease 2019 (COVID-19) outbreak on the Diamond Princess cruise ship, based on the joint observation of molecular sequences and case count data. In macroevolution, we infer the diversity through time in the cetacean clade, based on the joint observation of molecular data for extant species, morphological character data for some fossils and some extant species, and the record of fossil occurrences available on the Paleobiology Database.

Materials and Methods

Phylodynamic Model

We consider that a population of individuals starts at the time of origin t_{or} with one lineage, and evolves through time under a birth–death process with piecewise constant birth rate, λ_t, and death rate, μ_t. Three different sampling schemes are simultaneously applied along the process. First, individuals can be sampled through time and be included in the tree, with piecewise-constant sampling rate ψ_t. Second, they can be sampled through time as raw *occurrences* not included in the tree, with piecewise-constant sampling rate ω_t. Third, lineages reaching the present time are included in the tree with a fixed probability ρ. Finally, upon sampling, lineages are removed with a piecewise-constant probability of removal r_t. Note that time t is thus assumed to run backwards from t_{or} to 0.

As a result of these three sampling steps, we observe a reconstructed tree $\mathcal{T}$, which is the tree spanning all ψ-sampled and ρ-sampled individuals, as well as a record of occurrences $\mathcal{O}$, which is a timeline recording successive ω-sampling events. We aim at (i) computing the probability density of $(\mathcal{T},\mathcal{O})$, which will play the role of the phylodynamic likelihood in our Bayesian framework and (ii) compute the probability distribution of the total number of lineages in the process at time t, I_t, conditioned on the observed $(\mathcal{T},\mathcal{O})$. Note that the number of lineages in $\mathcal{T}$ at time t, denoted k_t, is an obvious lower bound of the total number of lineages in the process at time t, I_t. For this reason, we are targeting the probability distribution $\mathbb{P}(I_t = k_t + i)$, where i stands for the number of hidden lineages.

In our Supplementary Appendix available on Dryad at https://doi.org/10.5061/dryad.p8cz8w9rq, we extend the method introduced by Manceau et al. (2021) to include piecewise-constant parameters in computing two quantities. First, defining $(\mathcal{T}_t^{\uparrow},\mathcal{O}_t^{\uparrow})$ as the tree and record of occurrences constrained to $[t,t_{or}]$, we aim at numerically computing the joint probability of the partial tree and occurrence record between time t and the origin, and the total number of lineages at time t,

$$\forall i \in \mathbb{N},\ \ M_t^{(i)} := \mathbb{P}\left(\mathcal{T}_t^{\uparrow},\mathcal{O}_t^{\uparrow}, I_t = k_t + i\right) \tag{1}$$

which can be used to compute, upon reaching present day $t=0$, $\mathbb{P}(\mathcal{T},\mathcal{O}) = \sum_i M_0^{(i)}$.

Second, defining $(\mathcal{T}_t^{\downarrow},\mathcal{O}_t^{\downarrow})$ as the tree and record of occurrences constrained to $[0,t]$, we aim at numerically computing the probability of the partial tree and occurrence record between time t and the present, conditioned on the total number of lineages at time t,

$$\forall i \in \mathbb{N},\ \ L_t^{(i)} := \mathbb{P}\left(\mathcal{T}_t^{\downarrow},\mathcal{O}_t^{\downarrow} \mid I_t = k_t + i\right) \tag{2}$$

which can as well be used to compute, upon reaching the time of origin t_{or}, $\mathbb{P}(\mathcal{T},\mathcal{O}) = L_{t_{or}}^{(0)}$.

We derive initializing conditions and Master equations governing the evolution of M_t and L_t through time and compute these quantities by numerically evaluating the system ordinary differential equations (Supplementary Appendix available on Dryad). Note that in this numerical evaluation we have to make one approximation, namely to assume a maximal number of lineages N (while it could in theory become arbitrarily large). In practice, N must be chosen large enough to cover most of the high-density support of the L_t and M_t probability distributions to avoid biasing calculations.

Finally, provided we know both quantities at time t, the probability distribution K_t of the number of hidden lineages living at time t is given by,

$$\begin{aligned} K_t^{(i)} &:= \mathbb{P}(I_t = k_t + i \mid \mathcal{T},\mathcal{O}) \\ &\propto \mathbb{P}(I_t = k_t + i, \mathcal{T}_t^{\uparrow},\mathcal{O}_t^{\uparrow},\mathcal{T}_t^{\downarrow},\mathcal{O}_t^{\downarrow}) \\ &\propto \mathbb{P}(\mathcal{T}_t^{\downarrow},\mathcal{O}_t^{\downarrow} \mid I_t = k_t + i, \mathcal{T}_t^{\uparrow},\mathcal{O}_t^{\uparrow})\mathbb{P}(I_t = k_t + i, \mathcal{T}_t^{\uparrow},\mathcal{O}_t^{\uparrow}) \\ &\propto L_t^{(i)} M_t^{(i)}, \end{aligned} \tag{3}$$

where, in the last step, the first probability simplifies to $L_t^{(i)}$ thanks to the Markov property of the process. We summarize all the notation introduced above in Table 1.

Bayesian inference framework

We consider a Bayesian inference framework with additional model layers for character data evolution along the reconstructed tree $\mathcal{T}$. For epidemiological applications, we superimpose a model of molecular evolution, leading to the observation of a sequence alignment for both extant and extinct taxa in $\mathcal{T}$. For applications in macroevolution, we superimpose (i) a model of morphological evolution, leading to the

TABLE 1. Parameters and objects of the occurrence birth–death process

Parameter	Signification	Object	Signification
t_{or}	Time of origin	$\mathcal{T}$	Reconstructed tree
λ	Speciation rate	$\mathcal{O}$	Record of occurrence times
μ	Extinction rate	I_t	Total number of lineages
ψ	Fossil sampling rate	k_t	Number of sampled lineages
ω	Occurrence sampling rate	i	Number of hidden lineages
r	Removal probability at sampling	$(\mathcal{O}_t^{\uparrow}, \mathcal{T}_t^{\uparrow})$	Occurrences and tree before time t
ρ	Sampling probability at present	$(\mathcal{O}_t^{\downarrow}, \mathcal{T}_t^{\downarrow})$	Occurrences and subtrees after time t

observation of character data for both extant and extinct taxa in $\mathcal{T}$ and (ii) a model of molecular evolution, leading to the observation of a sequence alignment for extant taxa only. We summarize all model parameters as θ, and all (molecular and morphological) character data as $\mathcal{A}$, noting that $\mathcal{A}$ is independent of $\mathcal{O}$ given the tree and parameters. Thus, the target posterior distribution of reconstructed trees $\mathcal{T}$ and model parameters θ can be written as the product of the phylodynamic likelihood, the likelihood of character data given $\mathcal{T}$ and θ, and prior probabilities:

$$\mathbb{P}(\mathcal{T}, \theta | \mathcal{O}, \mathcal{A}) \propto \mathbb{P}(\mathcal{T}, \mathcal{O} | \theta)\, \mathbb{P}(\mathcal{A} | \mathcal{T}, \theta)\, \mathbb{P}(\theta). \quad (4)$$

First, we sample this posterior distribution using a Metropolis–Hastings Markov Chain Monte Carlo (MCMC). Second, the posterior probability distribution of the ancestral number of lineages can be written as,

$$K_t = \mathbb{P}(I_t | \mathcal{A}, \mathcal{O}) = \int_{\mathcal{T}, \theta} \mathbb{P}(I_t | \mathcal{T}, \mathcal{O}, \theta)\, d\mathbb{P}(\mathcal{T}, \theta | \mathcal{O}, \mathcal{A}) \quad (5)$$

and is thus numerically computed as the mean of K_t over the trace of the posterior of $(\mathcal{T}, \theta)$.

Numerical Implementation

We implement our model in RevBayes (Höhna et al. 2016, 2017), an open-source software for Bayesian inference in phylogenetics. RevBayes is fully based on graphical models (Höhna et al. 2014), a unified framework for representing complex probabilistic models in the form of graphs where nodes correspond to model variables and edges to their probabilistic relationships. It allows the user to construct interactively their own phylogenetic graphical model in the Rev language, by combining hundreds of available models of nucleotide substitution, rate variation across sites and along the tree, and tree priors proposed in the literature (see Fig. 6(b) for an illustration with our model). Our three key additions consist of (i) introducing the OBDP distribution (Fig. 6(a)) into RevBayes, so that it can be used by the community within other graphical models; (ii) implementing the core algorithms responsible for computing the quantities L_t and M_t through time and eventually the final log-likelihood; and (iii) including a function to generate the posterior probability distribution of the number of lineages through time.

Figure 2 summarizes the full workflow to go from the raw data $\mathcal{O}, \mathcal{A}$ to the inferred reconstructed tree $\mathcal{T}$, model parameters θ, and diversity trajectories I_t.

Validation of the Method

Direct likelihood comparison.—We verify that the phylodynamic likelihood computed using L_t or M_t coincides with (i) previous RevBayes implementations (Höhna et al. 2017; Heath et al. 2019) of linear birth–death processes that are special cases of our framework, when no occurrences are included and $r = \omega = 0$ and (ii) an earlier Python implementation of the likelihood with constant parameters (Manceau et al. 2021). We use a small fixed data set and compute the likelihood using (i), (ii), and our implementation, under a wide range of parameters which are listed in Figure 3.

Quantitative validation of the MCMC implementation.—We follow a procedure called simulation-based calibration (Talts et al. 2018) for validating our MCMC implementation. It consists in the following three steps: (i) we define priors (Table 2) for all the involved parameters and simulate 1000 parameter sets in Python, trees with sampled fossils, occurrences, and genetic sequences (100 nucleotides long), conditioning on the survival of at least two lineages to the present; (ii) for each simulated data set, we use the same priors to infer the posterior distribution of reconstructed trees and parameters; and (iii) we compute the proportion of data sets for which the true (simulated) parameter values fall within a $100\alpha\%$ credible interval of the posterior distribution, for a range of α values (19 evenly spaced between 0.05 and 0.95). If the MCMC is correctly sampling the posterior distribution, the proportion of posterior credible intervals recovering the truth should be close to α.

For this analysis, N is chosen for each simulation as the true maximum number of hidden lineages plus a margin of 20, to avoid unnecessary computations. In practice, this choice requires expert knowledge specific to each particular clade.

SARS-CoV-2 Data Analysis

Molecular and occurrence data set.—We use the model implementation with piecewise constant rates to

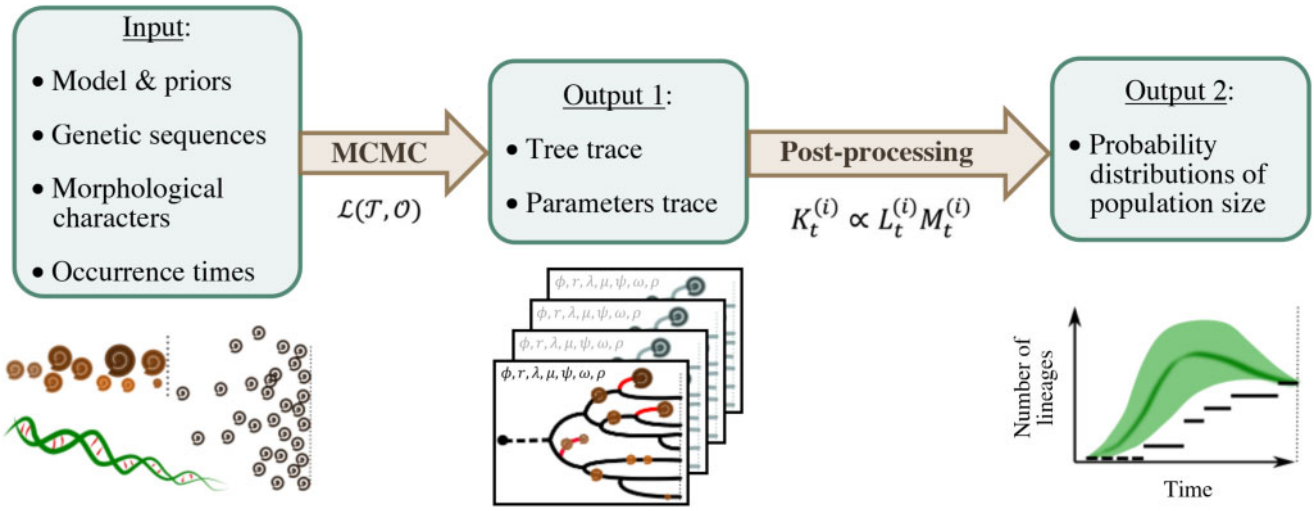

FIGURE 2. Workflow for using the occurrence birth–death (OBD) model for diversity inference. One first needs to specify a graphical model with priors and provide some empirical (molecular, morphological, and occurrence) data. A MCMC chain is run to sample the posterior distribution of trees and parameters, using joint likelihood $\mathcal{L}(\mathcal{T},\mathcal{O})$. Finally, these traces are used to compute the posterior distribution of the number of lineages through time (K_t).

perform a phylodynamic analysis of the spread of SARS-CoV-2 aboard the Diamond Princess cruise ship, a well-documented outbreak from February 2020. The outbreak is an example of a closely monitored, geographically constrained closed population, and thus constitutes an ideal case study of the disease dynamics and the mitigation policies undertaken.

The sequenced data used for this analysis consists of a set of 70 full-length viral genomes collected between February 15th and February 17th, all acquired from GISAID (Shu and McCauley 2017). Acknowledgements for laboratories that contributed the genome sequences used in this analysis are given in Supplementary Appendix G available on Dryad. All available sequences were aligned to reference genome MN908947, and sites subject to low sequencing accuracy were masked. Following the standard NextStrain pipeline, sites 13402, 24389, and 24390 as well as 150 bases at the ends of the genomes were masked, thought to be sequencing artifacts that would bias the alignment (Hadfield et al. 2018).

In this example, we define occurrences as patients testing positive for SARS-CoV-2 using reverse-transcription polymerase chain-reaction (RT-PCR) viral detection methods. Daily reports of new cases and total number of samples tested were published by the Japanese Ministry of Work throughout the outbreak and later compiled in the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong et al. 2020). Out of all 712 cases detected amongst passengers, we focus on the 705 cases detected while guests were still aboard the cruise ship, from the beginning of the cruise on January 20th until February 27th. Sequencing dates and case counts were communicated as daily reports throughout the outbreak. For all report entries, exact dates were uniformly assigned to all occurrences within each day. Additionally, we shift dates by a day to account for the delay between sampling and reporting of the PCR results. The full data set, in its original and processed formats, is presented in Supplementary Figure S16 available on Dryad.

Model assumptions.—The model parameterization allows us to examine two complementary aspects of the temporal change in epidemic spread. First, we estimate the effective reproductive number across all time intervals of interest. The reproductive number is the expected number of secondary cases produced by a single infected individual and is a standard epidemiological parameter, quantified in our model as $R_e = \frac{\lambda}{\mu + r(\omega+\psi)}$, with rate parameter μ encompassing either patient recovery or death in this application. Second, we infer the corresponding prevalence trajectories, to bring insight into the total infectious population throughout the outbreak. This includes potentially undetected asymptomatic patients, which are thought to make up a significant proportion of the total infected population (Mizumoto et al. 2020).

To achieve these two goals, we make full use of the skyline implementation of our model by fixing independent shifts for different rate parameters. In doing so, we closely follow the exact timeline of events of the outbreak. The testing strategy was initiated by Japanese authorities after a first guest was confirmed positive for SARS-CoV-2 on February 3rd. It was then extended to asymptomatic passengers from February 11th onward. Sequencing of some of the viral samples was then performed between February 15th and February 17th. To these four sampling parameters shifts, we additionally introduce another shift for the birth rate λ before the start of mandatory cabin isolation, on February 5th, producing the full timeline of $m = 5$ intervals.

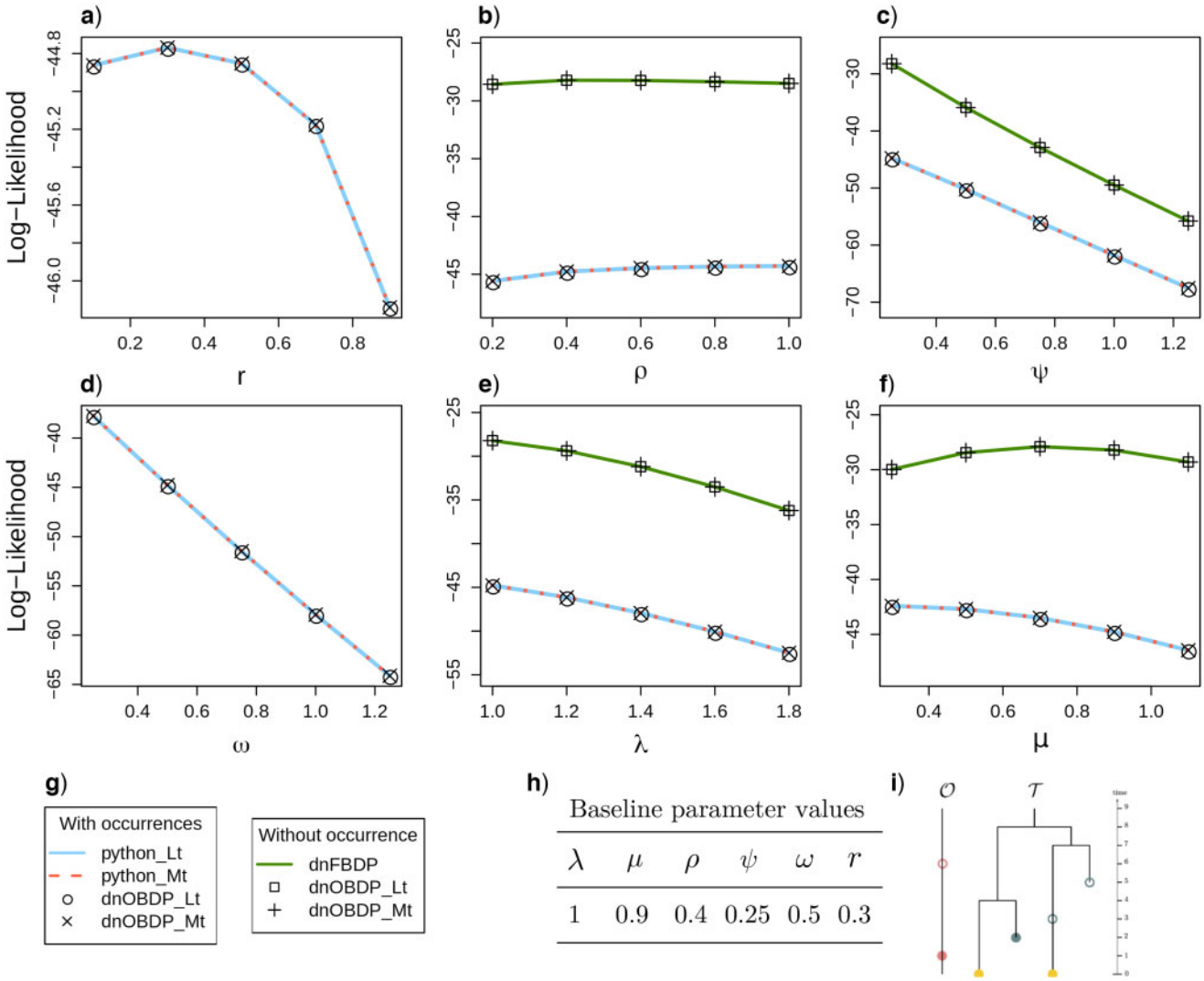

FIGURE 3. Validation of the likelihood calculation. Each parameter is varying a–f) while keeping the others at their baseline values h) and evaluating the likelihood of the toy data set i) where occurrences are shown on the right, and both samples at present and in the past are present on the tree. Filled dots are removed samples, unfilled are not removed. For all parameters a–f), our RevBayes implementation is compared to the Python code provided in Manceau et al. (2021) and whenever possible b, c, e, f), to earlier implementations of the FBDP available in RevBayes, fixing $r=0$, $\omega=0$, and $\mathcal{O}=\emptyset$.

TABLE 2. Prior distributions of the OBDP parameters for the quantitative validation test

Parameter	t_{or}	$\lambda-\mu$	μ	ψ	ω	r	ρ	Mutation rate
Prior or model	$\mathcal{U}(1,5)$	$\mathcal{E}(0.01)$	$\mathcal{E}(1)$	$\mathcal{E}(0.2)$	$\mathcal{E}(0.2)$	$\mathcal{U}(0,1)$	$\mathcal{U}(0.8,1)$	$\mathcal{E}(0.05)$

Note: $\mathcal{U}$ for Uniform distribution with given lower and upper bound, $\mathcal{E}$ for Exponential with given rate parameter. The model of molecular evolution is the Jukes–Cantor 1969 substitution model (JC69) with strict clock hypothesis.

Reports of the total number of samples tested were assembled to adjust prior means for $\omega+\psi$ on different time intervals and account for the extension of testing to asymptotic passengers. In total, testing efforts yielded 4066 samples over the entire period of interest, with 3622 of them being obtained after February 11th. All settings and priors used in this analysis are presented in detail in Supplementary Table S7 available on Dryad.

Cetacean Data Analysis

Context.—Cetaceans are a group of marine mammals, represented by 89 living species, that possess a remarkable and well-studied fossil record (Fordyce 2009). Their history can be summarized by three main phases (Marx et al. 2016), (i) starting 53 Ma, a 10 myr land-to-sea transition accompanied by drastic morphological transformations in the archaeocetes (stem cetaceans),

(ii) the emergence of neocetes (crown cetaceans, including filter-feeding mysticetes and echolocating odontocetes) at the Eocene–Oligocene boundary (∼34 Ma) and their radiation up to a Mid-Miocene peak (∼12 Ma) followed by (iii) a sharp decline in diversity in the last 4–6 myr.

Several studies have already attempted to estimate the diversity trajectory of cetaceans, using the fossil record (Uhen and Pyenson 2007), molecular phylogenies (Morlon et al. 2011), or both (Marx and Fordyce 2015) but even the latter total-evidence study did not include all fossil occurrences in its analyses. Although the initial huge discrepancies between the history inferred from the fossil record and from molecular phylogenies (Quental and Marshall 2010) have been partially bridged, including occurrences may help further provide a more reliable time-calibrated tree and a robust diversity trajectory estimation.

Molecular, morphological, and occurrence data sets.—The data can be subdivided into three parts: molecular, morphological, and occurrences. Data sets were collected and analyzed separately and are stored on the Open Science Framework (https://osf.io) (Aguirre-Fernández et al. 2020). Molecular data come from Steeman et al. (2009) and comprises 6 mitochondrial and 9 nuclear genes, for 87 of the 89 accepted extant cetacean species. Morphological data were obtained from Churchill et al. (2018), the most recent version of a widely used data set first produced by Geisler and Sanders (2003). After merging 2 taxa that are now considered synonyms on the Paleobiology Database (PBDB) and removing 3 outgroups that would have violated our model's assumptions, it now contains 327 variable morphological characters for 27 extant and 90 fossil taxa (mostly identified at the species level but 21 remain undescribed). In order to speed up the analysis, we further excluded the undescribed specimens and reduced this data set to the generic level by selecting the most complete specimen of each genus. Genus level has been the preferred unit of analysis in previous studies of cetacean diversity that include fossils (e.g., Marx and Uhen 2010; Dominici et al. 2020) since Uhen and Pyenson (2007) introduced it to counteract taxonomic inflation caused by the naming of extinct cetacean species based on fossil material of limited taxonomic value. Indeed, the computing cost increases rapidly with the maximum number of hidden lineages N (see Supplementary Appendix Figure S12 available on Dryad), to the point of becoming the bottleneck in our MCMC when $N > 100$. Given that a mid-Miocene peak diversity between 100 and 220 species is expected (Quental and Marshall 2010), with less than 100 observed lineages in our inferred tree at that time, N should therefore be about 150. Inferring instead the tree of cetacean genera allows us to reduce N to 70 hidden lineages. The final data set thus contains 41 extant and 62 extinct genera.

Occurrences come from the PBDB (data archive 9, M. D. Uhen) on May 11, 2020. The data set initially consisted of all 4678 cetacean occurrences, but the cetacean fossil record is known to be subject to several biases (Uhen and Pyenson 2007; Marx and Uhen 2010; Dominici et al. 2020). A detailed exploration (see Supplementary Appendix F available on Dryad) of this occurrence data set revealed several notable biases. First, an artifactual cluster of occurrences in very recent times, combined with other expected Pleistocene biases (Dominici et al. 2020), led us to remove all Late Pleistocene and Holocene occurrences. Second, we detected substantial variations in fossil recovery per time unit across lineages (see Supplementary Fig. S13 available on Dryad) resulting from oversampling of some species and localities, possibly due to greater abundance or spatiotemporal biases (Dominici et al. 2020). This observation violates our assumption of identical fossil sampling rates among taxa during a given interval. In order to reduce this bias, we retained occurrences identified at the genus level and further aggregated all occurrences belonging to an identical genus found at the same geological formation. In the case of occurrences for which the geological formation was not specified, we used geoplate data combined with stratigraphic interval as a proxy for geological formation. This resulted in a total of 968 occurrences retained for the analysis.

Model assumptions.—Each fossil comes along with a stratigraphic age uncertainty interval. Reducing this interval to either the midpoint, or a uniformly drawn point, has been shown to lead to serious biases in the divergence time estimates (Barido-Sottani et al. 2019). We instead follow the same procedure as Heath et al. (2019) and apply a uniform prior for the age of fossils with morphological characters, within the bounds of their stratigraphic age uncertainty. As a result, the age of a fossil included in the tree can slide within this interval during the MCMC.

Based on previous work showing huge discrepancies in mutation rates between odontocetes and mysticetes (Dornburg et al. 2012), and generally between nuclear and mitochondrial sequences (Allio et al. 2017) we considered a relaxed clock across the tree and partitioned between the two types of sequences. Much less biological knowledge is available about the dynamics of morphological characters (Wright 2019). We thus chose a minimal substitution model and partitioned the alignment in order to treat separately characters that are represented by a different number of states.

Moreover, we made the most of the piecewise-constant parameter framework to include 4four shifts in diversification and extinction rates at variable time points, using autocorrelated exponential priors in successive intervals. We also added shifts in fossilization rates in the Early Oligocene (Rupelian), Early Miocene (Aquitalian), and End Miocene (Messinian), in order to give more flexibility to the model during these geological periods of well-established low preservation (Marx et al. 2016). Additional flexibility could be introduced by increasing the number of rate shifts.

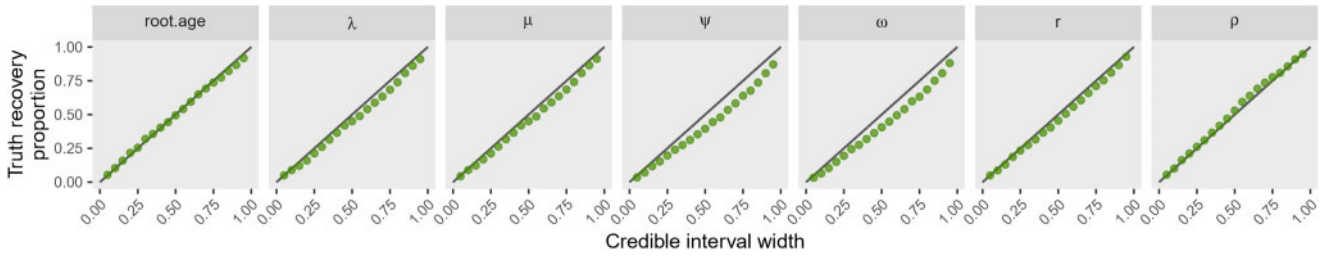

FIGURE 4. Results of the simulation-based calibration. Each dot corresponds to the proportion of simulated parameters (y-axis) falling within its inferred posterior credible interval with a given level (x-axis). The black line corresponds to the expected perfect match.

Finally, we added monophyly constraints on the mysticetes and odontocetes clades to further speed up the analysis.

All prior distributions are fully detailed in Supplementary Table S6 available on Dryad.

RESULTS

Validation of the Method

Direct likelihood comparison.—We illustrate in Figure 3 the perfect agreement with likelihood values computed using previous functions under a wide range of parameters, for both L_t and M_t traversal algorithms.

Quantitative validation of the MCMC implementation.—Figure 4 shows a good correspondence between the proportion of posterior credible intervals containing the true parameter value and the width of the credible interval. This indicates that the MCMC is properly calibrated, that is, samples adequately the targeted posterior distribution (see Supplementary Appendix E available on Dryad for more details).

Reproductive Number and Prevalence in the COVID Outbreak

Figure 5 shows the raw data, as well as the estimates of the total instantaneous prevalence and reproductive number through time.

The instantaneous prevalence is typically always slightly lower than the total number of new cases detected each day, or daily incidence. It also differs conceptually from the epidemiological prevalence in that detected cases are all assumed to be immediately removed from the infectious population upon sampling ($r=1$) through quarantining measures. For public health applications, all patients still infected with the virus are generally counted into the prevalence.

The reproductive number is inferred with very high uncertainty in the beginning of the epidemic, when very few cases were observed, and with a much higher precision in the second part of the process. It decreases synchronously with the launching of nonpharmaceutical interventions in early February (i.e., testing effort and cabin quarantine).

Total Diversity in the Cetacean Clade

Figure 6a shows the inferred diversity of cetacean genera over the past 50 million years. The diversity curve indicates an Early-Eocene origin followed by a monotonous diversification up to a first Mid-Miocene peak (12 Ma), before reaching its maximum in the Pliocene (between 2.6 and 5.3 Ma) with more than 100 inferred genera. The last few million years correspond to a sharp decline leading to the 41 extant genera.

This diversification history is reflected in Figure 6b by a 55 myr trend of slightly decreasing genus origination rates (from 0.16 to 0.10 genera per myr) and increasing genus extinction rates (from 0.035 to 0.085 genera per myr). Then, from the Pliocene to the present, both rates increase drastically to around 0.25 genera per myr, and the diversification rate reaches a minimum (but still non-negative, as expected with a lognormal prior). In parallel, the background fossil sampling rate is estimated between 0.020 and 0.025 samples per lineage and per myr, and lower values are recovered in the three geological periods with expected lower sedimentary record (more details in the Discussion). The inferred phylogenetic timetree is shown in the Supplementary Appendix Figure S14 available on Dryad.

DISCUSSION

Technical Achievements and Limitations

In this article, we extend the work of Gupta et al. (2020) and Manceau et al. (2021) to consider piecewise-constant rates through time, and implement the OBDP in the popular phylogenetic inference software RevBayes. This enables us to simultaneously incorporate numerous occurrences without character data, together with taxa for which we have genetic sequences and/or morphological characters. In addition to using the OBDP as a tree prior for inferring epidemiological or macroevolutionary parameters, tree topology and divergence dates, it allows users to compute the posterior probability distribution of the number of lineages through time, in a post-MCMC analysis (see workflow in Fig. 2). We validate the framework and illustrate its use in the fields of epidemiology and macroevolution.

The likelihood computation can be very fast when lineages become extinct upon sampling ($r=1$), relying

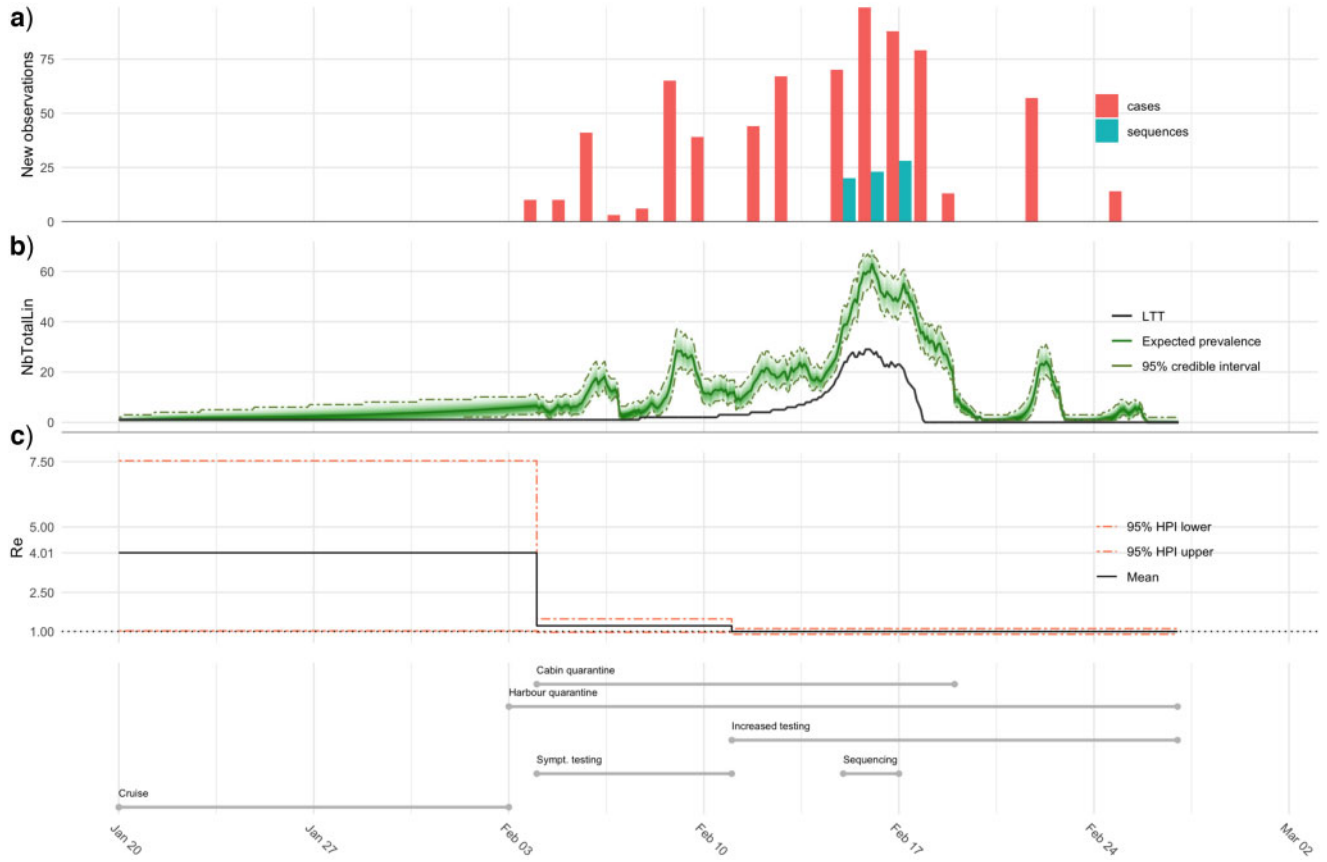

FIGURE 5. Analysis of the SARS-CoV-2 outbreak aboard the Diamond Princess cruise ship. a) Occurrence and sequenced data are plotted as daily new observations. We focus on infections detected while guests were still aboard, until the end of the harbor quarantine on February 27th. b) Posterior probability distribution of the instantaneous total infected population aboard the cruise ship. The 95% credible intervals (2.5% and 97.5% quantiles) are indicated in dashed lines, the expected prevalence with a solid green line and the inferred lineages through time (LTT) in black. c) Mean estimates and 95 % highest posterior density intervals (HPI) for the effective reproductive number (R_e) throughout a 38-day period starting at the beginning of the cruise.

on the results of Gupta et al. (2020), for data sets containing relatively few occurrences (in the order of a 100). This advantage is lost when including larger numbers of occurrences but can be recovered with the ingenious approximation developed by Zarebski et al. (2022). In practice, the assumption that these implementations rely on ($r=1$) only makes sense for some epidemiological applications, when infected individuals can self-quarantine and be safely assumed to be removed from the process. For macroevolutionary applications, the r parameter typically equals zero, and the likelihood computation relies on a more computationally intensive method to numerically solve Master equations (see details in Supplementary Appendix A available on Dryad). Supplementary Appendix Figure 12 available on Dryad provides an estimate of the run-time increase with N (on the order of $10^{-9}N^{7.5}$ min on a cluster) based on simulations. More work is thus needed to help speed up the likelihood computation when $r \neq 1$, on data sets for which a large number of hidden lineages is expected. This will be especially important for further applications in macroevolution and paleobiology, as many data sets feature thousands of fossils occurrences.

COVID-19 Diamond Princess Epidemic

The application of our method to the study of a thoroughly documented outbreak highlights the versatility of our model implementation. The ability to incorporate both incidence data and pathogen sequences, in combination with temporal information constitutes one of the first few instances of the use of a *combined-evidence* phylodynamic approach for the inference of epidemiological trajectories.

The conclusions that we draw from both our parameter estimates and the prevalence trajectory are consistent with other analyses. The basic reproductive number is inferred to be 4.01 in the absence of any intervention and detection, during the first 15 days of the cruise (see Fig. 5b), higher than most estimates for early global outbreaks (Lai et al. 2020; Nadeau et al. 2021). We then infer a decrease in reproductive number, which remains near or below 1 in the last 23 days of the time period of interest, suggesting that the epidemic was mostly contained by measures taken. These trends agree with other studies in magnitude, and although estimates ranging from 2.28 to 14.8 have been reported for the first

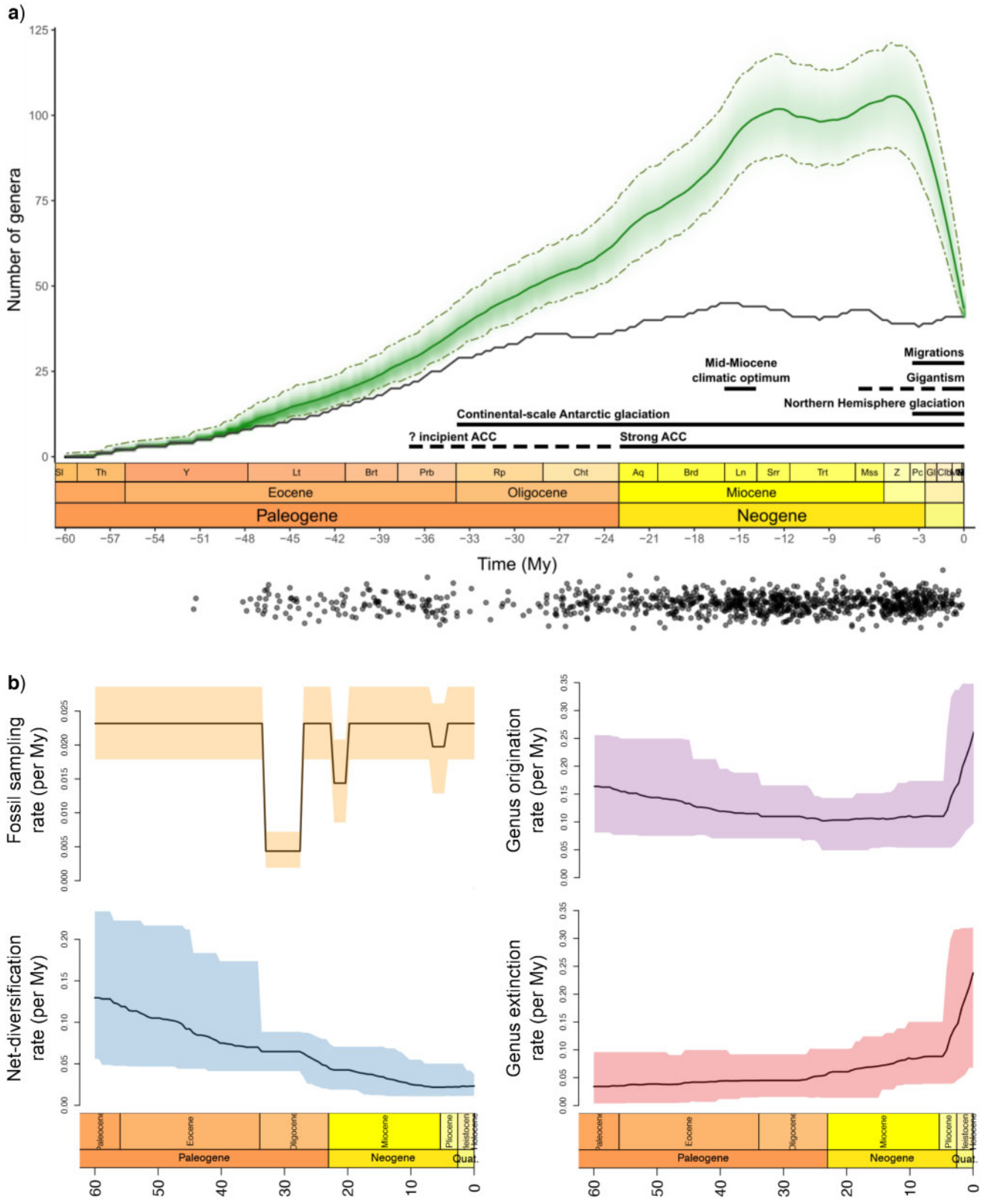

FIGURE 6. Inferred diversification history of Cetacea. a) Posterior probability distribution of the total number of genera over time. The 95% credible intervals (2.5% and 97.5% quantiles) are indicated in dashed lines, the expected diversity with a solid green line and the inferred LTT in black. Periods of biotic or abiotic factors that are hypothesized to have driven diversification changes are adapted from Marx and Fordyce (2015) and shown for information but had no influence on the analysis. ACC = Antarctic Circumpolar Current. Black dots below represent the occurrences used in the analysis. b) Posterior means and 95% credible intervals for extinction, origination, net diversification, and fossil sampling over time.

time period, most analyses infer a reproductive number below 2 in the subsequent time periods (Hoshino et al. 2021; Zhang et al. 2020; Rocklöv et al. 2020). Interestingly, we note that in our results the decrease in reproductive number is driven by the sampling and removal of individuals from the infectious population, with the total sampling rate going up from 1.76 to 2.14 days^{-1}, after February 11th (see Supplementary Fig. 17 available on Dryad for the detailed timeline). This extended sampling, which can be attributed to the decision to test asymptomatic passengers, results in occurrences strongly influencing the estimated instantaneous prevalence. This highlights the successful integration of both sequence and occurrence data, and the added value brought about by this new implementation.

Biases inherent to many epidemiological data sets indicate important areas for development. For instance, sampling of outbreaks is most often carried out with the aim of quickly monitoring the disease, without rigorously following a protocol. This can result in inconsistent sampling and reporting strategies, with gaps and/or missing data. We note for example that no new cases were reported between February 19th and February 22th. Although it is likely due to a delay in reporting or testing, we did not model a drop in sampling as no changes in the monitoring efforts were reported. Additionally, due to a 24 h reporting delay for this data set, the first detected case was originally placed after the start of the quarantine. We tried to meticulously remove as many such biases as possible, but our framework could also be improved in the future to explicitly account for these.

Other potential biases include the effect of population structure—the Diamond princess outbreak is likely to have spread in at least two distinct subpopulations: guests and crew members (Nishiura 2020)—and density dependence—the outbreak being in a closed, geographically constrained population (Rocklöv et al. 2020). Further developments of the method to cover these scenarios could provide even better insight into the dynamics of this outbreak.

Past Cetacean Diversity

Molecular and paleontological data come with their inherent limitations, for example, a lack of information about extinct lineages for the former and substantial spatiotemporal biases for the latter. Combining them into a single analysis may gather enough signal to mitigate these limitations, but special attention should be paid to model assumptions. We have endeavored to respect these constraints, by (i) correcting occurrence distribution sampling biases (see Supplementary Appendix E available on Dryad) and (ii) including skyline variation in diversification and preservation rates.

The emerging patterns of cetacean generic diversification in Figure 6 are consistent with previous estimates (Uhen and Pyenson 2007; Morlon et al. 2011; Marx and Fordyce 2015): (i) the "boom and bust" dynamics of prolonged diversification followed by a recent decline is recovered and (ii) estimated generic richness is higher than the incomplete raw generic counts, as expected. The diversification of cetaceans, starting in the Eocene and accelerating in the Neogene, has been associated by previous authors with the development of the Antarctic Circumpolar Current (ACC) that fueled a diatom radiation, via nutrient supply, prompting the diversification of bulk filtering cetaceans. The diversity drop in the last 4 myr has been linked to the global climate deterioration and the Northern Hemisphere glaciation, which coincides with the final establishment of modern mysticete gigantism and long-distance migration. Our inferred diversity trajectory (Fig. 6) is compatible with these hypotheses. On the other hand, the distinct second peak with maximum diversity in the Pliocene is unexpected and will require further investigation.

When it comes to the dated phylogeny, many of the deepest nodes are much older than what has been inferred previously in the literature. This observation could be explained by an unbalanced sampling of fossil genera included in the tree compared to fossil occurrences: 1–8–50 ratio of archaeocete–mysticete–odontocete specimens with character information versus a 1–2–8 ratio for occurrences (based on the PBDB taxonomy information). Indeed, the few archaeocete lineages included may then appear incoherent with the many Eocene occurrences, leading to an artificially early placement of the Neocete diversification. This pattern of differential preservation between younger and older lineages is to be expected and could be accounted for with increasing occurrence sampling rates over time. We therefore advise to account for similar sampling heterogeneity when performing similar analyses in the future. More broadly, simulations are required to examine in more detail the impact of spatiotemporal sampling biases (Close et al. 2020) on this and other approaches to estimating time tree and diversity.

New Avenues for Phylogenetics

Over the last decade, the field of phylogenetics has expanded considerably with the development of the FBDP and related extensions, of which the OBDP is the latest instance. As a result, the long-standing opposition between molecular-based and fossil-based macroevolutionary inferences is in the process of being bridged, and case count records can be analyzed jointly with sequencing data in epidemiology applications. Many extant clades with a relatively rich paleontological record—for example, turtles, sharks, and angiosperms—as well as outbreak surveillance data, could benefit from this new method to infer reliable phylogenies and diversity/prevalence trajectories.

Future progress could be made to couple birth rates with abiotic drivers, such as biogeography (see also work on multitype birth–death processes Scire et al. 2020), or biotic drivers such as density dependence (see also Etienne et al. 2012). Going even further down the

mechanistic road for macroevolutionary applications, stratigraphic palaeobiology could even become an explicit part of diversification models, by considering the accumulation of sediments over finer time and spatial scales (Patzkowsky and Holland 2012). We anticipate that these approaches will all benefit from combining paleontological and molecular data.

Overall, our two empirical applications demonstrate that a phylogenetic framework can be successfully applied to recover both the past outbreak prevalence and the past paleodiversity. In contrast to alternative approaches, it maximizes the use of available evidence, since it uniquely allows us to combine genetic and morphological character data, together with occurrences. Further, our inference method relies on a generating model, incorporating explicit assumptions about the processes giving rise to our data, including sampling. We argue that it is prospectively more flexible than alternative approaches to mitigating sampling biases.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.p8cz8w9rq.

Funding

This work was supported by a Postdoctoral Fellowship funded by Eidgenossische Technische Hochschule Zurich to M.M.

References

Aguirre-Fernández G., Warnock R., Benites-Palomino A.M., Andréoletti J., Manceau M. 2020. Cetacean timeline. Available from: osf.io/78xys.

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in Jawed vertebrates. Proc. Natl. Acad. Sci. USA 106:13410–13414.

Allio R., Donega S., Galtier N., Nabholz B. 2017. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. Mol. Biol. Evol. 34:2762–2772.

Barido-Sottani J., Aguirre-Fernández G., Hopkins M.J., Stadler T., Warnock R. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth-death process. Proc. R. Soc. B 286:20190685.

Churchill M., Geisler J.H., Beatty B.L., Goswami A. 2018. Evolution of cranial telescoping in echolocating whales (Cetacea: Odontoceti). Evolution 72:1092–1108.

Close R., Benson R.B., Saupe E., Clapham M., Butler R. 2020. The spatial structure of phanerozoic marine animal diversity. Science 368:420–424.

Dominici S., Danise S., Cau S., Freschi A. 2020. The awkward record of fossil whales. Earth-Sci. Rev. 205:103057.

Dong E., Du H., Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20:533–534.

Dornburg A., Brandley M.C., McGowen M.R., Near T.J. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol. Biol. Evol. 29:721–736.

Drummond A.J., Nicholls G.K., Rodrigo A.G., Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore A.B. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. R. Soc. B 279:1300–1309.

Featherstone L.A., Di Giallonardo F., Holmes E.C., Vaughan T.G., Duchêne, S. 2021. Infectious disease phylodynamics with occurrence data. Methods Ecol. Evol. 12:1498–1507.

Fordyce R.E. 2009. Cetacean fossil record. In: Perrin W.F., Würsig B., Thewissen J.G.M., eds. Encyclopedia of marine mammals. 2nd ed. London: Academic Press. p. 207–215.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2016. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66:57–73.

Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput. Biol. 10:e1003919.

Geisler J.H., Sanders A.E. 2003. Morphological evidence for the phylogeny of Cetacea. J. Mamm. Evol. 10:23–129.

Gupta A., Manceau M., Vaughan T., Khammash M., Stadler T. 2020. The probability distribution of the reconstructed phylogenetic tree with occurrence data. J. Theor. Biol. 488:110115.

Hadfield J., Megill C., Bell S.M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., Neher R.A. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

He X., Lau E.H., Wu P., Deng X., Wang J., Hao X., Lau Y.C., Wong J.Y., Guan Y., Tan X., et al. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat. Med. 26:672–675.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. USA 111:2957–2966.

Heath T.A., Wright A.M., Pett W. 2019. RevBayes: combined evidence analysis and the fossilized birth-death process for stratigraphic range data. Available from https://revbayes.github.io/tutorials/fbd/.

Höhna S., Heath T. 2019. RevBayes: simple diversification rate estimation. Available from https://revbayes.github.io/tutorials/divrate/simple.html.

Höhna S., Heath T.A., Boussau B., Landis M.J., Ronquist F., Huelsenbeck J.P. 2014. Probabilistic graphical model representation in phylogenetics. Syst. Biol. 63:753–771.

Höhna S., Landis M.J., Heath T.A. 2017. Phylogenetic inference using RevBayes. Curr. Protocols Bioinformatics 57:6.16.1–6.16.34.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

Hoshino K., Maeshiro T., Nishida N., Sugiyama M., Fujita J., Gojobori T., Mizokami M. 2021. Transmission dynamics of SARS-CoV-2 on the Diamond Princess uncovered using viral genome sequence analysis. Gene 779:145496.

Kendall D.G. 1948. On the generalized 'birth-and-death' process. Ann. Math. Stat. 19:1–15.

Kingman J.F.C. 1982. The coalescent. Stoch. Proc. Appl. 13:235–248.

Lai A., Bergna A., Acciarri C., Galli M., Zehender G. 2020. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. J. Med. Virol. 92:675–679.

Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.

Manceau M., Gupta A., Vaughan T., Stadler T. 2021. The probability distribution of ancestral population size under birth-death processes. J. Theor. Biol. 509:110400.

Marshall C.R. 2017. Five palaeobiological laws needed to understand the evolution of the living biota. Nat. Ecol. Evol. 1:1–6.

Marx F.G., Fordyce R.E. 2015. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. R. Soc. Open Sci. 2:140434.

Marx F.G., Lambert O., Uhen M.D. 2016. Cetacean paleobiology. Chichester, UK: Wiley Blackwell.

Marx F.G., Uhen M.D. 2010. Climate, critters, and cetaceans: Cenozoic drivers of the evolution of modern whales. Science 327:993–996.
McGowen M.R., Tsagkogeorga G., Álvarez Carretero S., dos Reis M., Struebig M., Deaville R., Jepson P.D., Jarman S., Polanowski A., Morin P.A., Rossiter S.J. 2020. Phylogenomic resolution of the Cetacean tree of life using target sequence capture. Syst. Biol. 69:479–501.
Mizumoto K., Kagaya K., Zarebski A., Chowell G. 2020. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance 25:2000180.
Morlon H., Parsons T.L., Plotkin J.B. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108:16327–16332.
Nadeau S.A., Vaughan T.G., Scire J., Huisman J.S., Stadler T. 2021. The origin and early spread of SARS-CoV-2 in Europe. Proc. Natl. Acad. Sci. USA 118:e2012008118.
Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. Philos. T. R. Soc. B. 344:305–311.
Nishiura H. 2020. Backcalculating the incidence of infection with covid-19 on the diamond princess. J. Clin. Med. 9:657.
O'Reilly J.E., Donoghue P.C. 2020. The effect of fossil sampling on the estimation of divergence times with the fossilized birth–death process. Syst. Biol. 69:124–138.
Parag K.V., du Plessis L., Pybus O.G. 2020. Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. Mol. Biol. Evol. 37:2414–2429.
Patzkowsky M.E., Holland S.M. 2012. Stratigraphic paleobiology: understanding the distribution of fossil taxa in time and space. Chicago:University of Chicago Press.
Quental T.B., Marshall C.R. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends Ecol. Evol. 25:434–441.
Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9:e89543.
Rasmussen D.A., Ratmann O., Koelle K. 2011. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Comput. Biol. 7:1002136.
Raup D.M. 1972. Taxonomic diversity during the phanerozoic. Science 177:1065–1071.
Rocklöv J., Sjödin H., Wilder-Smith A. 2020. COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. J. Travel. Med. 27:taaa030.
RStudio Team. 2020. RStudio: integrated development environment for R. Boston, MA: RStudio, PBC.
Scire J., Barido-Sottani J., Kühnert D., Vaughan T.G., Stadler T. 2020. Improved multi-type birth-death phylodynamic inference in BEAST 2. bioRxiv, 895532.
Sepkoski J.J., Bambach R.K., Raup D.M., Valentine J.W. 1981. Phanerozoic marine diversity and the fossil record. Nature 293:435–437.
Shu, Y. and McCauley J. 2017. GISAID: global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22:30494.
Silvestro D., Salamin N., Antonelli A., Meyer X. 2019. Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. Paleobiology 45:546–570.
Silvestro D., Salamin N., Schnitzler J. 2014. PyRate: a new program to estimate speciation and extinction rates from incomplete fossil data. Methods Ecol. Evol. 5:1126–1131.
Silvestro D., Warnock R.C., Gavryushkina A., Stadler T. 2018. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. Nat. Commun. 9:1–14.
Stadler T. 2010. Sampling-through-time in birth–death trees. J. Theor. Biol. 267:396–404.
Stadler T., Gavryushkina A., Warnock R.C., Drummond A.J., Heath T.A. 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. J. Theor. Biol. 447:41–55.
Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (HCV). Proc. Natl. Acad. Sci. USA 110: 228–233.
Steeman M.E., Hebsgaard M.B., Fordyce R.E., Ho S.Y., Rabosky D.L., Nielsen R., Rahbek C., Glenner H., Sørensen M.V., Willerslev E. 2009. Radiation of extant cetaceans driven by restructuring of the oceans. Syst. Biol. 58:573–585.
Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A. 2018. Validating Bayesian inference algorithms with simulation-based calibration. arXiv:1804.06788.
Tribble C.M., Freyman W.A., Landis M.J., Lim J.Y., Barido-Sottani J., Kopperud B.T., Höhna S., May M.R. 2021. Revgadgets: an R package for visualizing Bayesian phylogenetic analyses from RevBayes. Methods Ecol. Evol. 3:314–323.
Uhen M., Pyenson N. 2007. Diversity estimates, biases, and historiographic effects: resolving cetacean diversity in the tertiary. Palaeontol. Electron. 10:1–22.
Vaughan T.G., Leventhal G.E., Rasmussen D.A., Drummond A.J., Welch D., Stadler T. 2019. Estimating epidemic incidence and prevalence from genomic data. Mol. Biol. Evol. 36:1804–1816.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York:Springer.
Wright A.M. 2019. A systematist's guide to estimating Bayesian phylogenies from morphological data. Insect. Syst. Diver. 3:2.
Wright A.M. 2020. RevBayes: discrete morphology - multistate characters. Available from https://revbayes.github.io/tutorials/morph_tree/V2.html.
Xing Y., Onstein R.E., Carter R.J., Stadler T., Peter Linder H. 2014. Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected. Evolution 68:2821–2832.
Yule G.U. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos. Trans. R. Soc. Lond. B. 213:21–87.
Zarebski A.E., du Plessis L., Parag K.V., Pybus O.G. 2022. A computationally tractable birth-death model that combines phylogenetic and epidemiological data. PLoS Comput. Biol. 18:e1009805.
Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2015. Total-evidence dating under the fossilized birth–death process. Syst. Biol. 65:228–249.
Zhang S., Diao M., Yu W., Pei L., Lin Z., Chen D. 2020. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: a data-driven analysis. Int. J. Infect. Dis. 93:201–204.