A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences

Jérémy Andréoletti, Antoine Zwaans, Rachel C M Warnock, Gabriel Aguirre-Fernández, Joëlle Barido-Sottani, Ankit Gupta, Tanja Stadler, Marc Manceau

HAL Id: hal-03123037
https://hal.science/hal-03123037

Preprint submitted on 27 Jan 2021

A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences

Jérémy Andréoletti*, Antoine Zwaans*,
Rachel C. M. Warnock, Gabriel Aguirre-Fernández, Joëlle Barido-Sottani,
Ankit Gupta, Tanja Stadler, Marc Manceau

October 27, 2020

Affiliations :

Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Paleontological Institute and Museum, University of Zürich, Zürich, Switzerland
Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, USA

Keywords :

skyline birth-death process; epidemiology; macroevolution; phylogenetics; fossils

Abstract

Phylodynamic models generally aim at jointly inferring phylogenetic relationships, model parameters, and more recently, population size through time for clades of interest, 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, i.e. lacking morphological and molecular data, which we call *occurrences*, provide invaluable information to reconstruct past population sizes.

Here, we present new methodological developments around the Fossilized Birth-Death Process enabling us to (i) efficiently incorporate occurrence data while remaining computationally tractable and scalable; (ii) consider piecewise-constant birth, death and sampling rates; and (iii) reconstruct past population sizes, 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 datasets stemming from the fields of epidemiology and macroevolution. In epidemiology, we apply our model to the Covid-19 outbreak on the Diamond Princess ship. We infer the total prevalence throughout the outbreak, by taking into account jointly the case count record (occurrences) along with viral sequences for a fraction of infected individuals. In macroevolution, we present an empirical case study of cetaceans. We infer the diversity trajectory using molecular and morphological data from extant taxa, morphological data from fossils, as well as numerous fossil occurrences. Our case studies highlight that the advances we present allow us to further bridge the gap between between epidemiology and pathogen genomics, as well as paleontology and molecular phylogenetics.

1 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 individuals appear, and the rate at which individuals 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 twentieth 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, i.e. the tree obtained by pruning all extinct lineages from the full genealogical history of the process (see Fig. 1A). 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 largely 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 scarce, could much better inform extinction estimates (Quental and Marshall 2010). An extension introduced by Stadler (2010) and dubbed the *Fossilized Birth-Death Process* (FBD) (Heath et al. 2014) has allowed us to model jointly 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 throughout its lifetime at a fixed rate, and appear in the reconstructed tree (see Fig. 1B). The probability density of the resulting phylogeny is derived in closed-form, and has been successfully used as a prior in Bayesian phylodynamic analyses to study the diversification history of hymenopterans (Zhang et al. 2015), as well as the penguins (Gavryushkina et al. 2016). The same model was also used in the context of epidemiology, where infected individuals can as well be sampled throughout the infectious period and appear in the reconstructed tree Stadler et al. (2013). Finally, model extensions have been introduced to help take into account the age of species (i.e. stratigraphic ranges) or, in the context of epidemiology, an extended period of infection (Stadler et al. 2018).

An important feature of many standard paleontological datasets, 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 paper, 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. This data, while providing no useful information regarding the topology of the tree, still contain invaluable information regarding the underlying population size (see Fig. 1C). 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, 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 (Heath et al. 2014; Gavryushkina et al. 2014; O'Reilly and Donoghue 2020). However, in the event both samples with character data and occurrence data are available, applying the standard FBD model requires making the assumption that both sets of data were generated under the same process and with the same rate. 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. 1D). 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) presents 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. (2019) 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 individuals sampled through time are removed from the process upon sampling. Finally, under the same model assumptions, Manceau et al. (2019) presented a method to compute the population size distribution conditioned on a reconstructed tree and a record of occurrences.

In this paper, 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

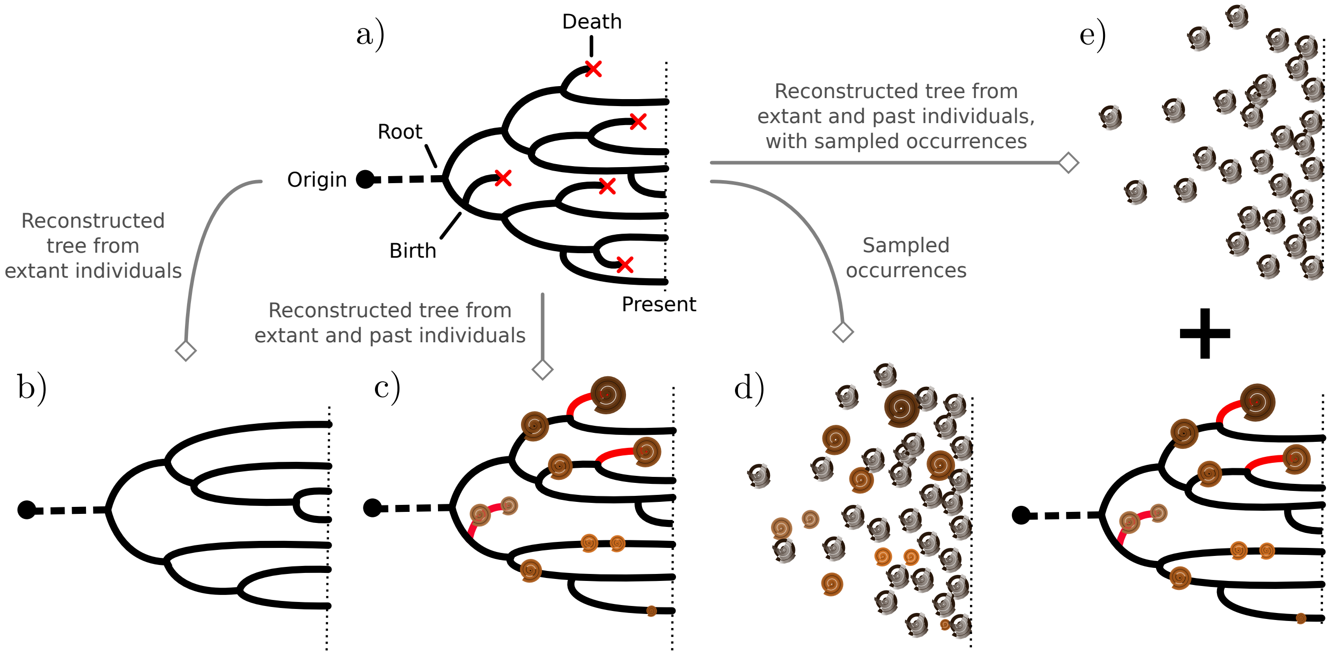

a)
Death
Root
Origin
Birth
Present
Reconstructed tree from extant and past individuals, with sampled occurrences
e)
Reconstructed tree from extant individuals
Reconstructed tree from extant and past individuals
Sampled occurrences
+
b)
c)
d)

Figure 1: Different approaches to infer past history and population sizes. a) The full unknown history of the population. Different types of data can be used in order to infer the past history and population sizes of the population. b) Genetic sequencing data and character data for present day individuals allows to infer the reconstructed phylogenetic tree and to estimate the past number of lineages. c) This tree can be enriched with past individuals with documented character data (either morphological or molecular data), adding information on some extinct lineages (red). d) Past population sized can be obtained from sampled occurrences alone. Finally, e) A more comprehensive total evidence method integrates extant genetic sequences, samples with character data and the occurrences in a unified framework.

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 population sizes 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 datasets coming from the fields of epidemiology and macroevolution. In epidemiology, we infer the prevalence through time for the 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.

2 Material and methods

2.1 Phylodynamic model

We consider that a population of individuals starts at the time of origin t_{or} with one individual, 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 successively 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, individuals reaching present time are included in the tree with a fixed probability ρ. Finally, upon sampling, individuals are removed with a piecewise-constant probability of removal r_t.

As a result of these three sampling steps, we observe a reconstructed tree $\mathcal{T}$, which is the tree spanning all

Table 1: Parameters and objects of the Occurrence Birth-Death Process.

Parameter	Signification
t_{or}	Time of origin
λ	Speciation rate
μ	Extinction rate
ψ	Fossil sampling rate
ω	Occurrence sampling rate
r	Removal probability at sampling
ρ	Sampling probability at present

Object	Signification
$\mathcal{T}$	Reconstructed tree
$\mathcal{O}$	Record of occurrence times
I_t	Total number of lineages
k_t	Number of sampled lineages
i	Number of hidden lineages
$(\mathcal{O}_t^\uparrow, \mathcal{T}_t^\uparrow)$	Occurrences and tree before time t
$(\mathcal{O}_t^\downarrow, \mathcal{T}_t^\downarrow)$	Occurrences and subtrees after time t

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 individuals 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 individuals.

In Appendix A, we extend the method introduced by Manceau et al. (2019) 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 $[t, t_{\text{or}}]$, 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,

$$\forall i \in \mathbb{N}, \quad M_t^{(i)} := \mathbb{P}\left(\mathcal{T}_t^\uparrow, \mathcal{O}_t^\uparrow, I_t = k_t + i\right) \tag{2.1}$$

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,

$$\forall i \in \mathbb{N}, \quad L_t^{(i)} := \mathbb{P}\left(\mathcal{T}_t^\downarrow, \mathcal{O}_t^\downarrow \mid I_t = k_t + i\right) \tag{2.2}$$

which can as well be used to compute, upon reaching the time of origin t_{or}, $\mathbb{P}(\mathcal{T}, \mathcal{O}) = L_{t_{\text{or}}}^{(0)}$.
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 (Appendix A). Note that in this numerical evaluation, we have to make one approximation namely, assume a maximal population size N (while in theory the population size may 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 total number of individuals living at time t is given by,

$$\begin{aligned} K_t^{(i)} &:= \mathbb{P}(I_t = k_t + i \mid \mathcal{T}, \mathcal{O}) \\ &\propto \mathbb{P}(I_t = k_t + i, \mathcal{T}_t^\uparrow, \mathcal{O}_t^\uparrow, \mathcal{T}_t^\downarrow, \mathcal{O}_t^\downarrow) \\ &\propto \mathbb{P}(\mathcal{T}_t^\downarrow, \mathcal{O}_t^\downarrow \mid I_t = k_t + i, \mathcal{T}_t^\uparrow, \mathcal{O}_t^\uparrow)\mathbb{P}(I_t = k_t + i, \mathcal{T}_t^\uparrow, \mathcal{O}_t^\uparrow) \\ &\propto L_t^{(i)} M_t^{(i)} \end{aligned} \tag{2.3}$$

where the last equality is due to the Markov property of the process. We summarize all the notation introduced above in Table 1.

2.2 Bayesian inference framework

We consider a Bayesian inference framework with additional model layers for character data evolution along the reconstructed tree $\mathcal{T}$. For epidemiology 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 macroevolution applications, we superimpose (i) a model of morphological evolution, leading to the 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. Summarizing all (molecular and morphological) character data together as $\mathcal{A}$, and all model parameters as θ, 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:

$$\mathbb{P}(\mathcal{T},\theta|\mathcal{O},\mathcal{A}) \propto \mathbb{P}(\mathcal{T},\mathcal{O}|\theta)\mathbb{P}(\mathcal{A}|\mathcal{T},\theta)\mathbb{P}(\theta) \; . \tag{2.4}$$

First, we sample this posterior distribution using a Metropolis-Hastings MCMC. Second, the posterior probability distribution of the ancestral population size can be written as,

$$\mathbb{P}(I_t|\mathcal{A},\mathcal{O}) = \int_{\mathcal{T},\theta} \mathbb{P}(I_t|\mathcal{T},\mathcal{O},\theta)d\mathbb{P}(\mathcal{T},\theta|\mathcal{O},\mathcal{A}) \tag{2.5}$$

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

2.3 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 of their probabilistic relationships. It allows the user to construct interactively their own phylogenetic graphical model in the Rev language, by combining the hundreds of available models of nucleotide substitution, rate variation across sites and along the tree, and tree priors proposed in the literature (see Supp Fig. S6B for an illustration with our model). Our three key additions consist of (i) introducing the OBDP distribution (Supp Fig S6A) into RevBayes, so that everyone can use it with their own 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 ancestral population size 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.

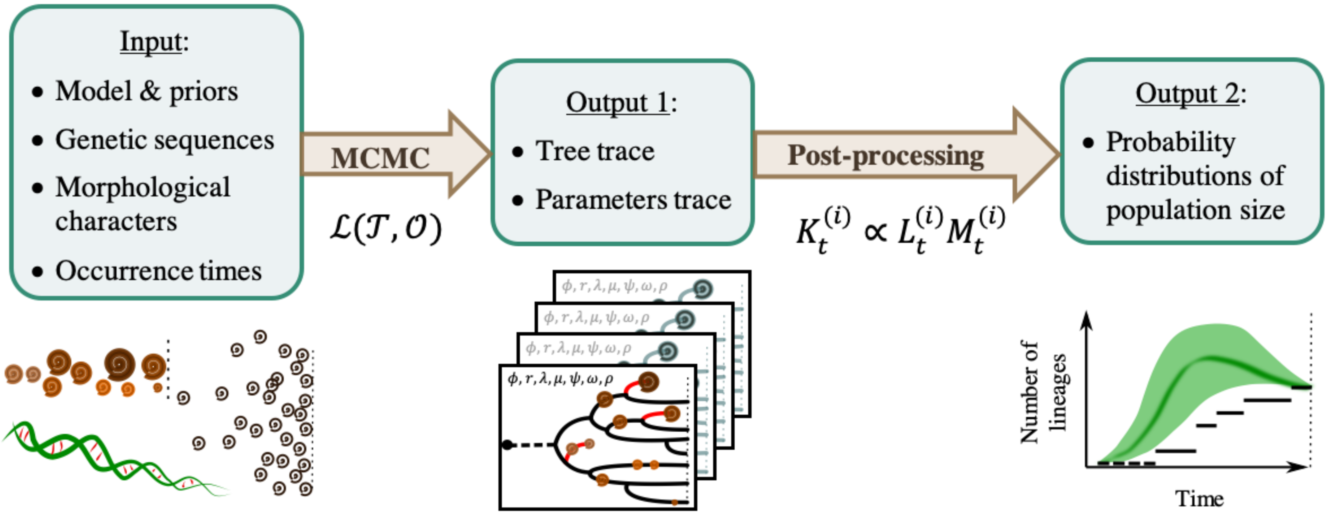

Input:
Model & priors
Genetic sequences
Morphological characters
Occurrence times
MCMC
$\mathcal{L}(\mathcal{T},\mathcal{O})$
Output 1:
Tree trace
Parameters trace
$\phi, r, \lambda, \mu, \psi, \omega, \rho$
Post-processing
$K_t^{(i)} \propto L_t^{(i)} M_t^{(i)}$
Output 2:
Probability distributions of population size
Number of lineages
Time

Figure 2: Workflow for using the OBD model for diversity inference. One first needs to specify a graphical model with priors and provide some empirical (molecular, morphological, occurrence) data. A MCMC is run to sample the posterior distribution of trees and parameters. Finally, these traces are used to compute the posterior distribution of diversity through time.

2.4 Validation of the method

2.4.1 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. 2019). We use a small fixed dataset and compute the likelihood using (i), (ii) and our implementation, under a wide range of parameters which are listed in Figure 3.

2.4.2 Qualitative validation on simulated datasets

Time-forward simulations of two full OBD processes were performed to produce two datasets. On the first one (dataset 1), we simulated morphological data for past samples, and both morphological and molecular data for extant samples, mimicking a macroevolution scenario. On the second one (dataset 2), we simulated only molecular data for all samples. Parameter values and the full model specifications are described in Supp Mat C. We used the RevBayes implementation to infer back the parameter values along with the reconstructed tree and population sizes. We performed this validation as a blind test, with two of us being in charge of conducting the analysis, and remaining ignorant of the true simulated scenarios.

2.4.3 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, trees with sampled fossils, occurrences and genetic sequences (100 nucleotides long); (ii) for each simulated dataset, we use the same priors to infer the posterior distribution of reconstructed trees and parameters; and (iii) we compute the proportion of datasets 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.

Table 2: Prior distributions of the OBDP parameters for the quantitative validation test. Notations: U for Uniform distribution with given lower and upper bound, Exp for Exponential with given rate parameter. The model of molecular evolution is the Jukes-Cantor 1969 substitution model (JC69) with strict clock hypothesis.

Parameter	t_{or}	μ	$\lambda - \mu$	r	ψ	ω	ρ	Mutation rate
Prior or Model	$\mathcal{U}(1,5)$	$Exp(1)$	$Exp(100)$	$\mathcal{U}(0,1)$	$Exp(5)$	$Exp(5)$	$\mathcal{U}(0.8,1)$	$Exp(100)$

2.5 Covid data analysis

2.5.1 Molecular and occurrence dataset

We use the model implementation with piecewise constant rates to perform a phylodynamic analysis of the SARS-CoV-2 epidemic 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 (Mallapaty 2020).

The sequenced data used for this analysis consists of a set of 71 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 Supp. Mat. E.1. All available sequences were aligned to reference genome MN908947 and sites subject to low sequencing accuracy were masked.

Following the standard NextStrain (Hadfield et al. 2018) pipeline, sites 13402, 24389 and 24390 as well as 150 bases at the ends of the genomes were masked, thought to be sequencing artefacts that would bias the alignment.

In this example, we define occurrences as patients testing positive for SARS-CoV-2 using polymerase chain-reaction (PCR) detection methods, recorded as case counts. These case counts, i.e. the daily reports of new cases, along with the total number of samples tested were published by the Japanese Ministry of Work throughout the outbreak; case counts were then compiled in the JHU CSSE database (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 randomly 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 dataset, in its original and processed formats, is presented in Figure S12.

2.5.2 Model assumptions

The model parametrisation 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 $R_e = \frac{\lambda}{\mu + r(\omega + \psi)}$. Second, we infer the corresponding prevalence trajectories, aiming to gain insight into the number of potentially undetected patients that are thought to make up a significant proportion of the infected population (Mizumoto et al. 2020).

To achieve these two goals, we make full use of the skyline implementation of our model by allowing independent shifts for each rate parameter. In doing so, we closely follow the exact timeline of events of the outbreak. We first incorporate information regarding the testing and sequencing procedures. 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 4 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.

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 other settings and priors used in this analysis are presented in detail in Supp. Table S7.

2.6 Cetacean data analysis

2.6.1 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 Ma.

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. The initial huge discrepancies between the history inferred from the fossil record and from molecular phylogenies (Quental and Marshall 2010) have been partially bridged, but including occurrences may help provide a more reliable time-calibrated tree and a robust diversity trajectory estimation.

2.6.2 Molecular, morphological and occurrence datasets

The data can be subdivided in three parts: molecular, morphological, and occurrences. Datasets were collected and analysed separately and are stored on the Open Science Framework (https://osf.io) (Aguirre-Fernández et al. 2020). Molecular data comes from Steeman et al. (2009), and comprises 6 mitochondrial and 9 nuclear genes, for 87 of the 89 accepted extant cetacean species. Morphological data was obtained from Churchill et al. (2018), the most recent version of a widely-used dataset 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 dataset to the generic level by selecting the most complete specimen in each genera. Indeed, the computing cost increases quadratically with the maximum number of hidden lineages N, 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 dataset thus contains 41 extant and 62 extinct genera.

Occurrences come from the PBDB (data archive 9, M. D. Uhen) on May 11, 2020. The dataset 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 et al. 2016; Dominici et al. 2020). A detailed exploration (see Supp. Mat. D) of this occurrence dataset revealed several notable biases. First, an artefactual 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 Supp. Fig. S10) resulting from oversampling of some species and localities, possibly due to greater abundance or spatio-temporal 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. Occurrences for which the geological formation was not specified, we used geoplate data combined with stratigraphic interval. This resulted in a total of 968 occurrences retained for the analysis.

2.6.3 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 partitioned between the two types of sequences and considered a relaxed clock across the tree.

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.

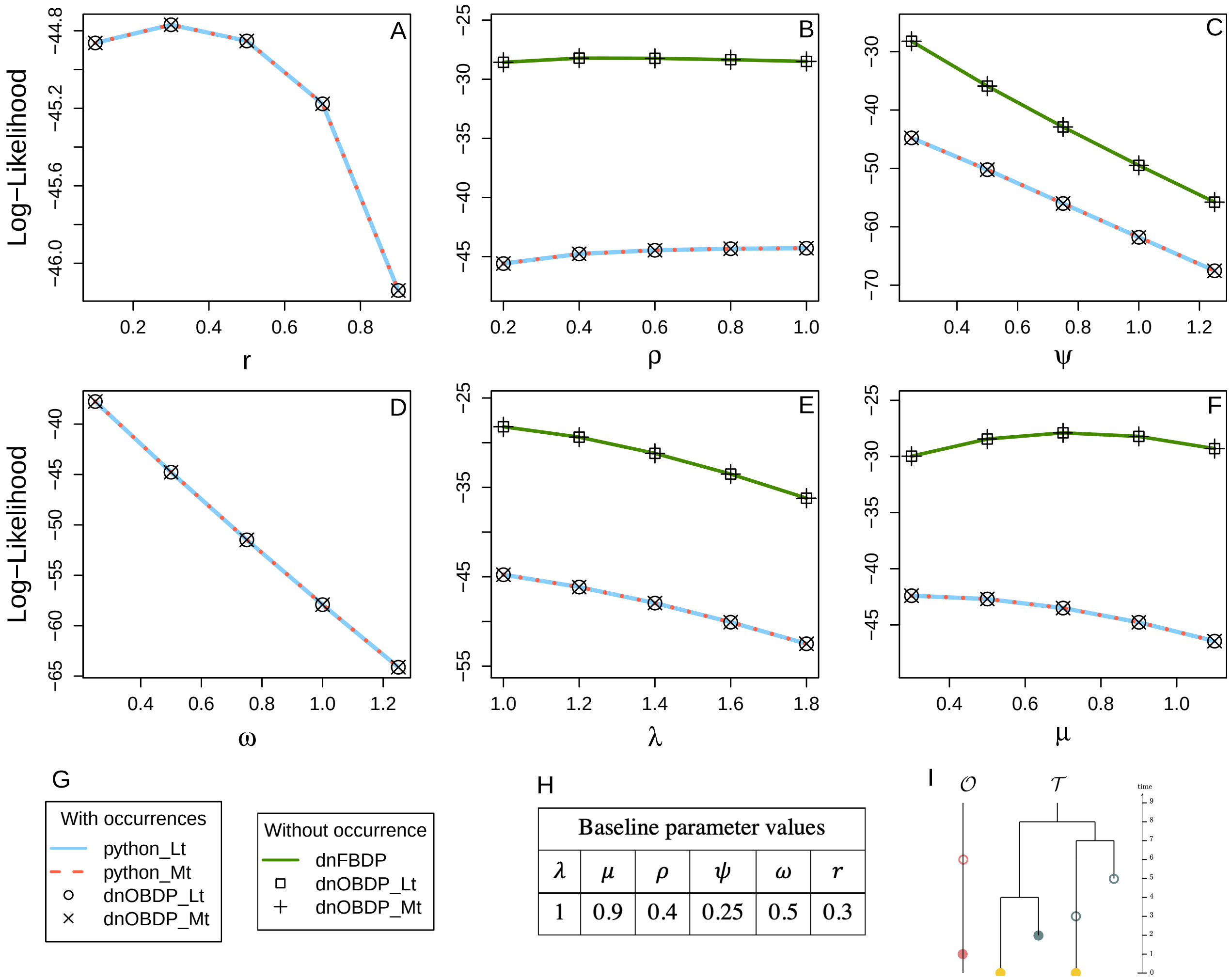
All prior distributions are fully detailed in Supp. Table S6.

3 Results

3.1 Validation of the method

3.1.1 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.



Baseline parameter values					
λ	μ	ρ	ψ	ω	r
1	0.9	0.4	0.25	0.5	0.3

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 dataset (I) where pink dots are occurrences, blue dots at past samples, and yellow dots are extant samples. 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. (2019) and whenever possible (B,C,E,F), to earlier FBDP implementations available in RevBayes, fixing $r = 0$, $\omega = 0$ and $\mathcal{O} = \emptyset$.

3.1.2 Qualitative Validation on simulated datasets

On Figure 4, we superimpose the true, simulated, trajectory of the total number of individuals, together with the inferred posterior distribution. Most importantly, the true trajectory falls within, or is very close to the boundaries, of the 95% posterior credibility interval at any point in time.

On both datasets, the topology of the tree was also well recovered but divergence dates do not always perfectly match (see Supp. Mat. Figure S9). Due to a greater amount of data in genetic sequences of both past and extant individuals, the divergence dates have been better inferred on dataset 2 as compared to dataset 1.

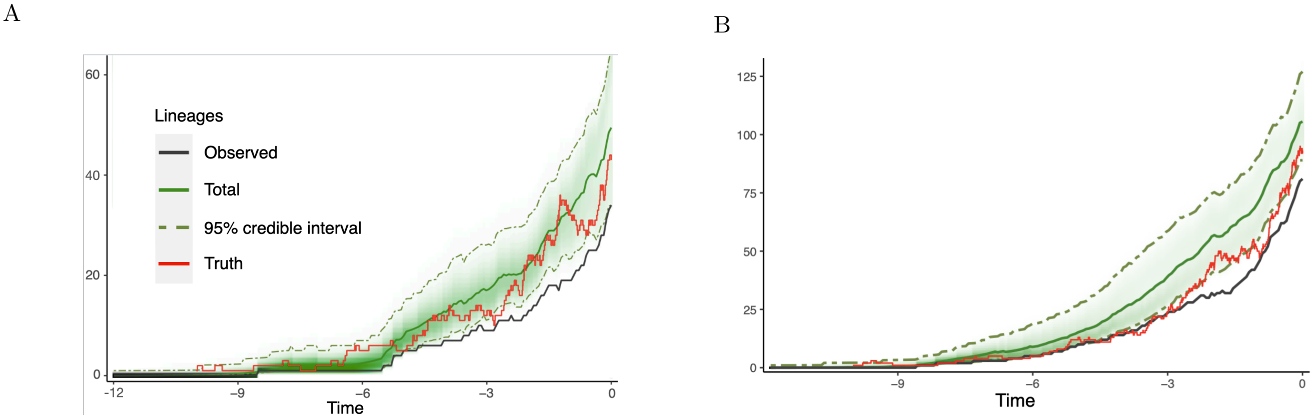


Figure 4: Results of the blind test analysis. The posterior distribution of the number of lineages through time is in green. The inferred LTT plot, showing the number of lineages in the tree through time, is in black. The true simulated number of individuals is in red.

3.1.3 Quantitative Validation of the Diversity Inference

Figure 5 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, i.e. samples adequately the targeted posterior distribution.

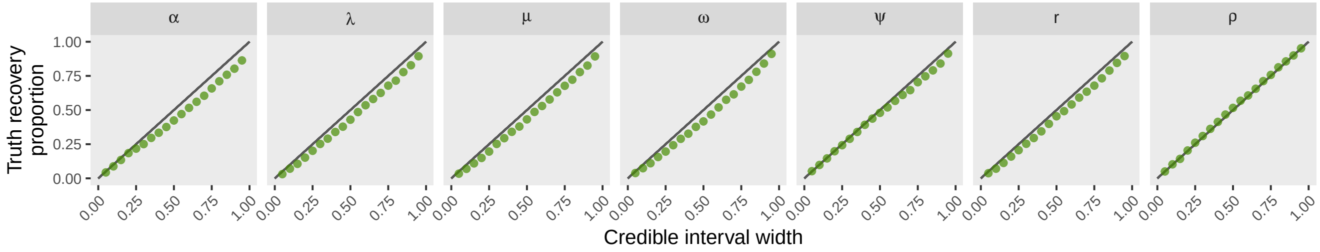


Figure 5: Results of the Simulation-Based Calibration. Each dot corresponds to the proportion of simulated parameters (y axis) falling within its inferred posterior credible interval with a given level (x axis). The black line corresponds to the expected perfect match.

3.2 Reproductive number and prevalence in the Covid outbreak

Figure 6 shows the raw data, as well as the estimates of the total prevalence and reproductive number through time. The instantaneous prevalence is typically always slightly lower than the total number of cases sampled each day, which corresponds to the sum of all sampled (and likely removed) individuals over a one-day period. 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 non-pharmaceutical interventions in early February (i.e. testing effort and cabin quarantine).

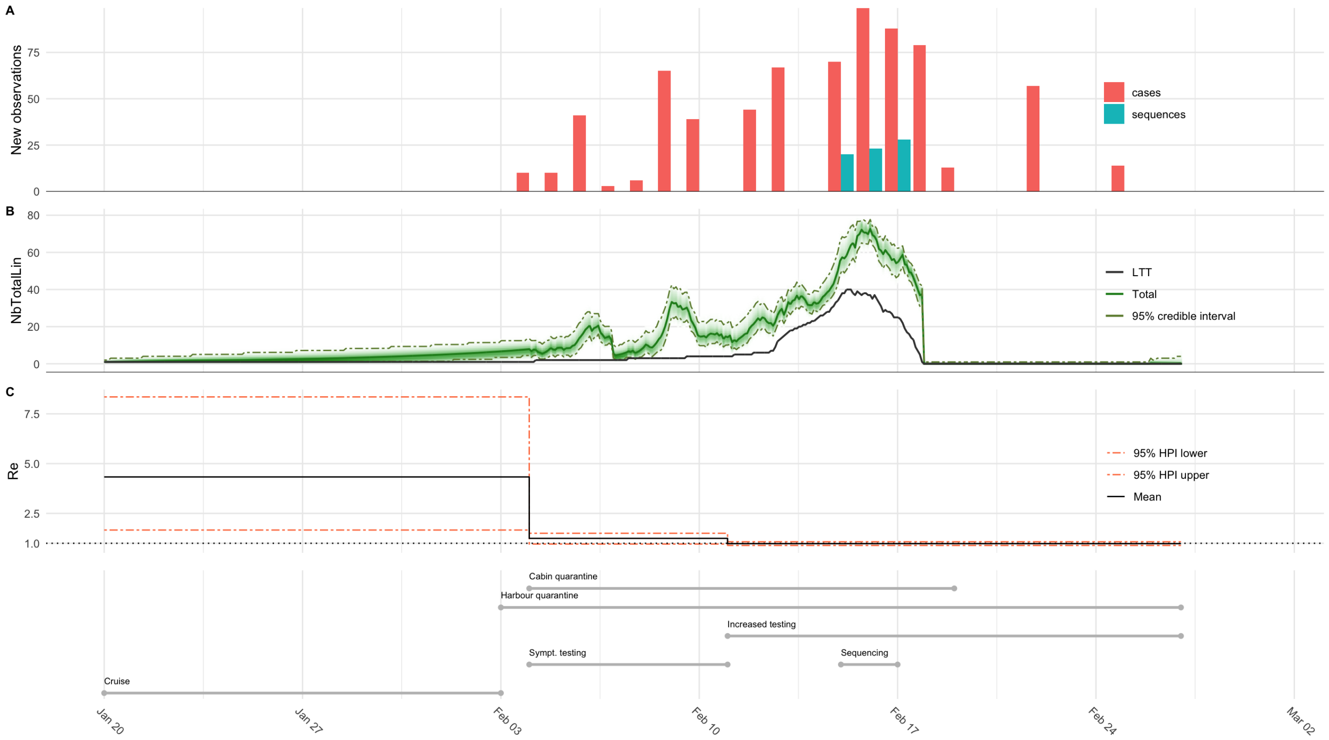


Figure 6: Analysis of the SARS-2 COVID-19 outbreak dynamics aboard the Diamond Princess cruise ship. (A) Occurrence and sequenced data are plotted as daily new observations. Although passengers were monitored until July 2020 (Ministry of Health and Welfare 2020), we focus on infections detected while guests were still aboard, until the end of the harbour quarantine. (B) Posterior probability distribution of the instantaneous total infected population aboard the cruise ship and inferred LTT. (C) Estimates of the effective reproductive number (R_e) throughout a 38 day period starting at the beginning of the cruise.

3.3 Total diversity in the Cetacean clade

Figure 7 shows the inferred diversity of cetacean genera over the past 50 Myr. 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 (3.5 Ma) with almost 70 inferred genera. The last million years correspond to a sharp decline leading to the 41 extant genera.

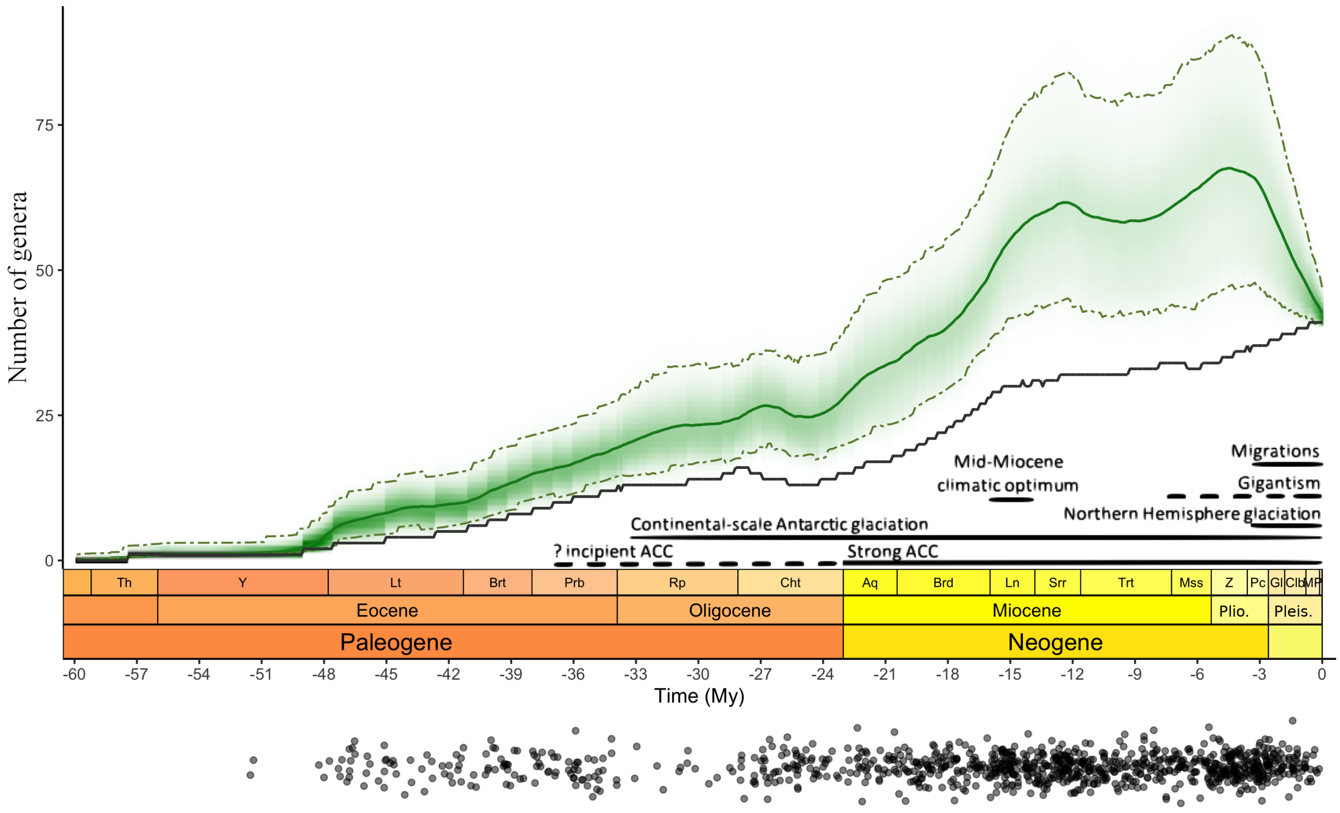


Figure 7: Posterior probability distribution of the total number of genera over time. The 95% credible intervals are indicated in dashed lines, the expected diversity is in green and the inferred LTT is in black. Periods of biotic or abiotic factors that are hypothesized to have driven diversification changes are adapted from Marx and Fordyce (2015) and shown in black for information but had no influence on the analysis. ACC = Antarctic Circumpolar Current. Black dots below represent the occurrences used in the analysis.

4 Discussion

4.1 Technical achievements and limitations

In this paper, we extend the work of Gupta et al. (2019) and Manceau et al. (2019) to consider piecewise-constant rates through time, and implement the Occurrence Birth-Death Process 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 population size through time, in a post-MCMC analysis (see workflow in Fig. 2). We validate the framework, both qualitatively and quantitatively, 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 on the results of Gupta et al. (2019). In practice, this assumption 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 Supp. Mat. A).

More work is thus needed to help speed up the likelihood computation when $r \neq 1$, on datasets 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.

4.2 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 *total evidence* approach for the inference of epidemiological trajectories.

In fact, encouraging conclusions can be drawn from both our parameter estimates and the corresponding trajectory inference. First, the reproductive number is inferred to be 4.33 in the absence of any intervention and detection, during the first 15 days of the cruise (see Fig. 6B). These values are consistent with previous studies estimating the basic reproductive number R_0 between 2.5 and 3.5 for the global pandemic (Stadler 2020a), and 3.53 for the Diamond Princess outbreak before quarantine (Stadler 2020b). Second, we infer a decrease in the reproductive number, that remains near or below 1 in the last 23 days of the time period of interest, suggesting that the epidemic was contained by measures taken. Interestingly, we note that this decrease is driven by the sampling and removal of individuals from the infectious population, with the total sampling rate going up from 1.41 to 1.79 days^{-1}, after February 11th (see Figure S13 for detailed timeline). This extended sampling, stemming partly from the decision to test asymptomatic passengers, results in occurrences having a visible impact on the reconstructed prevalence curve. This underlines 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. Due to a 24 hours reporting delay for this dataset, the first detected case was, for example, originally placed after the start of the quarantine. We tried to meticulously remove as many such biases as possible, but our framework could be improved in the future to 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 sub-populations: 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.

4.3 Past cetacean diversity

Molecular and paleontological data come with their inherent limitations, e.g. 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 endeavoured to respect these constraints, by (i) correcting occurrence distribution sampling biases, and (ii) making the most of the piecewise-constant parameter framework to include shifts in diversification rates, as detected previously (Rabosky 2014), as well as shifts in fossilization rates corresponding to the well-established low preservation rates in the Early Oligocene (Rupelian), Early Miocene (Aquitalian) and End Miocene (Messinian) (Marx et al. 2016).

The emerging patterns of cetacean generic diversification in Figure 7 are coherent 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 fuelled a diatom radiation, via nutrient supply, prompting the diversification of bulk filtering cetaceans. The diversity drop in the last 4 million years 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. 7) 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. Similar to applications in epidemiology, insights into macroevolution based on our novel framework will also benefit from developments that account for biases in fossil sampling, e.g. spatial and temporal biases (Close et al. 2020).

4.4 New avenues for phylogenetics

Over the last decade, the field of phylogenetics has expanded considerably with the development of the Fossilized Birth-Death Process and related extensions, of which the Occurrence Birth-Death Process 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 – e.g. turtles, sharks, 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). Birth-death process models also exist for the analysis of stratigraphic ranges, or paleontological data only (Stadler et al. 2018; Silvestro et al. 2019), further expanding the potential of phylogenetic models in quantitative paleobiology. 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 maximises the use of available evidence, since it uniquely allows us to combine genetic an 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, and is prospectively much more flexible than alternative approaches to mitigating sampling biases.

References

Aguirre-Fernández, G., R. Warnock, A. M. Benites-Palomino, J. Andréoletti, and M. Manceau. 2020. Cetacean timeline. 2.6.2

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. P. Natl. Acad. Sci. USA 106:13410–13414. 1

Allio, R., S. Donega, N. Galtier, and B. Nabholz. 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. Molecular Biology and Evolution 34:2762–2772 publisher: Oxford Academic. 2.6.3, S6

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. Warnock. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. Proceedings of the Royal Society B: Biological Sciences 286:20190685 publisher: Royal Society. 2.6.3

Churchill, M., J. H. Geisler, B. L. Beatty, and A. Goswami. 2018. Evolution of cranial telescoping in echolocating whales (cetacea: Odontoceti). Evolution 72:1092–1108. 2.6.2

Close, R., R. B. Benson, E. Saupe, M. Clapham, and R. Butler. 2020. The spatial structure of phanerozoic marine animal diversity. Science 368:420–424. 4.3

Dominici, S., S. Danise, S. Cau, and A. Freschi. 2020. The awkward record of fossil whales. Earth-Science Reviews . 2.6.2

Dong, E., H. Du, and L. Gardner. 2020. An interactive web-based dashboard to track covid-19 in real time. The Lancet infectious diseases 20:533–534. 2.5.1

Dornburg, A., M. C. Brandley, M. R. McGowen, and T. J. Near. 2012. Relaxed Clocks and Inferences of Heterogeneous Patterns of Nucleotide Substitution and Divergence Time Estimates across Whales and Dolphins (Mammalia: Cetacea). Molecular Biology and Evolution 29:721–736 publisher: Oxford Academic. 2.6.3, S6

Etienne, R. S., B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society B: Biological Sciences 279:1300–1309. 4.4

Fordyce, R. E. 2009. Cetacean Fossil Record. Pages 207–215 *in* Encyclopedia of Marine Mammals (Second Edition) (W. F. Perrin, B. Würsig, and J. G. M. Thewissen, eds.). Academic Press, London. 2.6.1

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. 2016. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic Biology 66:57–73. 1

Gavryushkina, A., D. Welch, T. Stadler, and A. J. Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS computational biology 10:e1003919. 1

Geisler, J. H. and A. E. Sanders. 2003. Morphological Evidence for the Phylogeny of Cetacea. Journal of Mammalian Evolution 10:23–129. 2.6.2

Gupta, A., M. Manceau, T. Vaughan, M. Khammash, and T. Stadler. 2019. The probability distribution of the reconstructed phylogenetic tree with occurrence data. bioRxiv Page 679365. 1, 4.1, 4.4

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. 2.5.1

He, X., E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, et al. 2020. Temporal dynamics in viral shedding and transmissibility of covid-19. Nature medicine 26:672–675. S7

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:2957–2966. 1

Heath, T. A., A. M. Wright, and W. Walker. 2019. RevBayes: Combined-Evidence Analysis and the Fossilized Birth-Death Process for Stratigraphic Range Data. 2.4.1, 2.6.3, S7, S6

Höhna, S. and T. Heath. 2019. RevBayes: Simple Diversification Rate Estimation. S6

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014. Probabilistic Graphical Model Representation in Phylogenetics. Systematic Biology 63:753–771. 2.3

Höhna, S., M. J. Landis, and T. A. Heath. 2017. Phylogenetic Inference Using RevBayes. Current Protocols in Bioinformatics 57:6.16.1–6.16.34 _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.22. 2.3, 2.4.1

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Systematic Biology 65:726–736. 1, 2.3

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

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56:701–710. 1

Mallapaty, S. 2020. What the cruise-ship outbreaks reveal about covid-19. Nature 580:18–18. 2.5.1

Manceau, M., A. Gupta, T. Vaughan, and T. Stadler. 2019. The probability distribution of ancestral population size under birth-death processes. bioRxiv Page 679365. 1, 2.1, 2.4.1, 3, 4.1, 4.4, A.1

Marshall, C. R. 2017. Five palaeobiological laws needed to understand the evolution of the living biota. Nature Ecology & Evolution 1:1–6. 1

Marx, F. G. and R. E. Fordyce. 2015. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. Royal Society Open Science 2:140434 publisher: Royal Society. 2.6.1, 7, 4.3

Marx, F. G., O. Lambert, and M. D. Uhen. 2016. Cetacean Paleobiology. John Wiley & Sons google-Books-ID: ADLOCwAAQBAJ. 2.6.1, 2.6.2, 4.3

McGowen, M. R., G. Tsagkogeorga, S. Álvarez Carretero, M. dos Reis, M. Struebig, R. Deaville, P. D. Jepson, S. Jarman, A. Polanowski, P. A. Morin, and S. J. Rossiter. 2020. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. Systematic Biology 69:479–501 publisher: Oxford Academic. S6

Ministry of Health, L. and Welfare. 2020. Situation of the covid-19 in the cruise ship diamond princess. 6, S7

Mizumoto, K., K. Kagaya, A. Zarebski, and G. Chowell. 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. 2.5.2

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011. Reconciling molecular phylogenies with the fossil record. P. Natl. Acad. Sci. USA 108:16327–16332. 1, 2.6.1, 4.3

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344:305–311. 1

Nishiura, H. 2020. Backcalculating the incidence of infection with covid-19 on the diamond princess. 4.2

O'Reilly, J. E. and P. C. Donoghue. 2020. The effect of fossil sampling on the estimation of divergence times with the fossilized birth–death process. Systematic biology 69:124–138. 1

Patzkowsky, M. E. and S. M. Holland. 2012. Stratigraphic paleobiology: understanding the distribution of fossil taxa in time and space. University of Chicago Press. 4.4

Quental, T. B. and C. R. Marshall. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution 25:434–441. 1, 2.6.1, 2.6.2

Rabosky, D. L. 2014. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. PLOS ONE 9:e89543 publisher: Public Library of Science. 4.3, S6

Rasmussen, D. A., O. Ratmann, and K. Koelle. 2011. Inference for nonlinear epidemiological models using genealogies and time series. PLoS computational biology 7:–1002136. 1

Raup, D. M. 1972. Taxonomic Diversity during the Phanerozoic. Science 177:1065–1071 publisher: American Association for the Advancement of Science. 1

Rocklöv, J., H. Sjödin, and A. Wilder-Smith. 2020. Covid-19 outbreak on the diamond princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. Journal of travel medicine 27:taaa030. 4.2

Scire, J., J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler. 2020. Improved multi-type birth-death phylodynamic inference in beast 2. bioRxiv . 4.4

Sepkoski, J. J., R. K. Bambach, D. M. Raup, and J. W. Valentine. 1981. Phanerozoic marine diversity and the fossil record. Nature 293:435–437 number: 5832 Publisher: Nature Publishing Group. 1

Shu, Y. and J. McCauley. 2017. Gisaid: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22:30494. 2.5.1

Silvestro, D., N. Salamin, A. Antonelli, and X. Meyer. 2019. Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. Paleobiology 45:546–570 publisher: Cambridge University Press. 1, 4.4

Silvestro, D., N. Salamin, and J. Schnitzler. 2014. PyRate: a new program to estimate speciation and extinction rates from incomplete fossil data. Methods in Ecology and Evolution 5:1126–1131 _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12263. 1

Silvestro, D., R. C. Warnock, A. Gavryushkina, and T. Stadler. 2018. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. Nature Communications 9:1–14. 1

Stadler, T. 2010. Sampling-through-time in birth–death trees. Journal of theoretical biology 267:396–404. 1

Stadler, T. 2020a. Phylodynamic analyses based on 128 sequences. 4.2, S7

Stadler, T. 2020b. Phylodynamic analyses of outbreaks in china, italy, washington state (usa), and the diamond princess. 4.2

Stadler, T., A. Gavryushkina, R. C. Warnock, A. J. Drummond, and T. A. Heath. 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. Journal of theoretical biology 447:41–55. 1, 4.4

Stadler, T., D. Kühnert, S. Bonhoeffer, and A. J. Drummond. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (HCV). Proceedings of the National Academy of Sciences 110:228–233. 1

Steeman, M. E., M. B. Hebsgaard, R. E. Fordyce, S. Y. Ho, D. L. Rabosky, R. Nielsen, C. Rahbek, H. Glenner, M. V. Sørensen, and E. Willerslev. 2009. Radiation of extant cetaceans driven by restructuring of the oceans. Systematic biology 58:573–585. 2.6.2

Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. 2018. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat] ArXiv: 1804.06788. 2.4.3

Uhen, M. and N. Pyenson. 2007. Diversity estimates, biases, and historiographic effects: Resolving cetacean diversity in the Tertiary. Palaeontologia Electronica 10. 2.6.1, 2.6.2, 4.3

Vaughan, T. G., G. E. Leventhal, D. A. Rasmussen, A. J. Drummond, D. Welch, and T. Stadler. 2019. Estimating epidemic incidence and prevalence from genomic data. Molecular Biology and Evolution . 1

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer google-Books-ID: XgFkDAAAQBAJ. B.2

Wright, A. M. 2019. A systematist's guide to estimating bayesian phylogenies from morphological data. Insect Systematics and Diversity 3:2. 2.6.3

Wright, A. M. 2020. RevBayes: Discrete morphology - Multistate Characters. S6

Xing, Y., R. E. Onstein, R. J. Carter, T. Stadler, and H. Peter Linder. 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. 1

Yule, G. U. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. R. Soc. Lond. B . 1

Zhang, C., T. Stadler, S. Klopfstein, T. A. Heath, and F. Ronquist. 2015. Total-evidence dating under the fossilized birth–death process. Systematic biology 65:228–249. 1

– Appendix –
A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences

This appendix presents the detailed derivation of the model used in "A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences" by Andréoletti, Zwaans et al, as well as supplementary results and figures. We extend results of Gupta et al. (2019) and Manceau et al. (2019) to piecewise-constant parameters, describe our implementation in the RevBayes software, and give detailed information on all priors used for simulation or inference in our analyses.

A Method extension to piecewise-constant parameters **2**

A.1 Notation and outline of the general strategy . . . 2

A.2 Temporal setup for piecewise constant parameters . . . 3

A.3 Master equations governing L_t and M_t . . . 4

A.4 Updates at punctual events . . . 5

A.5 Numerical approximation of the ODEs . . . 6

B RevBayes implementation **9**

B.1 Core algorithms . . . 9

B.2 RevGadgets . . . 11

C Qualitative validation: "blind test" on simulated data **13**

D Macroevolution application: Inferring past cetacean diversity **17**

D.1 Preliminary analysis of the cetacean occurrence fossil record . . . 17

D.2 Detailed priors used for Bayesian inference . . . 17

E Epidemiology application: the Diamond Princess SARS-2 COVID-19 outbreak dynamics **19**

E.1 Data acquisition on GISAID . . . 19

E.2 Pre-processing the data . . . 19

E.3 Detailed priors . . . 22

A Method extension to piecewise-constant parameters

A.1 Notation and outline of the general strategy

We first recall in Figure S1 the notation that we introduced in the main text with the three different sampling (ψ-sampling for sampling of fossils with inclusion in the tree, ω-sampling for occurrences and ρ-sampling at present).

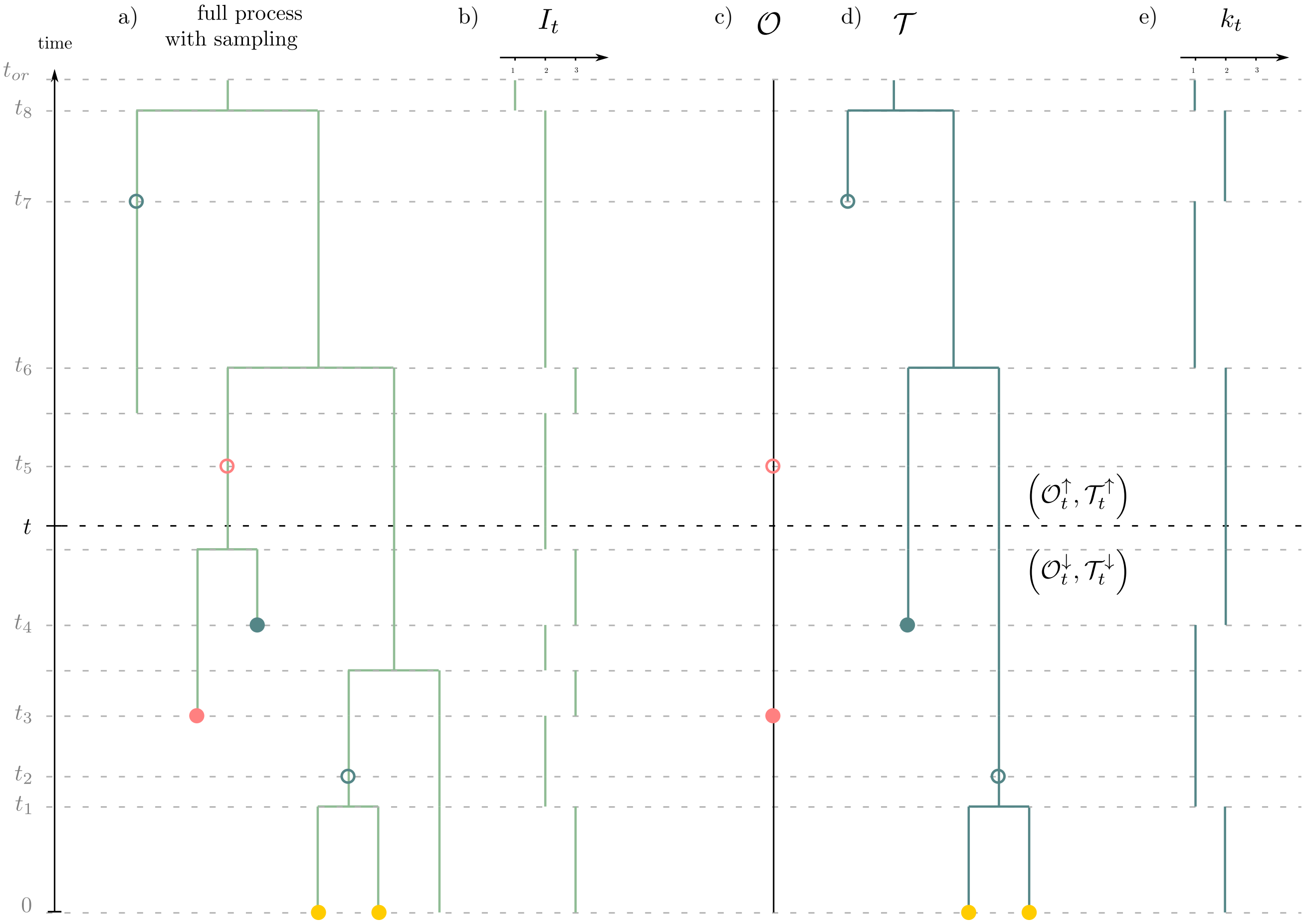


Figure S1: General setting of the method. a) the full process with sampling. Pink dots correspond to ω-sampling (sampling through time without sequencing), blue dots correspond to ψ-sampling (sampling through time with sequencing) and yellow dots correspond to ρ-sampling at present. Filled or unfilled dots correspond respectively to sampling with or without removal. b) Total number of individuals through time. c) Record of occurrences. d) Reconstructed tree spanning ψ- and ρ-samples. e) Number of lineages through time in the reconstructed tree (i.e. LTT plot).

To compute the likelihood of $(\mathcal{T}, \mathcal{O})$ under this process, we will slice horizontally our observations and perform a breadth-first traversal of these. We thus introduce now,

$$\begin{aligned}
\mathcal{T}_t^\uparrow &:= \text{the tree } \mathcal{T} \text{ cut at time } t \\
\mathcal{T}_t^\downarrow &:= \text{the collection of trees (or forest) obtained by cutting } \mathcal{T} \\
&\quad\ \text{at time } t\text{, and considering all subtrees descending from cut lineages} \\
k_t &:= \text{number of sampled lineages in } \mathcal{T} \text{ at time } t \\
\mathcal{O}_t^\uparrow &:= \mathcal{O}_{|(t,+\infty)} \\
\mathcal{O}_t^\downarrow &:= \mathcal{O}_{|(0,t)}
\end{aligned}$$

We can now recall the definition of our two key probability densities,

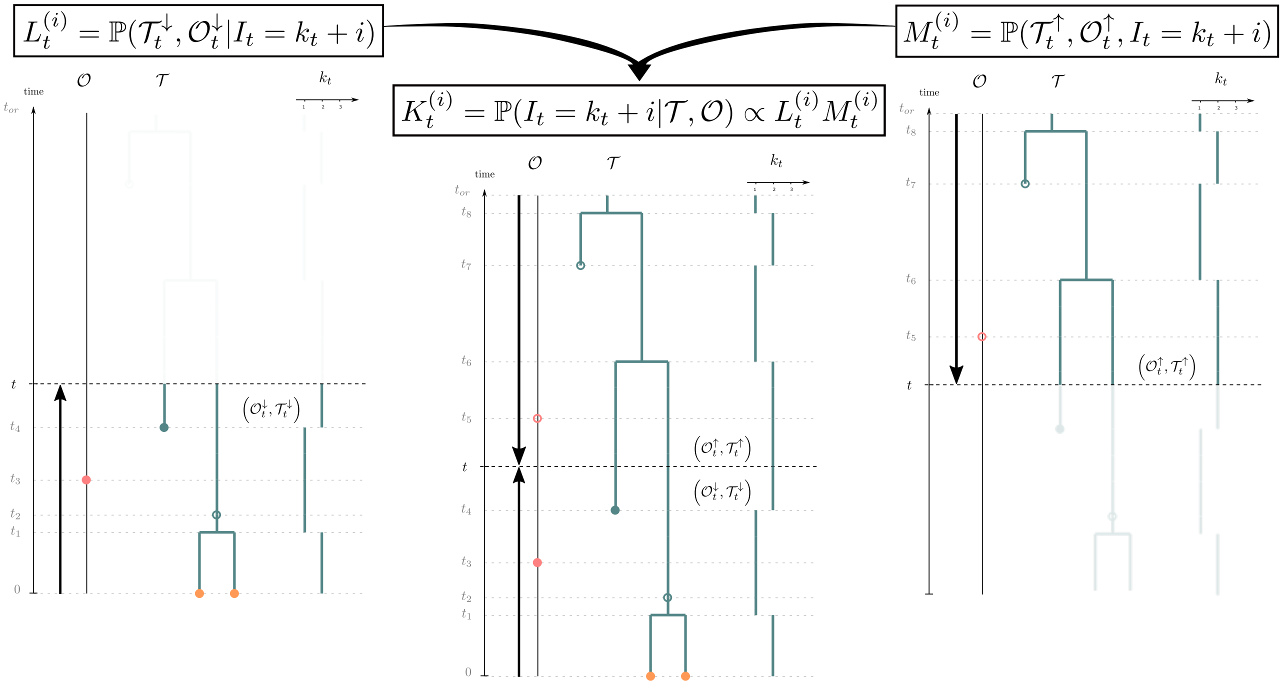
$$\forall i \in \mathbb{N}, \quad L_t^{(i)} := \mathbb{P}(T_t^{\downarrow}, \mathcal{O}_t^{\downarrow} \mid I_t = k_t + i) \tag{S1}$$

$$\forall i \in \mathbb{N}, \quad M_t^{(i)} := \mathbb{P}(T_t^{\uparrow}, \mathcal{O}_t^{\uparrow}, I_t = k_t + i) \tag{S2}$$

These probability densities have been introduced in Manceau et al. (2019) as a way to target the probability distribution K_t of the total number of individuals given the data. Indeed,

$$\begin{aligned} K_t^{(i)} &:= \mathbb{P}(I_t = k_t + i \mid \mathcal{T}, \mathcal{O}) \\ &\propto \mathbb{P}(I_t = k_t + i, T_t^{\uparrow}, \mathcal{O}_t^{\uparrow}, T_t^{\downarrow}, \mathcal{O}_t^{\downarrow}) \\ &\propto \mathbb{P}(T_t^{\downarrow}, \mathcal{O}_t^{\downarrow} \mid I_t = k_t + i, T_t^{\uparrow}, \mathcal{O}_t^{\uparrow}) \mathbb{P}(I_t = k_t + i, T_t^{\uparrow}, \mathcal{O}_t^{\uparrow}) \\ &\propto L_t^{(i)} M_t^{(i)} \end{aligned} \tag{S3}$$

The general strategy of the methods consists of (i) traversing the data backward in time to compute L_t; (ii) traversing the data forward in time to compute M_t; (iii) using the results to compute K_t. This scheme is illustrated in Figure S2.



$L_t^{(i)} = \mathbb{P}(\mathcal{T}_t^{\downarrow}, \mathcal{O}_t^{\downarrow} | I_t = k_t + i)$
$K_t^{(i)} = \mathbb{P}(I_t = k_t + i | \mathcal{T}, \mathcal{O}) \propto L_t^{(i)} M_t^{(i)}$
$M_t^{(i)} = \mathbb{P}(\mathcal{T}_t^{\uparrow}, \mathcal{O}_t^{\uparrow}, I_t = k_t + i)$
time
$\mathcal{O}$
$\mathcal{T}$
k_t
1 2 3
t_{or}
t_8
t_7
t_6
t_5
t
t_4
t_3
t_2
t_1
0
$(\mathcal{O}_t^{\downarrow}, \mathcal{T}_t^{\downarrow})$
$(\mathcal{O}_t^{\uparrow}, \mathcal{T}_t^{\uparrow})$

Figure S2: Inferring the posterior distribution of the number of individuals (K_t) in the OBDP. The probability distribution of the past number of lineages K_t is obtained at each time t by combining the quantity L_t obtained from the backward traversal algorithm (left) and the quantity M_t obtained from the forward traversal algorithm (right). See Table 1 for notations.

In the rest of this appendix section, we present the Master equations governing the evolution of these densities through time in a setup with piecewise-constant parameters.

A.2 Temporal setup for piecewise constant parameters

We partition time into two distinct units.
First, we define periods of time with no observations or sampling events, coined *epochs*, which allow for the basic derivation of Master equations of L_t and M_t. Epochs are delimited by all n *punctual events* times (i.e. branching

and sampling events) in $\mathcal{O}$ and $\mathcal{T}$ pooled in an ordered list $(t_h)_{h=1}^{n}$. Epoch h is thus defined as the time interval (t_h, t_{h+1}).
Second, we account for all rate shift events, which define *constant rate time intervals*. If we have m such intervals, we pool all $m+1$ rate shift events in an ordered list $(\tau_l)_{l=0}^{m+1}$, where by convention we consider that $\tau_0 = 0$ and $\tau_{m+1} = t_{or}$. Rate time interval l is defined as (τ_l, τ_{l+1}), with parameter set $(\lambda_l, \mu_l, \psi_l, \omega_l, r_l)$. We illustrate this setup in Figure S3 below.

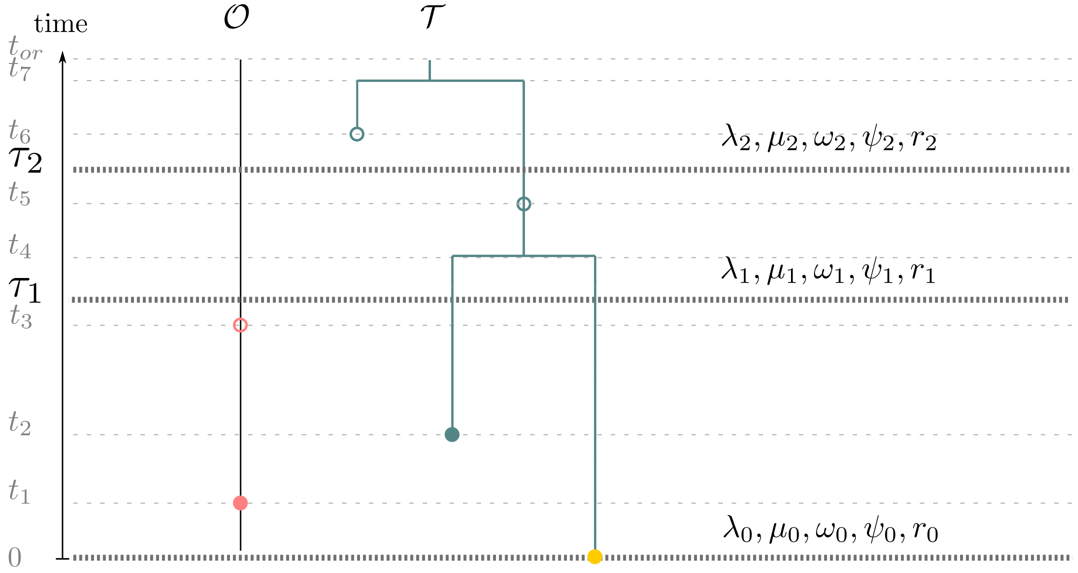

time
$\mathcal{O}$
$\mathcal{T}$
t_{or}
t_7
t_6
τ_2
t_5
t_4
τ_1
t_3
t_2
t_1
0
$\lambda_2, \mu_2, \omega_2, \psi_2, r_2$
$\lambda_1, \mu_1, \omega_1, \psi_1, r_1$
$\lambda_0, \mu_0, \omega_0, \psi_0, r_0$

Figure S3: Temporal setup of the method.

A.3 Master equations governing L_t and M_t

Probability densities L_t and M_t satisfy different Master equations obtained by studying their evolution through time along any given epoch. These are ordinary differential equations (ODE) that can be approximated numerically. Here, we assume $\tau_l \le t < \tau_{l+1}$ meaning that parameters have values $(\lambda_l, \mu_l, \psi_l, \omega_l, r_l)$.

First, we can initialize L_t and M_t respectively at present time 0 and at the time of origin t_{or}. At present, ρ sampling of extant tips yields,

$$\forall i \in \mathbb{N}, \quad L_0^{(i)} = \rho^{k_0}(1-\rho)^i \tag{S4}$$

while at the time of origin, the process starts with only one individual $k_{t_{or}} = 1$, which yields,

$$\forall i \in \mathbb{N}, \quad M_{t_{or}}^{(i)} = \mathbb{P}(I_{t_{or}} = 1 + i) = \mathbb{1}_{i=0} \tag{S5}$$

We now consider all events happening in an infinitesimal time step δt in the full underlying process which do not result in observations or samplings. Three scenarios correspond to this case:

1. nothing happened with probability $(1 - \gamma_l(k+i)\delta t)$, where $\gamma_l = \lambda_l + \mu_l + \psi_l + \omega_l$
2. a birth event happened :
   (a) among the k sampled lineages in $T_t^{\downarrow}$, and it leads to an extinct or unsampled subtree to the left or to the right with probability $2\lambda_l k \delta t$
   (b) among the i other individuals with probability $\lambda_l i \delta t$.
3. a death event happened among the i particles, with probability $\mu_l i \delta t$

We combine these to write, $\forall i \in \mathbb{N}$,

$$L^{(i)}_{t+\delta_t} = (1 - \gamma_l(k+i)\delta t)L^{(i)}_t + \lambda_l(2k+i)\delta t)L^{(i+1)}_t + \mu_l i\delta t)L^{(i-1)}_t \tag{S6}$$

Letting $\delta t \to 0$ yields the following differential equation for L_t,

$$\forall i \in \mathbb{N},\ \ L^{(i)}_0 = \rho^{k_0}(1-\rho)^i \tag{S7}$$

$$\dot{L}^{(i)}_t = -\gamma_l(k+i)L^{(i)}_t + \lambda_l(2k+i)L^{(i+1)}_t + \mu_l i L^{(i-1)}_t \tag{S8}$$

Similarly, M_t is the solution of the following ODE,

$$\forall i \in \mathbb{N},\ \ M^{(i)}_{t_{or}} = \mathbb{P}(I_{t_{or}} = 1+i) = \mathbb{1}_{i=0} \tag{S9}$$

$$\dot{M}^{(i)}_t = -\gamma_l(k+i)M^{(i)}_t + \lambda_l(2k+i-1)M^{(i-1)}_t + \mu_l(i+1)M^{(i+1)}_t \tag{S10}$$

A.4 Updates at punctual events

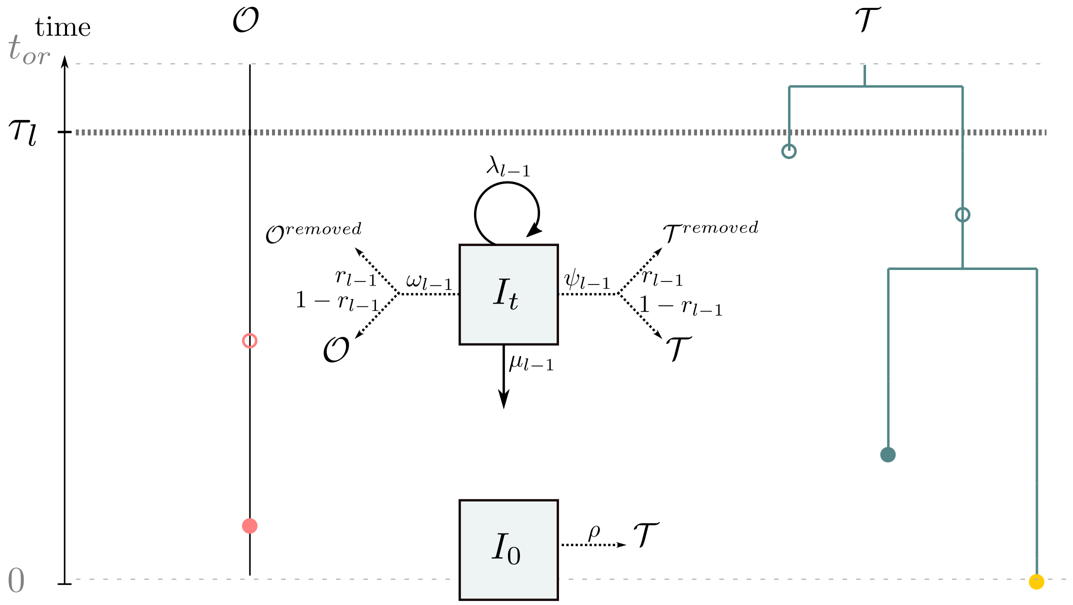


Figure S4: Updated sampling scheme of the method.

There are 6 types of punctual events in $\mathcal{T}$ and $\mathcal{O}$ that affect the probability densities M_t and L_t. These correspond to all different sampling options along $\mathcal{T}$ and $\mathcal{O}$ as illustrated in Figure S4. We denote as M_{t^-} and L_{t^-} the probability densities immediately prior to the event and M_{t^+} and L_{t^+} immediately after each event. We emphasise that the expressions differ when considering the process forward in time for M_t or backward in time, for L_t. These cases are the following :

1. sampling of a leaf:
   (a) in $\mathcal{T}^{\downarrow}_t$, $L^{(i)}_{t^+} = \psi_l(1-r_l)L^{(i+1)}_{t^-}$
   (b) in $\mathcal{T}^{\uparrow}_t$, $M^{(i)}_{t^-} = \psi_l(1-r_l)M^{(i-1)}_{t^+}$
2. removed sampled leaf:
   (a) in $\mathcal{T}^{\downarrow}_t$, $L^{(i)}_{t^+} = \psi_l r_l L^{(i)}_{t^-}$
   (b) in $\mathcal{T}^{\uparrow}_t$, $M^{(i)}_{t^-} = \psi_l r_l M^{(i)}_{t^+}$

3. sampling along a branch:
   (a) in $\mathcal{T}_t^{\downarrow}$, $L_{t^+}^{(i)} = \psi_l(1-r_l)L_{t^-}^{(i)}$
   (b) in $\mathcal{T}_t^{\uparrow}$, $M_{t^-}^{(i)} = \psi_l(1-r_l)M_{t^+}^{(i)}$
4. occurrence:
   (a) in $\mathcal{O}_t^{\downarrow}$, $L_{t^+}^{(i)} = (k+i)\omega_l(1-r_l)L_{t^-}^{(i)}$
   (b) in $\mathcal{O}_t^{\uparrow}$, $M_{t^-}^{(i)} = (k+i)\omega_l(1-r_l)M_{t^+}^{(i)}$
5. removed occurrence:
   (a) in $\mathcal{O}_t^{\downarrow}$, $L_{t^+}^{(i)} = \omega_l r_l i L_{t^-}^{(i-1)}$
   (b) in $\mathcal{O}_t^{\uparrow}$, $M_{t^-}^{(i)} = \omega_l r_l (i+1) M_{t^+}^{(i+1)}$
6. branching event:
   (a) in $\mathcal{T}_t^{\downarrow}$, $L_{t^+}^{(i)} = \lambda_l L_{t^-}^{(i)}$
   (b) in $\mathcal{T}_t^{\uparrow}$, $M_{t^-}^{(i)} = \lambda_l M_{t^+}^{(i)}$

A.5 Numerical approximation of the ODEs

As described in A.3, for any constant rate time interval where $\tau_l \leq t < \tau_{l+1}$, M_t and L_t are defined along epochs as the solution to systems of differential equations S8 and S10 for $t_h \leq t < t_{h+1}$. Numerically, the solution to such systems of equations is approximated by truncating the system at a fixed integer N as follows:

$$L_{t_{h+1}} = e^{A_l(t-t_h)}L_{t_h} \tag{S11}$$

$$M_{t_h} = e^{A_l^{'}(t-t_{h+1})}M_{t_{h+1}} \tag{S12}$$

Where A_l and A_l' are $N \times N$ tridiagonal matrices with ODE coefficients. When there is a rate shift τ_l within an epoch (t_h, t_{h+1}), the epoch is cut in two parts and L_t and M_t are simply computed as,

$$L_{t_{h+1}} = e^{A_{l+1}(t_{h+1}-\tau_l)}e^{A_l(\tau_l-t_h)}L_{t_h} \tag{S13}$$

$$M_{t_h} = e^{A_l'(t_h-\tau_l)}e^{A_{l+1}^{'}(\tau_l-t_{h+1})}M_{t_{h+1}} \tag{S14}$$

This can be extended to any number of rate changes within an epoch. This strategy of solving for L_t and M_t yields the following two algorithms. Because exponential matrices are computationally intensive to calculate, these algorithms are only used in the most general cases, when no other analytical formula is available (i.e. when $\omega \neq 0$ and $r \neq 1$).

Algorithm 1 Computes a numerical approximation of L_t for a specific set of times with known rate shift events

```
Input:
    Observed tree and occurrence data $(\mathcal{T}, \mathcal{O})$,
    extant sampling probability $\rho$,
    set of times of rate shift events $(\tau_l)_{l=0}^{m+1}$,
    and corresponding sets of parameters :
    vector $\lambda = (\lambda_l)_{l=0}^{m}$ where $\lambda_l$ is the birth rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\mu = (\mu_l)_{l=0}^{m}$ where $\mu_l$ is the death rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\psi = (\psi_l)_{l=0}^{m}$ where $\psi_l$ is the sampling rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\omega = (\omega_l)_{l=0}^{m}$ where $\omega_l$ is the rate of occurence sampling in time interval $[\tau_l, \tau_{l+1})$
    vector $r = (r_l)_{l=0}^{m}$ where $r_l$ is the removal probability in time interval $[\tau_l, \tau_{l+1})$
    set of time points $(d_j)_{j=1}^{S}$ for which we want to compute the density, and
    the truncation $N$ setting the accuracy of the algorithm.
Output: A numerical approximation of $L_t$ at times $(d_j)_{j=1}^{S}$, $(\widetilde{L}_t^{(i)})_{\substack{i\in\{0,1,\ldots,N\}\\ j\in\{1,2,\ldots,S\}}}$.
Pool all $(d_j)_{j=1}^{S}$, all branching and sampling times of $(\mathcal{T}, \mathcal{O})$ and rate shift times $(\tau_l)_{l=0}^{m+1}$ in an ordered list $(t_h)_{h=1}^{n+m+1}$.
Set $j = 1$ and initialize $B$ as a $S \times N + 1$ empty matrix.
Set $l = 0$ and $\lambda = \lambda_0$, $\mu = \mu_0$, $\psi = \psi_0$, $\omega = \omega_0$, $r = r_0$, $\gamma_0 = \lambda_0 + \mu_0 + \psi_0 + \omega_0$.
Set $\forall i \in \{0, 1, \ldots, N\}$, $\widetilde{L}_0^{(i)} = \rho^{k_0}(1-\rho)^i$.
for  $h = 1, 2, \ldots, n + m + 1$  do
    Numerically solve the ODE $\dot{\widetilde{L}}_t = A\widetilde{L}_t$ on $(t_h, t_{h+1})$, where matrix $A$ is a $N \times N$ tridiagonal matrix with entries given by,
        $$\forall i \in \{0,1,\ldots,N\} \quad A^{(i,i)} = \gamma(k+i)$$
        $$\forall i \in \{0,1,\ldots,N-1\} \quad A^{(i,i+1)} = \lambda(2k+i)$$
        $$\forall i \in \{1,2,\ldots,N\} \quad A^{(i,i-1)} = \mu i$$
    if $t_h = d_j$ then
        Set $B^{(j,i)} = \widetilde{L}_{t_h}^{(i)}$ and
        Set $j = j + 1$.
    end if
    if $t_h = t_{or}$ or $t_h = d_S$ then
        return $B$
    else if $t_h = \tau_l$ then
        Set $\lambda = \lambda_l$, $\mu = \mu_l$, $\psi = \psi_l$, $\omega = \omega_l$, $r = r_l$, $\gamma_l = \lambda_l + \mu_l + \psi_l + \omega_l$
        Set $l = l + 1$
    else if $t_h$ is a removed leaf then
        Set $\widetilde{L}_{t_h^+} = \psi r \widetilde{L}_{t_h^-}$
    else if $t_h$ is a non-removed leaf then
        Set $\forall i < N, \widetilde{L}_{t_h^+}^{(i)} = \psi(1-r)\widetilde{L}_{t_h^-}^{(i+1)}$ and $\widetilde{L}_{t_h^+}^{(N)} = 0$
    else if $t_h$ is a sampled ancestor then
        Set $\widetilde{L}_{t_h^+} = \psi(1-r)\widetilde{L}_{t_h^-}$
    else if $t_h$ is a removed occurrence then
        Set $\forall i > 0, \widetilde{L}_{t_h^+}^{(i)} = \omega r i \widetilde{L}_{t_h^-}^{(i-1)}$ and $\widetilde{L}_{t_h^-}^{(0)} = 0$.
    else if $t_h$ is a non-removed occurrence then
        Set $\widetilde{L}_{t_h^+}^{(i)} = \omega(1-r)(k+i)\widetilde{L}_{t_h^-}^{(i)}$
    else $t_h$ is a branching event
        Set$\widetilde{}_{t_h^+} = \lambda\widetilde{L}_{t_h^-}$
    end if
end for
```

Algorithm 2 Computes a numerical approximation of M_t for a specific set of times with known rate shift events

```
Input:
    Observed tree and occurrence data $(\mathcal{T}, \mathcal{O})$,
    parameters $t_{or}, \rho$
    set of times of rate shift events $(\tau_l)_{l=0}^{m+1}$,
    and corresponding sets of parameters :
    vector $\lambda = (\lambda_l)_{l=0}^{m}$ where $\lambda_l$ is the birth rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\mu = (\mu_l)_{l=0}^{m}$ where $\mu_l$ is the death rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\psi = (\psi_l)_{l=0}^{m}$ where $\psi_l$ is the sampling rate in time interval $[\tau_l, \tau_{l+1})$
    vector $\omega = (\omega_l)_{l=0}^{m}$ where $\omega_l$ is the rate of occurence sampling in time interval $[\tau_l, \tau_{l+1})$
    vector $r = (r_l)_{l=0}^{m}$ where $r_l$ is the removal rate in time interval $[\tau_l, \tau_{l+1})$
    set of time points $(d_j)_{j=1}^{S}$ for which we want to compute the density,
    and the truncation $N$ setting the accuracy of the algorithm.
Output: A numerical approximation of $M_t$ at times $(d_j)_{j=1}^{S}$, $(\widetilde{M}_t^{(i)})_{i \in \{0,1,\dots,N-1\}, j \in \{1,2,\dots,S\}}$.
Pool all $(d_j)$, rate shift times $(\tau_l)$ and all branching and sampling times of $(\mathcal{T}, \mathcal{O})$ in an ordered list $(t_h)_{h=1}^{n}$.
Set $j = S$, $k = m$ and $B'$ as a $S \times N$ empty matrix.
Set $\forall i \in \{0, 1, \dots, N-1\}$, $\widetilde{M}_{t_n}^{(i)} = \mathbb{1}_{i=0}$.
Set $l = m$ and $\lambda = \lambda_m$, $\mu = \mu_m$, $\psi = \psi_m$, $\omega = \omega_m$, $r = r_m$.
for $h = n-1, n-2, \dots, 0$ do
    Numerically solve the ODE $\dot{\widetilde{M}}_t = A'\widetilde{M}_t$ on $(t_h, t_{h+1})$, where matrix $A'$ is a $N \times N$ tridiagonal matrix with entries given by,
        $\forall i \in \{0, 1, \dots, N-1\} \quad A'^{(i,i)} = \gamma(k+i)$
        $\forall i \in \{0, 1, \dots, N-2\} \quad A'^{(i,i+1)} = -\mu(i+1)$
        $\forall i \in \{1, 2, \dots, N-1\} \quad A'^{(i,i-1)} = -\lambda(2k+i-1)$
    if $t_h = \tau_j$ then
        Set $B'^{(j,i)} = \widetilde{M}_{t_h}^{(i)}$ and $j = j-1$.
    end if
    if $t_h = 0$ or $t_h = \tau_S$ then
        return $B'$
    else if $t_h = \tau_l$ then
        Set $\lambda = \lambda_k$, $\mu = \mu_k$, $\psi = \psi_k$, $\omega = \omega_k$, $r = r_k$, $\gamma_l = \lambda_l + \mu_l + \psi_l + \omega_l$
        Set $l = l-1$
    else if $t_h$ is a removed leaf then
        Set $\widetilde{M}_{t_h^-} = \psi r \widetilde{M}_{t_h^+}$
    else if $t_h$ is a non-removed leaf then
        Set $\forall i \in \{1, 2, \dots, N-1\}, \widetilde{M}_{t_h^-}^{(i)} = \psi(1-r)\widetilde{M}_{t_h^+}^{(i-1)}$ and $\widetilde{M}_{t_h^-}^{(0)} = 0$
    else if $t_h$ is a sampled ancestor then
        Set $\widetilde{M}_{t_h^-} = \psi(1-r)\widetilde{M}_{t_h^+}$
    else if $t_h$ is a removed occurrence then
        Set $\forall i \in \{0, 1, \dots, N-2\}, \widetilde{M}_{t_h^-}^{(i)} = \omega r(i+1)\widetilde{M}_{t_h^+}^{(i+1)}$ and $\widetilde{M}_{t_h^-}^{(N-1)} = 0$.
    else if $t_h$ is a non-removed occurrence then
        Set $\widetilde{M}_{t_h^-}^{(i)} = \omega(1-r)(k+i)\widetilde{M}_{t_h^+}^{(i)}$
    else $t_h$ is a branching event
        Set $\widetilde{M}_{t_h^-} = \lambda \widetilde{M}_{t_h^+}$
    end if
end for
```

B RevBayes implementation

B.1 Core algorithms

To enable great flexibility and ensure fast computation, RevBayes is constructed around a mirror structure (Figure 5) in which all the core functions coded in C++ are reflected in the revlanguage section that links with the Rev language interface.

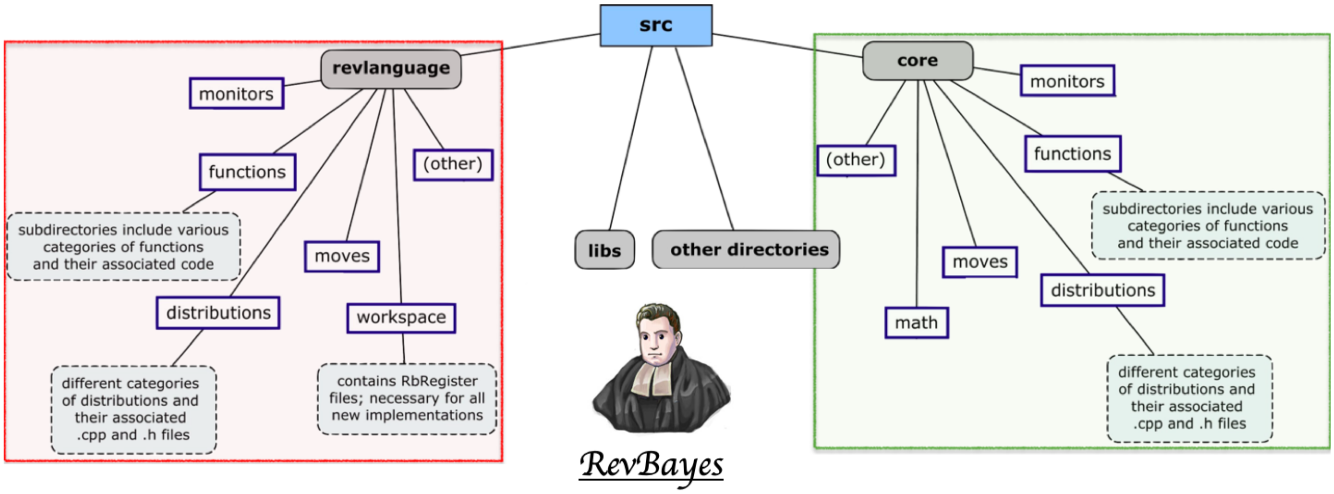

src
revlanguage
monitors
functions
(other)
subdirectories include various categories of functions and their associated code
moves
distributions
workspace
different categories of distributions and their associated .cpp and .h files
contains RbRegister files; necessary for all new implementations
libs
other directories
RevBayes
core
monitors
(other)
functions
subdirectories include various categories of functions and their associated code
moves
distributions
math
different categories of distributions and their associated .cpp and .h files

Figure S5: Simplified representation of the RevBayes structure. Modified from the RevBayes website, keeping only descriptions of the folders we modified. Note the organizational symmetry between the core directory containing the hard-coded features and the revlanguage directory matching the Rev syntax.

Due the multiple advantages of RevBayes and its increasing use, particularly for macroevolutionary research, we chose this software to implement the OBDP. All our modifications have been carried out in a separate copy of its development branch on GitHub (`https://github.com/revbayes/revbayes/tree/dev-cevo-lab`), and are aimed to be integrated in a future stable release. They consist in 3 key additions detailed in Table S1.

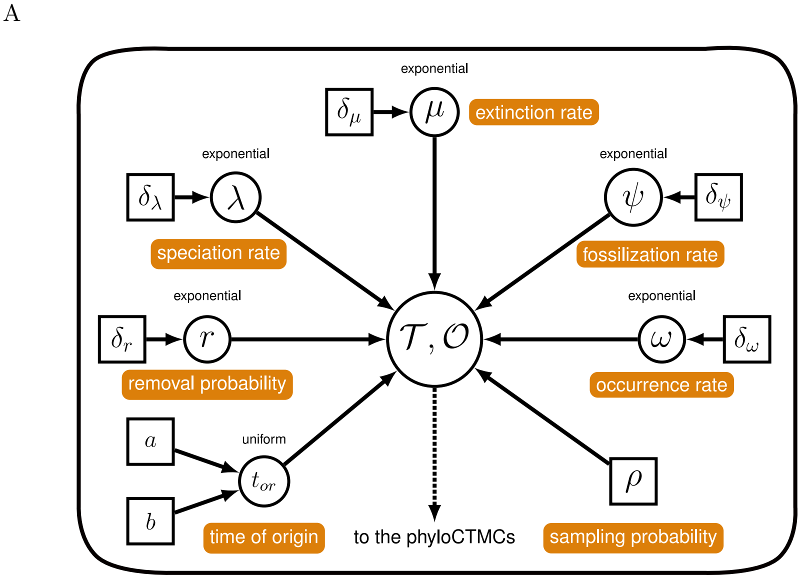
The necessary first step was to implement the core algorithms responsible for computing the quantities L_t and M_t through time. The final organisation is as follows: from outside of the *ComputeLikelihoodsLtMt.cpp* file (see Supp. Table S1) the only functions called are *ComputeLnProbabilityDensitiesOBDP* – returning L_t and M_t through time – or *ComputeLnLikelihoodOBDP* – returning only the final likelihood. Those functions will themselves call the appropriate internal function (*ForwardsTraversalMt* or *BackwardsTraversalLt*) with the correct parameters. Those rely on a key function, *PoolEvents*, the role of which is to construct the vector containing all the events that will be browsed by the traversal algorithms, namely branching times, ψ- and ω-sampling times, and time points for which we want to store the probability distribution.

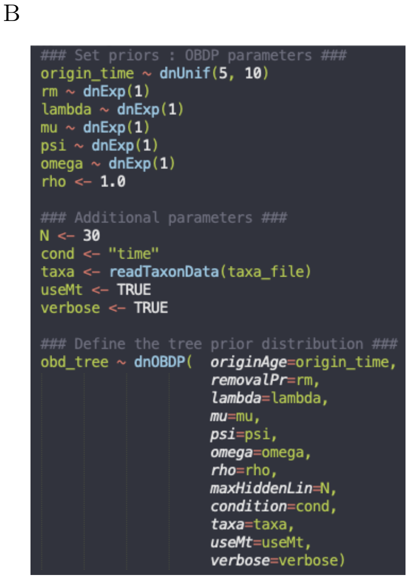
Because the densities computed during the traversals very quickly reached excessively small or elevated values, to the point of exceeding the maximum number of recorded decimals, a correction term is added at each step to bring the densities closer to 1. At the end of the traversal, the recorded correction terms plus the factorizable factors are added to the log-transformed densities.

In addition, the Occurrence Birth-Death Process and the traversal algorithms not only allow us to perform a MCMC phylogenetic inference incorporating the occurrences, they can also be used to output the probability distribution of the population size through time, K_t. We introduced this functionality into RevBayes through *InferAncestralPopSizeFunction*, which can be called directly from the Rev interface. As with the OBDP distribution, we had to design the parameter loading procedure, then call the *ComputeLnProbabilityDensitiesOBDP* function to get the $\log(L_t)$ and $\log(M_t)$) matrices and finally combine and normalize them to obtain the $\log(K_t)$) matrix.

Table S1: Overview of the implementations carried out to incorporate the Occurrence Birth-Death Process and the associated Diversity Inference method into RevBayes. It lists for each of our goals the associated C++ files, along with their assignment in the RevBayes structure.

Objectives	Location	File names	Major new functions
1. Perform Forwards and Backwards traversal algorithms	core/ functions	*ComputeLikelihoods LtMt.h* *ComputeLikelihoods LtMt.cpp*	*ComputeLnProbabilityDensitiesOBDP* *ComputeLnLikelihoodOBDP* *PoolEvents* *ForwardsTraversalMt* *BackwardsTraversalLt*
2. Encode the OBDP distribution	core/ distributions	*OccurrenceBirthDeath Process.h* *OccurrenceBirthDeath Process.cpp*	*OccurrenceBirthDeathProcess* *computeLnProbabilityDivergenceTimes*
	revlanguage/ distributions	*Dist_ occurrenceBirth DeathProcess.cpp* *Dist_ occurrenceBirth DeathProcess.h*	*createDistribution* *getParameterRules*
3. Infer past diversity	core/ distributions	*InferAncestralPop SizeFunction.h* *InferAncestralPop SizeFunction.cpp*	*InferAncestralPopSizeFunction*
	revlanguage/ distributions	*Func_ inferAncestral PopSize.h* *Func_ inferAncestral PopSize.cpp*	*createFunction* *getArgumentRules*





```
### Set priors : OBDP parameters ###
origin_time ~ dnUnif(5, 10)
rm ~ dnExp(1)
lambda ~ dnExp(1)
mu ~ dnExp(1)
psi ~ dnExp(1)
omega ~ dnExp(1)
rho <- 1.0

### Additional parameters ###
N <- 30
cond <- "time"
taxa <- readTaxonData(taxa_file)
useMt <- TRUE
verbose <- TRUE

### Define the tree prior distribution ###
obd_tree ~ dnOBDP(  originAge=origin_time,
                    removalPr=rm,
                    lambda=lambda,
                    mu=mu,
                    psi=psi,
                    omega=omega,
                    rho=rho,
                    maxHiddenLin=N,
                    condition=cond,
                    taxa=taxa,
                    useMt=useMt,
                    verbose=verbose)
```

Figure S6: A graphical model of the OBDP and its translation into the Rev language. (A) Graphical model, modified from the RevBayes FBD tutorial, representing the OBDP parameters – labelled in orange – generating a reconstructed tree $\mathcal{T}$ and a record of occurrences $\mathcal{O}$. (B) Rev script corresponding to this graphical model. Note the distinction between the $\sim$ notation attributing a distribution to a stochastic node and the $\leftarrow$ notation defining a constant node.

B.2 RevGadgets

The postprocessing step consists in computing the posterior probability of the total number of individuals through time. It can be performed independently of the previous steps, given that one has at least a tree, a set of parameters and optionally occurrence times. It comprises 2 steps, the first one uses the *fnInferAncestralPopSize* function, implemented in RevBayes, to obtain the matrix of diversity densities K_t for each tree in the MCMC trace. Then, in order to convert K_t matrices into a nicely rendered plot we added two functions in the auxiliary R package RevGadgets (`https://github.com/revbayes/RevGadgets/tree/dev-plotDiversity`). Starting from the trace of posterior trees, parameters, and K_t matrices one first needs to execute the *rev.process.nbLineages* function that will organize the required information into the Kt_mean data frame. The goal is to incorporate all the uncertainty concerning the inferred parameter values and tree topologies into the diversity trajectory estimation. Afterwards, this averaged Kt_mean is used by the function *rev.plot.nbLineages* to realize the final plot using ggplot2 (Wickham 2016). Here it is possible to alter most of the display options, such as the types of lineages to be shown (observed, hidden, total), as well as their colours and shapes (see e.g. Fig. S8).

Table S2: Description of two novel RevGadgets functions for visualizing OBDP diversity-through-time estimations. The input objects and display parameters are detailed, those with an asterisk always have to be provided while the others have default values.

Function	**Option**	**Type**	**Description**
rev.process .nbLineages	start_ time_ trace_ file*	character	MCMC trace of the starting times.
	popSize_ distribution _ matrices_ file*	character	Matrices computed with fnInferAncestralPopSize in RevBayes.
	trees_ trace_ file*	character	MCMC trace of the trees.
	weight_ trees_ posterior	Boolean	Whether to combine trees uniformly or weighted by their posterior probabilities.
rev.plot .nbLineages	Kt_ mean*	data.frame	Processed output for plotting.
	xlab / ylab	character	Label of the x-axis / y-axis.
	line.size / interval.line.size	numeric	Width of the lineage plot / credible interval line.
	col.Hidden / col.Observed / col.Total / col.Hidden.interval / col.Total.interval	character	Color of the hidden / observed / total lineages plot line. Color of the credible interval for hidden / total lineages.
	palette.Hidden / palette.Total	character	Palette of the hidden / total lineages distribution.
	show.Hidden / show.Observed / show.Total / show.intervals / show.densities / show.expectations	Boolean	Whether to show the plot for hidden / observed / total lineages / credible intervals / diversity densities / diversity expectations.
	use.interpolate	Boolean	Whether to interpolate densities.

C Qualitative validation: "blind test" on simulated data

Parameter values used to simulate the two datasets used in the blind test are presented in Table S3. Two trees with occurrences have been simulated under the OBDP (parameters 1-6). For "dataset 1", genetic sequences along the first tree are simulated according to a K80 model of molecular evolution (parameters 7-9) and recorded only for extant taxa. Binary traits are simulated according to a Markov process with symmetrical rates (parameters 10-12) and are recorded for both extant and extinct taxa. This corresponds to a classic macroevolution scenario. For "dataset 2", genetic sequences along the second tree are simulated according to a K80 model of molecular evolution (parameters 7-9) and recorded for extant and extinct individuals. This allows us to have a better resolution of the underlying tree than in the first dataset. Moreover, getting genetic sequences for individuals sampled in the past corresponds more to an epidemiology scenario.

Table S3: Parameter values used to simulate two datasets and test our OBDP inference workflow.

Parameter values											
λ	μ	ψ	ω	r	ρ	m_{nt}	α_{nt}	β_{nt}	m_{morpho}	q_{01}	q_{10}
1	0.9	0.2	0.3	0	0.8	10000	0.01	0.02	60	0.03	0.03

Two of us, ignorant of the ~~priors~~ Joelle: values used for simulation, designed the inference protocol and conducted the analysis, taking as input the occurrences, sequences, and morphological data only. Priors used for inference on "dataset 1" are presented in Table S4 and the general setup for analysis is illustrated in Figure S7. Priors used for inference on "dataset 2" were very similar, except for the absence of a model of morphological evolution, and they are presented in Table S5.

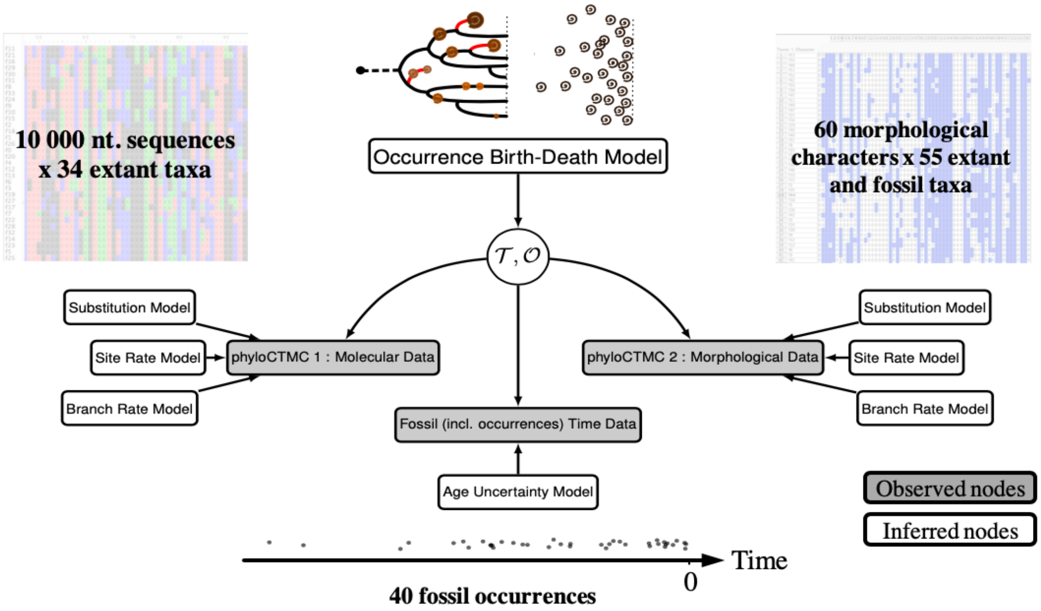


Figure S7: Modular representation of the graphical models used in the qualitative validation analysis. Modified from Heath et al. (2019). The simulated data, noted in the grey nodes are used to deduce the posterior distributions of all other random variables noted in the white nodes.

Table S4: Prior distributions on the OBDP parameters and models for the "Blind Test" analysis on dataset 1. Notations: $\mathcal{U}$ for the Uniform distribution, Exp for Exponential, Dir for Dirichlet, GTR for the General Time Reversible substitution model and MK for the Mk model, the analog of JC69 for an arbitrary number of character states.

Parameter	**Prior**
λ	$Exp(10)$
μ	$Exp(10)$
ψ	$Exp(10)$
ω	$Exp(5)$
ρ	$\mathcal{U}(0,1)$
r	0
t_{or}	$\mathcal{U}(7.7,12)$

Model	**Prior**
Molecular evolution: $GTR+\Gamma$	Strict clock rate: $Exp(10)$ Exchangeability rates: $Dir(1,1,1,1,1,1)$ Stationary frequencies: $Dir(1,1,1,1)$ Gamma distribution shape: $Exp(1)$
Morphological evolution: $MK+\Gamma$	Strict clock rate: $Exp(1)$ Gamma distribution shape: $Exp(1)$

Table S5: Prior distributions of the OBDP parameters and models for the "Blind Test" analysis on dataset 2. Notations: $\mathcal{U}$ for the Uniform distribution, B for the Beta distribution, Exp for Exponential, Dir for Dirichlet, GTR for the General Time Reversible substitution model.

Parameter	**Prior**
λ	$Exp(10)$
μ	$Exp(10)$
ψ	$Exp(10)$
ω	$Exp(10)$
ρ	$B(1.0,1.0)$
r	0
t_{or}	$\mathcal{U}(7.7,12)$

Model	**Prior**
Molecular evolution: $GTR+\Gamma$	Strict clock rate: $Exp(10)$ Exchangeability rates: $Dir(1,1,1,1,1,1)$ Stationary frequencies: $Dir(1,1,1,1)$ Gamma distribution shape: $Exp(1)$

In our blind inferences, we recovered posterior distribution of diversity trajectories (Fig. S8) and trees (Fig. S9) which are very close to the real data from the simulations. The true number of hidden lineages is most of the time near the expectation of the inferred posterior distribution and more importantly always in the 95% posterior credible interval. When looking at the total number of lineages – i.e. species richness in macroevolution or prevalence in epidemiology – the estimates remains very close to the truth and almost always in the 95% credible interval.

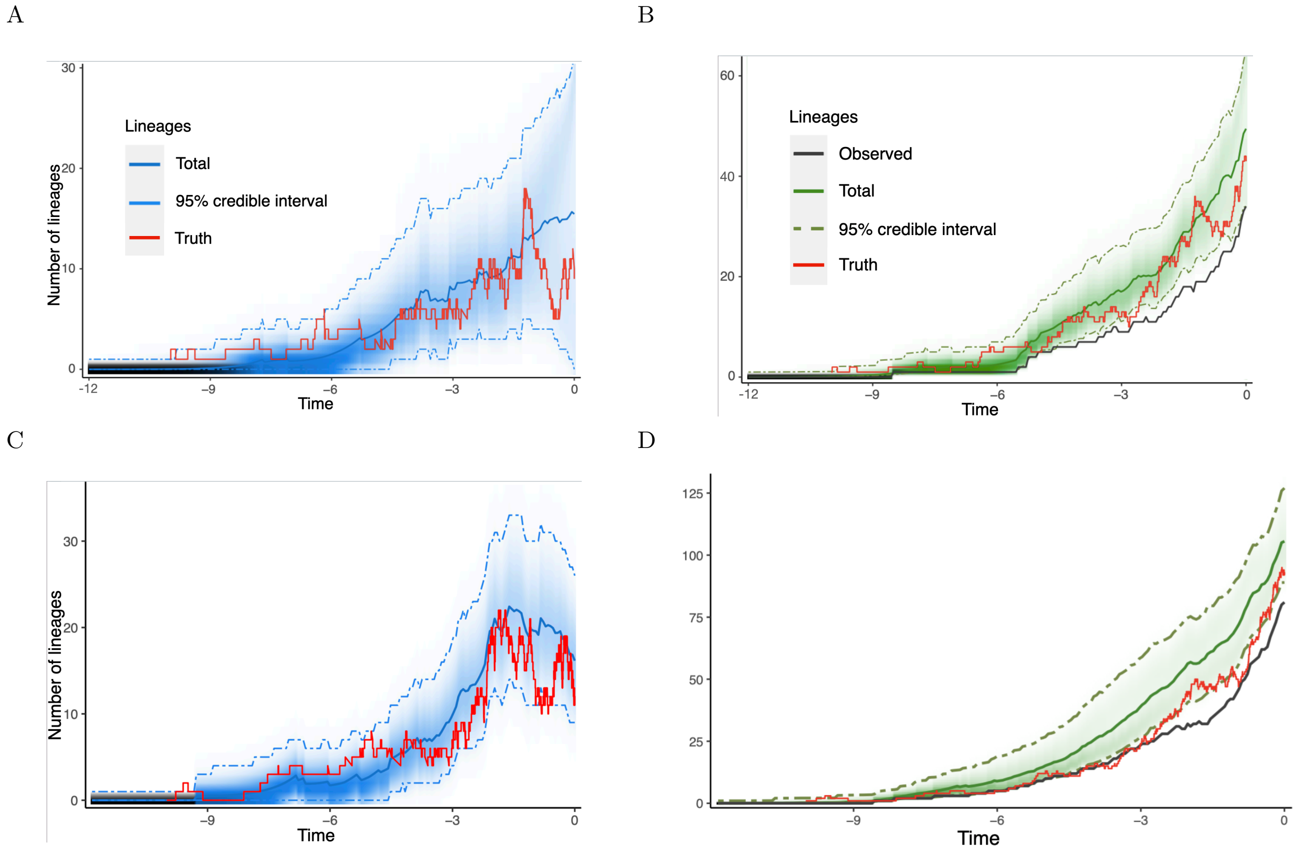


Figure S8: Validation of the diversity dynamics inferred by OBDP compared to the true simulated data. (A) Posterior probability distribution of the number of hidden lineages through time for “dataset 1”, plotted with the new RevGadgets utilities. (B) Posterior probability distribution of the total number of lineages through time for “dataset 1”. (CD) Same as (AB), but for “dataset 2”. The 95% credible intervals are indicated in dashed lines, the expected number of lineages is in blue or green and the true, simulated, trajectory in red. The black line represents the inferred Lineages Through Time (LTT) plot, note that the total diversity equals the LTT plus the hidden diversity.

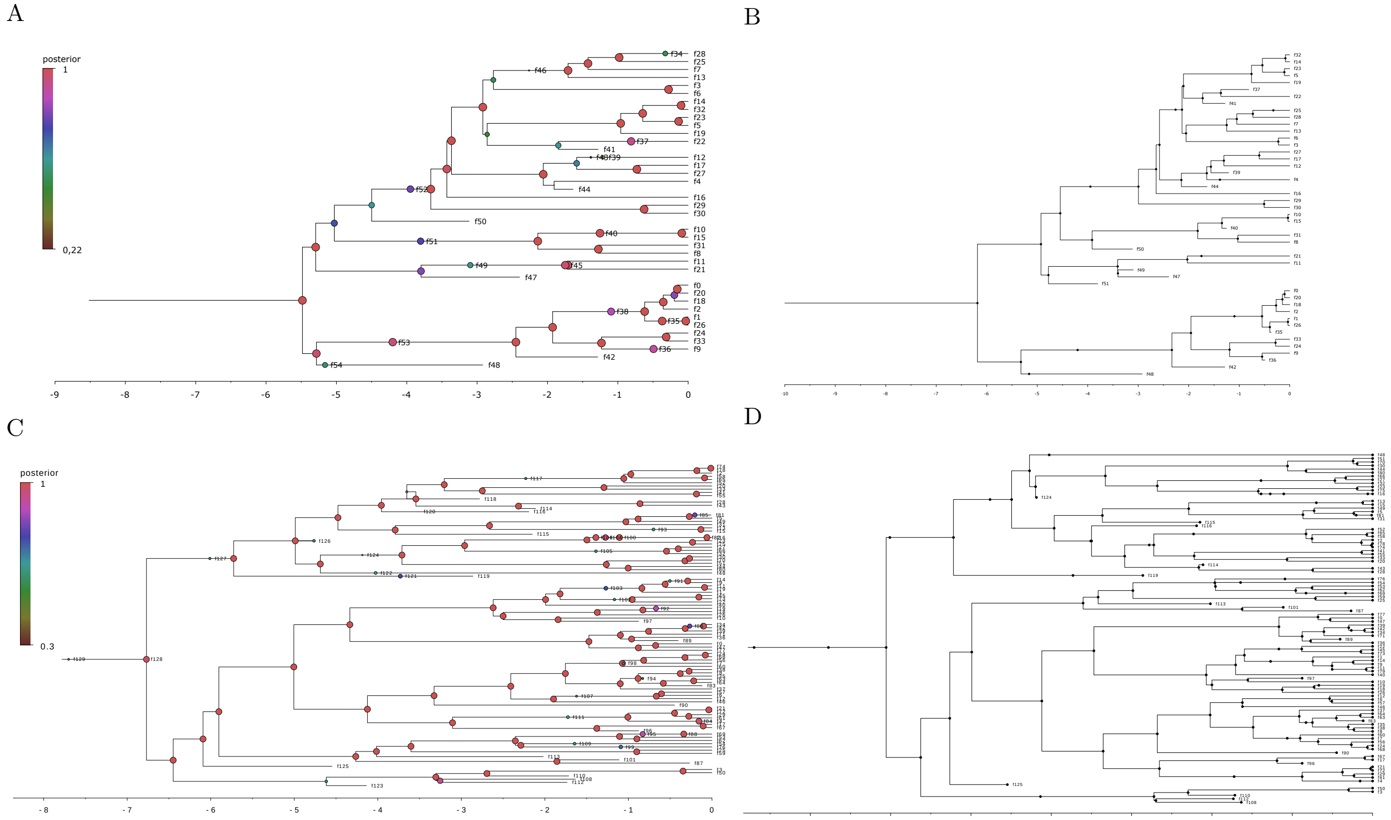


Figure S9: Validation of the inferred trees against the true simulated ones. (A) Inferred phylogenetic tree for "dataset 1", visualized in FigTree 1.4.4. The node colors refer to their posterior probability. (B) Original simulated tree for "dataset 1", aligned on the same temporal scale. Note that the topology is well recovered but divergence dates do not always perfectly match. (CD) Same as (AB) but on "dataset 2". Due to a greater amount of data in genetic sequences of both past and extant individuals, the divergence dates tend to be better inferred.

D Macroevolution application: Inferring past cetacean diversity

D.1 Preliminary analysis of the cetacean occurrence fossil record

A detailed notebook is available at `https://github.com/Jeremy-Andreoletti/Cetacea_PBDB_Occurrences` to follow our exploration of the cetacean dataset. We identified several biases in their fossil record, in particular much more variable occurrence densities – defined as the number of occurrences by unit of time in the stratigraphic range of a clade – than expected from our model (see Figure S10).

Since OBDP assumes that only one individual of a species will be sampled at a time, we subsampled the dataset to aggregate all occurrences of the same taxon found in the same geological formation. This subsampling also reduced the observed discrepancy in occurrence densities. The final subsampled dataset was composed of 968 occurrences.

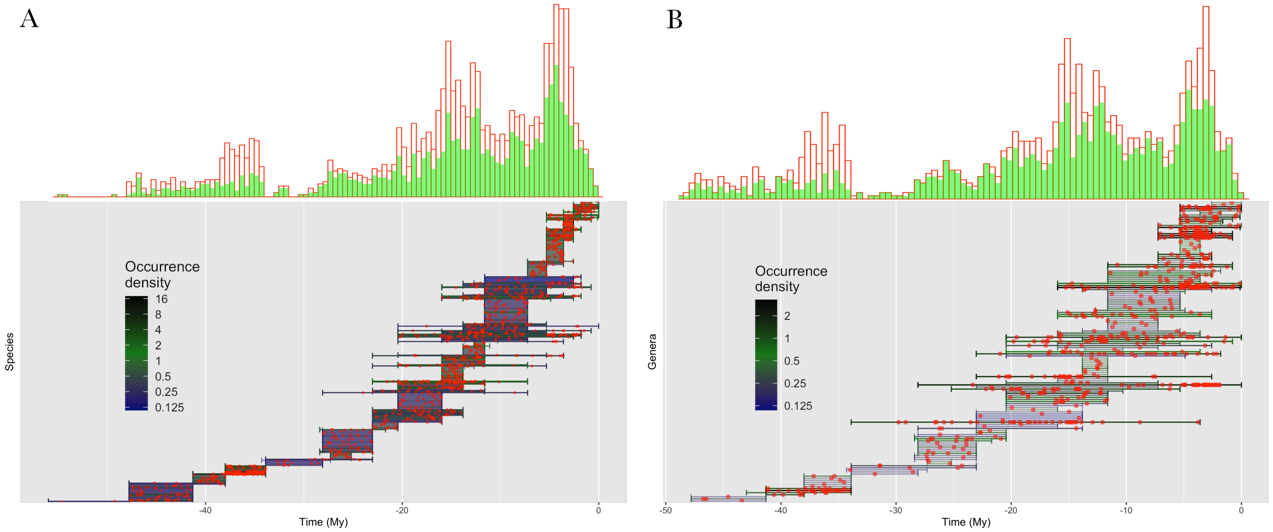


Figure S10: Occurrence distributions and bias correction, for cetacean species (A) and genera (B). At the top, occurrence distributions are compared before (red) and after (green) aggregating in geological formations. Below, stratigraphic ranges are displayed over time and colored according to the density of occurrences (red dots).

D.2 Detailed priors used for Bayesian inference

We detail in Table S6 all priors used for the inference on the cetacean dataset.

Table S6: Prior distributions for parameters and models of the Cetacea analysis. For each parameter its prior distribution, its initial value at the origin of the MCMC chain (set to speed up convergence) and the references that support these choices are indicated. Notations: $\mathcal{U}$ for the Uniform distribution, Exp for Exponential, $Log\mathcal{N}$ for Log-Normal, Dir for Dirichlet, GTR for General Time Reversible and $JC69$ for the Jukes-Cantor 1969.

Component	Prior	Initial	Justification
t_{or}	$\mathcal{U}(max($ occurrence_ages), 60)	54.0	Origin after the last occurrence. Initialised close to the estimated Whippomorpha root age from McGowen et al. (2020)
μ	$Exp(10)$	0.1	Initialized according to estimations by Rabosky (2014)
$\lambda - \mu$	$Log\mathcal{N}(\ln[\frac{\ln 89}{t_{or}}],$ 0.587405)	$\frac{\ln 89}{t_{or}}$	Expected number of species under a Birth-Death process centred around the observed number of species. Lognormal distribution with 95% prior probability spanning exactly one order of magnitude (Höhna and Heath 2019)
r	0	0	Removal probability at sampling, irrelevant in macroevolution
$\psi + \omega$	$Exp(10)$	Random	Unknown sampling rate for all fossils (including occurrences)
$\omega/(\psi + \omega)$	$\mathcal{U}(0, 1)$	Empirical	Unknown probability that morphological characters are available for a given fossil. Initialized at the empirical proportion of fossils with morphology among all fossils
ρ	$\mathcal{U}(0.95, 1)$	Random	Sequences or morphology is used for the 89 accepted extant cetacean species, but we allow for some still unknown species
Fossil age uncertainty	$\mathcal{U}(min, max)$	Minimum age	Moves shifting a fossil age outside of its range are rejected (Heath et al. 2019)
Mean molecular clock rate	Nuclear: $\mathcal{U}(0, 0.01)$ Mitochondrial: $\mathcal{U}(0, 0.1)$	0.0005 0.02	Priors based on rates of molecular evolution for all mammals in Allio et al. (2017). Initialised at an intermediate rate between mysticetes and odontocetes as estimated by Dornburg et al. (2012).
Clock rate relaxation	Uncorrelated: $Exp(1/mean)$	Random	Independent and identically distributed exponential rates are defined for each branch
Molecular substitution model: $GTR + \Gamma$	Exchangeability rates: $Dir(1, 1, 1, 1, 1, 1)$ Stationary frequencies: $Dir(1, 1, 1, 1)$ Gamma distribution shape: $Exp(1)$		Sophisticated nucleotide evolution model with rate variation across sites according to a discretized Gamma distribution. The Dirichlet distributions constrain vectors to sum to one (Heath et al. 2019)
Morphological substitution model: $JC69$	Strict clock rate: $Exp(1)$ Gamma distribution shape: $Exp(1)$		Simpler character evolution model. Characters are partitioned according to their number of states (Wright 2020)

E Epidemiology application: the Diamond Princess SARS-2 COVID-19 outbreak dynamics

E.1 Data acquisition on GISAID

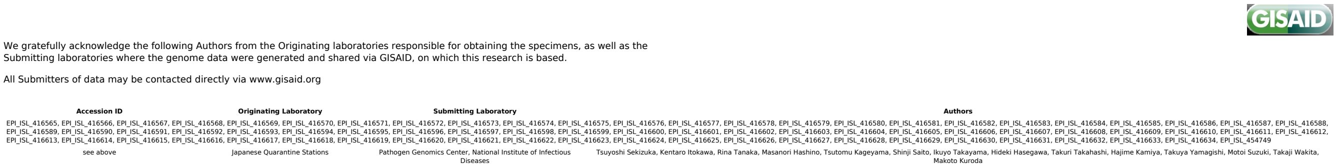
We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based. All Submitters of data may be contacted directly via www.gisaid.org

accession ID EPI_ISL_416565, EPI_ISL_416566, EPI_ISL_416567, EPI_ISL_416568, EPI_ISL_416569, EPI_ISL_416570, EPI_ISL_416571, EPI_ISL_416572, EPI_ISL_416573, EPI_ISL_416574, EPI_ISL_416575, EPI_ISL_416576, EPI_ISL_416577, EPI_ISL_416578, EPI_ISL_416579, EPI_ISL_416580, EPI_ISL_416581, EPI_ISL_416582, EPI_ISL_416583, EPI_ISL_416584, EPI_ISL_416585, EPI_ISL_416586, EPI_ISL_416587, EPI_ISL_416588, EPI_ISL_416589, EPI_ISL_416590, EPI_ISL_416591, EPI_ISL_416592, EPI_ISL_416593, EPI_ISL_416594, EPI_ISL_416595, EPI_ISL_416596, EPI_ISL_416597, EPI_ISL_416598, EPI_ISL_416599, EPI_ISL_416600, EPI_ISL_416601, EPI_ISL_416602, EPI_ISL_416603, EPI_ISL_416604, EPI_ISL_416605, EPI_ISL_416606, EPI_ISL_416607, EPI_ISL_416608, EPI_ISL_416609, EPI_ISL_416610, EPI_ISL_416611, EPI_ISL_416612, EPI_ISL_416613, EPI_ISL_416614, EPI_ISL_416615, EPI_ISL_416616, EPI_ISL_416617, EPI_ISL_416618, EPI_ISL_416619, EPI_ISL_416620, EPI_ISL_416621, EPI_ISL_416622, EPI_ISL_416623, EPI_ISL_416624, EPI_ISL_416625, EPI_ISL_416626, EPI_ISL_416627, EPI_ISL_416628, EPI_ISL_416629, EPI_ISL_416630, EPI_ISL_416631, EPI_ISL_416632, EPI_ISL_416633, EPI_ISL_416634, EPI_ISL_454749

Originating Laboratory Japanese Quarantine Stations

Submitting Laboratory Pathogen Genomics Center, National Institute of Infectious Diseases

Authors Tsuyoshi Sekizuka, Kentaro Itokawa, Rina Tanaka, Masanori Hashino, Tsutomu Kageyama, Shinji Saito, Ikuyo Takayama, Hideki Hasegawa, Takuri Takahashi, Hajime Kamiya, Takuya Yamagishi, Motoi Suzuki, Takaji Wakita, Makoto Kuroda



We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based.

All Submitters of data may be contacted directly via www.gisaid.org

Accession ID	Originating Laboratory	Submitting Laboratory	Authors
EPI_ISL_416565, EPI_ISL_416566, EPI_ISL_416567, EPI_ISL_416568, EPI_ISL_416569, EPI_ISL_416570, EPI_ISL_416571, EPI_ISL_416572, EPI_ISL_416573, EPI_ISL_416574, EPI_ISL_416575, EPI_ISL_416576, EPI_ISL_416577, EPI_ISL_416578, EPI_ISL_416579, EPI_ISL_416580, EPI_ISL_416581, EPI_ISL_416582, EPI_ISL_416583, EPI_ISL_416584, EPI_ISL_416585, EPI_ISL_416586, EPI_ISL_416587, EPI_ISL_416588, EPI_ISL_416589, EPI_ISL_416590, EPI_ISL_416591, EPI_ISL_416592, EPI_ISL_416593, EPI_ISL_416594, EPI_ISL_416595, EPI_ISL_416596, EPI_ISL_416597, EPI_ISL_416598, EPI_ISL_416599, EPI_ISL_416600, EPI_ISL_416601, EPI_ISL_416602, EPI_ISL_416603, EPI_ISL_416604, EPI_ISL_416605, EPI_ISL_416606, EPI_ISL_416607, EPI_ISL_416608, EPI_ISL_416609, EPI_ISL_416610, EPI_ISL_416611, EPI_ISL_416612, EPI_ISL_416613, EPI_ISL_416614, EPI_ISL_416615, EPI_ISL_416616, EPI_ISL_416617, EPI_ISL_416618, EPI_ISL_416619, EPI_ISL_416620, EPI_ISL_416621, EPI_ISL_416622, EPI_ISL_416623, EPI_ISL_416624, EPI_ISL_416625, EPI_ISL_416626, EPI_ISL_416627, EPI_ISL_416628, EPI_ISL_416629, EPI_ISL_416630, EPI_ISL_416631, EPI_ISL_416632, EPI_ISL_416633, EPI_ISL_416634, EPI_ISL_454749			
see above	Japanese Quarantine Stations	Pathogen Genomics Center, National Institute of Infectious Diseases	Tsuyoshi Sekizuka, Kentaro Itokawa, Rina Tanaka, Masanori Hashino, Tsutomu Kageyama, Shinji Saito, Ikuyo Takayama, Hideki Hasegawa, Takuri Takahashi, Hajime Kamiya, Takuya Yamagishi, Motoi Suzuki, Takaji Wakita, Makoto Kuroda

Figure S11: Genome sequences used, originating and submitting labs generated on GISAID. Content is reproduced above.

E.2 Pre-processing the data

All case count and sequencing data were available at a resolution of days.

In order to use the main method described in this article, the case count record had to be pre-processed so that occurrences are spread throughout the days. For a day with a case count of n newly infected individuals, we drew n time points uniformly distributed throughout the day. The resulting dataset is shown in Figure S12.

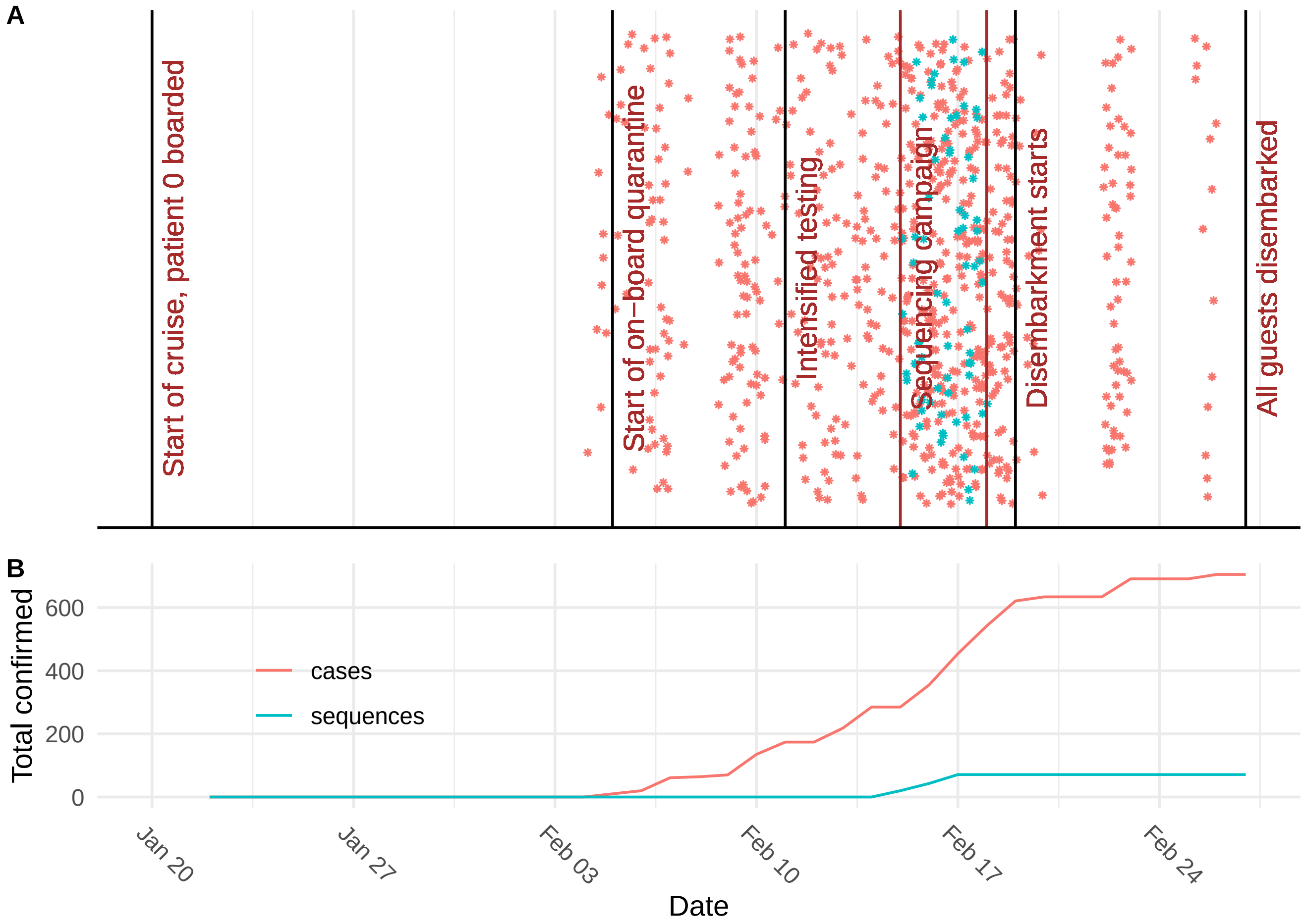


Figure S12: Pre-processed dataset for the Diamond Princess outbreak analysis. (A) Exact dates assigned to occurrences and sequences for the analysis. (B) Total case counts and sequences through time.

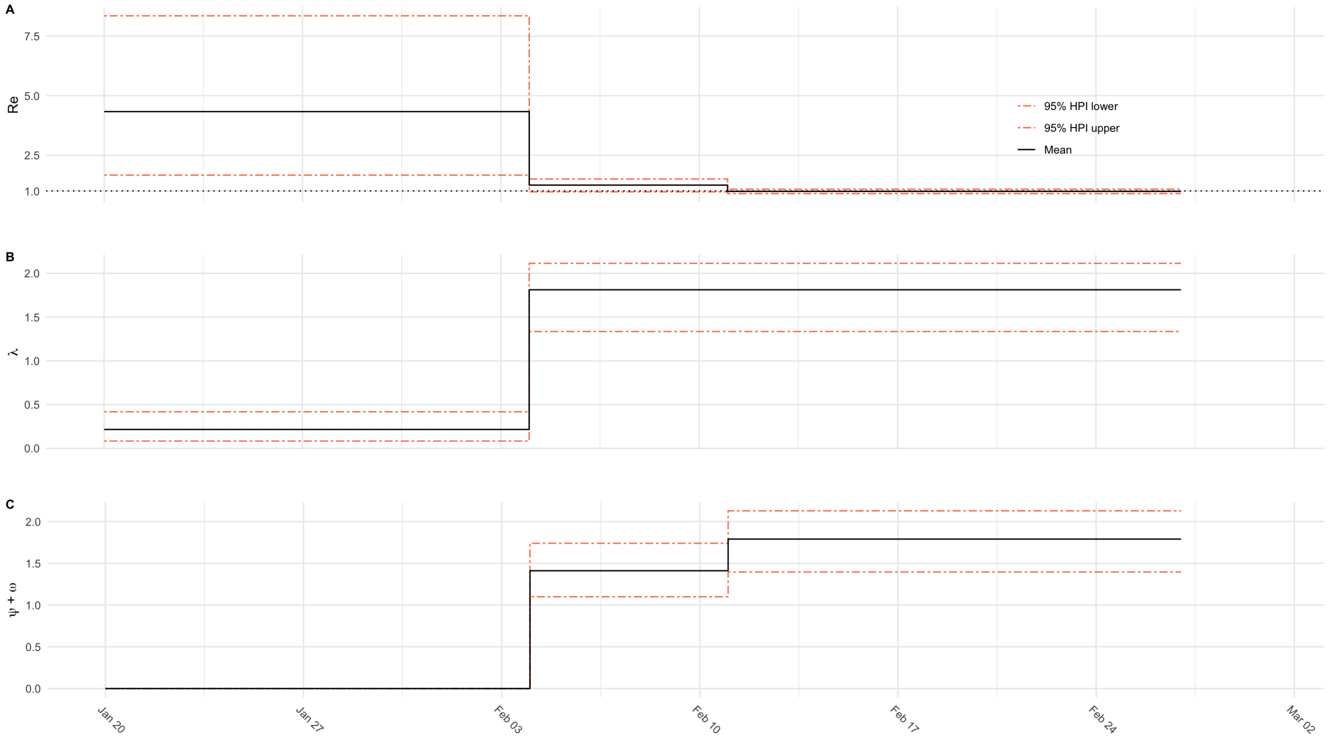


Figure S13: Detailed parameter estimates obtained from the COVID-19 outbreak analysis. (A) Reproductive number estimates. (B) Birth rate estimates. (C) Total sampling (sequencing and PCR testing) rate estimates.

E.3 Detailed priors

We detail in Table S7 all priors used for the inference on the outbreak dataset of COVID-19 aboard the Diamond Princess.

The mean of the prior distribution of $\psi + \omega$ is set up to be the number of tests used on the ship, per day and per passenger, on the two periods.

- Within the first 7 days period, from February 4th to February 11th, there were 439 tests carried out, on 3711 passengers, leading to $\frac{439}{7 \times 3711} \approx 1.7 \times 10^{-2}$ tests per day per passenger.
- on the following 15 days period, from February 11th to February 27th, there were 3622 tests carried out, on 3711 passengers, leading to $\frac{3622}{15 \times 3711} \approx 6.5 \times 10^{-2}$ tests per day per passenger.

Table S7: Prior distributions for parameters and models of the SARS-2 COVID-19 analysis. For each parameter its prior distribution or value and the references that support these choices are indicated.

Component	**Prior/Value**	**Shifts**	**Justification**
t_{or}	38	N/A.	We study the outbreak from the start of the cruise on January 20, until February 27, when all guests were confirmed to have disembarked the ship, (Ministry of Health and Welfare 2020) spanning a total period of 38 days.
μ	1/20 day^{-1}	None.	In the absence of sampling and removal,infected individuals patients are assumed to become uninfectious on average 20 days after infection. (He et al. 2020)
λ	$\mathcal{U}(0, 24)$ $\mathcal{U}(0, 10)$	$t_m = (04.02.2020)$	The upper bound is set to 1 transmission per hour per infected individual before cabin isolation and lowered to 10 individuals after (maximal cabin size), from February 4th onward.
$\psi + \omega$	$Log\mathcal{N}\left(\frac{3622}{15*3711}, 0.5\right)$ $Log\mathcal{N}\left(\frac{439}{7*3711}, 0.5\right)$	$t_m = (11.02.2020, 04.02.2020)$	Testing started on February 4th and was intensified from February 11th onward, yielding two periods of 7 days and 15 days each. For each time period, the mean for the LogNormal distribution is set as the number of tests taken per passenger per day. The total numbers of tests carried out throughout the quarantine were communicated in press releases (Ministry of Health and Welfare 2020)
r	1	None.	Quarantine measures are assumed to have minimised contact between guests aboard. Patients testing positive were disembarked from the ship to a separate medical facility.
ρ	0	None.	No samples were sequenced after February 17th.
$\frac{\psi}{\omega+\psi}$	$\frac{71}{328}$	None.	Set to the fraction of the samples testing positive for COVID-19 that were sequenced.
Clock rate	8×10^{-4} substitutions per site per year	N/A.	(Stadler 2020a)
Molecular substitution model: $GTR + \Gamma$	Exchangeability rates: $Dir(1, 1, 1, 1, 1, 1)$ Stationary frequencies: $Dir(1, 1, 1, 1)$ Gamma distribution shape: $Exp(1)$		We allow for site rate heterogeneity, and assume unequal base frequencies and transition/transversion rates.