A theoretical framework linking the birth-death and protracted birth-death models, with implications for the phylogenetic analysis of diversification

Pierre Veron [1,2,*], Jérémy Andréoletti [1], Tatiana Giraud [2], and Hélène Morlon [1]

[1]Institut de Biologie, École Normale Supérieure, Université PSL, CNRS, INSERM, Paris, France
[2]Écologie Systématique et Évolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France
[*]Corresponding author, pveron@bio.ens.psl.eu

July 1, 2024

Keywords speciation, macroevolution, microevolution, phylogeny

Abstract

While speciation takes time, standard birth-death processes used to infer diversification from phylogenies assume that it is instantaneous. This limits the interpretation of phylogenetic estimates of speciation and extinction rates, and our understanding of the microevolutionary processes that modulate these rates. Here, we consider the protracted birth-death model, which accounts for the fact that speciation takes time. We compute an "equivalent" standard birth-death process, of which diversification dynamics approximate those induced by the protracted process. We find that the equivalent birth rate declines close to the present, and that the duration of speciation influences when this decays occurs more than the equivalent rates themselves. The equivalent rates are relatively constant in the past. The birth rate in the past scales with the rate at which speciation is initiated, with a scaling factor that depends mostly on the population extinction rate. Our results suggest that the rates of speciation initiation and population extinction may often play a larger role than the speed at which populations acquire reproductive isolation in modulating speciation rates. Our study establishes a theoretical framework for understanding how microevolutionary processes combine to explain the diversification of species groups on macroevolutionary time scales.

1 Introduction

Birth-death models are widely used to understand the diversification of species groups; in this context, births represent speciation events, i.e. the emergence of two daughter species from a ancestral one, and deaths represent species extinction. This specific use of birth-death models is particularly widespread to interpret both fossil (Silvestro et al. 2014) and phylogenetic data (Stadler 2013; Morlon, Andréoletti, et al. 2024) in terms of diversification dynamics. In birth-death models, speciation is considered to be an instantaneous phenomenon, represented as a branching event following a Poisson point process.

Despite the widespread use of birth-death models to simulate speciation and extinction events, speciation is not instantaneous. Speciation requires an initial isolation of populations (speciation initiation) followed by the accumulation of genetic barriers to gene flow until speciation is complete. Speciation can initiate and fail before completion, for example because of secondary contact, or because isolated populations go extinct before speciation completion (Coyne and Orr 2004; Dynesius and Jansson 2014). The whole speciation process may take hundreds of thousands up to several millions of years (Benton and Pearson 2001; Etienne and Rosindell 2012; Etienne, Morlon, et al. 2014; Hua et al. 2022).

Ignoring the fact that speciation takes time by using standard birth-death models has non trivial consequences for our understanding of diversification dynamics. For example, when standard birth-death models are used in combination with phylogenetic trees of extant species to estimate speciation and extinction rates, the "protracted" nature of speciation may be misinterpreted as a speciation rate slowdown towards the present (Etienne and Rosindell 2012; Moen and Morlon 2014).

Etienne and Rosindell 2012 pioneered the development of the so-called protracted birth-death model (PBD). Instead of assuming that speciation is instantaneous as in the standard birth-death model, this model assumes that there are events of speciation *initiation* corresponding to the formation of *incipient* species that eventually become *good* species after a random *completion* time. An incipient lineage is a lineage that is not yet considered as a different species from the ancestral lineage. For sexually reproducing organisms, the completion time is the time it takes for lineages to achieve reproductive isolation. Each lineage is thus subject to initiation, extinction and completion events.

The protracted birth-death model has several advantages over the the standard birth-death model. First, it is biologically more realistic, and it thus unsurprisingly produces phylogenies that are closer to empirical phylogenies than those produced by the standard birth-death model (Etienne and Rosindell 2012). Specifically, it produces phylogenies that are less tippy (fewer recent speciation events) that those arising from the standard birth-death model. Second, it allows the integration of intraspecific processes that lead to speciation. For example, the matching competition birth-death model (MCBD, Aristide and Morlon 2019) integrates the effect of intraspecific competition on character displacement leading to speciation by modeling character displacement in incipient lineages.

Despite the advantages of the protracted birth-death model, the overwhelming majority of phylogenetic analyses of diversification use the standard birth-death model. Most available models for phylogenetic analyses of diversification are versions of the standard birth-death model, with birth and death rates that can vary in time and/or across lineages (Morlon, Andréoletti, et al. 2024). The protracted birth-death model can be fitted to empirical phylogenies, however not all of its parameters can be reliably estimated from a phylogeny (Etienne, Morlon, et al. 2014), which limits its usefulness. Recently, Hua et al. showed that the PBD parameters can be accurately estimated from population-level (rather than species-level) phylogenies, however such phylogenies remain rare (Hua et al. 2022). Fitting standard birth-death models thus remains the rule in phylogenetic analyses of diversification.

If speciation takes time but it is estimated by fitting standard birth-death models to phylogenies, which assume instantaneous speciation, what do resulting speciation and extinction rate estimates actually represent? We can expect that speciation rate estimates will be higher when rates of speciation initiation and completion are higher, and rates of extinction of incipient species are lower, but precisely answering this question requires to establish an analytical relationship between the parameters of the protracted and standard birth-death models. To our knowledge, such a relationship has not yet been established.

Elucidating the relationship between the parameters of the protracted and standard birth-death models is important not only to clarify the meaning of speciation rates estimated from phylogenies, but also to understand the microevolutionary processes that modulate these rates (Morlon, Andréoletti, et al. 2024). Indeed, macroevolutionary speciation rates (estimated from phylogenies) vary by orders of magnitude (Maliet, Hartig, et al. 2019; Quintero et al. 2024), but the processes underlying this variation remain unclear. Efforts to find empirical correlations between macroevolutionary speciation rates and rates of population formation or evolution of reproductive isolation have not been conclusive; a proposed explanation is that this expected correlation is erased by the frequent extinction of incipient species (Rabosky and Matute 2013; Singhal, Colli, et al. 2022; Singhal, Huang, et al. 2018). That is, population survival rather than population formation and the accumulation of reproductive barriers may be the factor "limiting" speciation. More generally, each of speciation initiation, speciation completion and population survival may be the process limiting macroevolutionary speciation rates in some situations but not others (Rabosky 2016). For instance, a lineage that has a propensity to accumulate fast reproductive isolation but does not experience frequent population splits might not have a high speciation rate: here, the rate of speciation completion is not limiting. Hence, factors acting on speciation initiation, the accumulation of reproductive isolation, and the extinction of incipient species can have non trivial outcomes in terms of speciation rate.

Here, we obtain a mathematical link between the parameters of the protracted and standard birth-death diversification models by computing "equivalent" speciation and extinction rates that generate some of the outcomes expected under the protracted birth-death model. In spirit, these equivalent rates are meant to represent the macroevolutionary outcomes (macroevolutionary speciation and extinction rates) of the protracted

birth-death process, as we may estimate them by fitting standard birth-death model to reconstructed phylogenies. We compare phylogenies simulated under the protracted birth-death model and the equivalent birth-death model and discuss the consequences of our results for the phylogenetic analysis of diversification.

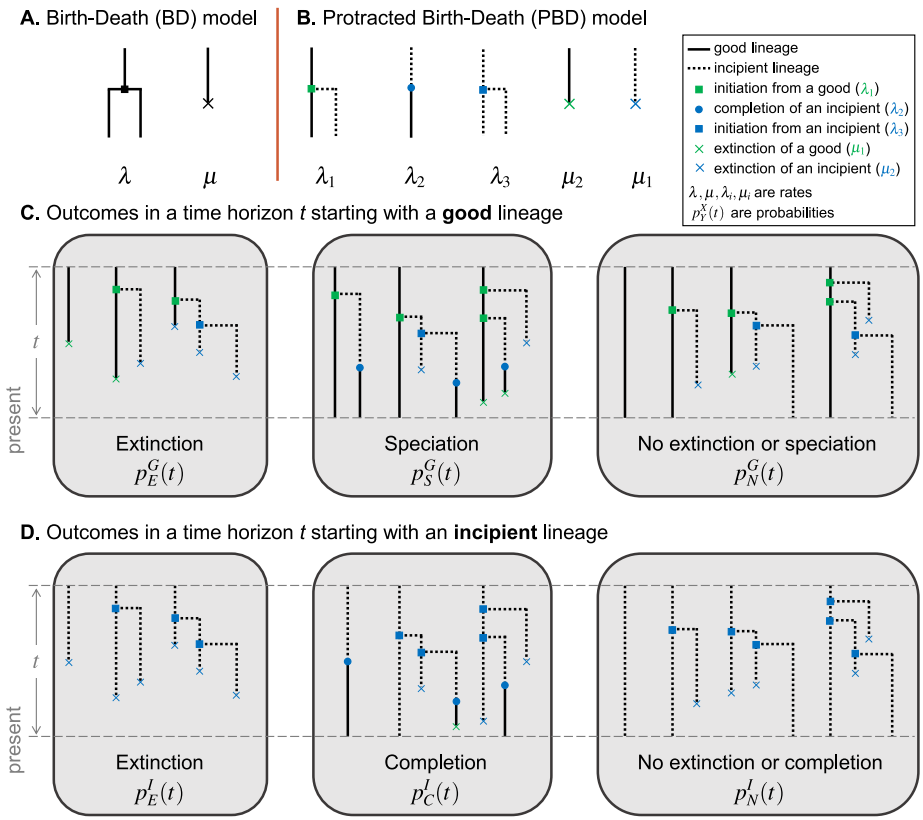

Figure 1. The birth-death (BD) and protracted birth-death (PBD) models. (**A, B**) Illustrations of the rates involved in the BD model (**A**), and the PBD model (**B**). (**C, D**) Possible outcomes of the PBD process in a fixed time horizon, starting from a good lineage (**C**) or an incipient lineage (**D**); non exhaustive examples of events leading to these outcomes are shown.

2 Materials and methods

2.1 Birth-death and protracted birth-death models

The standard birth-death (BD) model involves two rates (figure 1**A**): the birth (speciation) rate λ and the death (extinction) rate μ. The protracted birth-death (PBD) model as defined in (Etienne, Morlon, et al. 2014) involves 5 rates (figure 1**B**) : the rate of speciation initiation from a good species λ_1, the rate of completion λ_2, the rate of speciation initiation from an incipient species λ_3, the rate of extinction of a good species

μ_1, and the rate of extinction of an incipient species μ_2.

By convention, time elapses in the direction of increasing values. The process begins at time 0 with one good lineage and runs until the present at time T. We now consider an intermediate time $T-t$, and introduce $p_E^G(t)$ and $p_S^G(t)$ the probabilities that, for any given good lineage, the first event is respectively an extinction or a speciation event, and $p_N^G(t)$ the probability that none of these events occur during a time interval of length t (figure 1**C**). Similarly, $p_E^I(t)$ and $p_C^I(t)$ are the probabilities that, during a time interval of length t, the first event is respectively an extinction, or a completion event, when starting with an incipient lineage; $p_N^I(t)$ is the probability that none of these events occur during the time interval (figure 1**D**). By "extinction" we mean the direct extinction of the lineage in question, or the extinction of all possible descendants of this lineage in the time interval considered. Similarly, by "completion" we mean the direct completion of the lineage in question, or the completion of an incipient daughter lineage in the time interval under consideration. In case of multiple events occurring in the time interval (e.g. speciation followed by extinction), we consider only the first event. For instance the third case of speciation in figure 1**C** (where the lineages all go extinct after speciation) is considered as a speciation event and is recorded in $p_S^G(t)$.

2.2 Equivalent time-dependent BD rates

Definitions of equivalent birth and death rates, and relation with probabilities of speciation and extinction under the PBD

Given a PBD process with fixed parameter values running from 0 to T, we assume that the probabilities of speciation and extinction within an infinitesimal time interval $[t-\mathrm{d}t,t]$ can be written as $\hat{\lambda}(t)\,\mathrm{d}t$ and $\hat{\mu}(t)\,\mathrm{d}t$, for any time t. We call the quantities $\hat{\lambda}(t)$ and $\hat{\mu}(t)$ the time-dependent equivalent birth and death rates. Given a good lineage alive at time $T-t-\mathrm{d}t$, the probability that the first event occurring within the time interval $[T-t-\mathrm{d}t,T]$ is a speciation event is given by:

$$p_S^G(t+\mathrm{d}t) = \underbrace{\hat{\lambda}(T-t)\,\mathrm{d}t}_{(i)} + \underbrace{(1-\hat{\lambda}(T-t)\,\mathrm{d}t-\hat{\mu}(T-t)\,\mathrm{d}t)}_{(ii)}\underbrace{p_S^G(t)}_{(iii)}$$

with (i) the probability of speciation within the small time interval $[T-t-\mathrm{d}t,T-t]$, (ii) the probability of no speciation nor extinction within the small time interval $[T-t-\mathrm{d}t,T-t]$ and (iii) the probability that the first event occurring within the time interval $[T-t,T]$ is a speciation event, conditioned on existence of the lineage at the time $T-t$. Similarly we have:

$$p_E^G(t+\mathrm{d}t) = \hat{\mu}(T-t)\,\mathrm{d}t + (1-\hat{\lambda}(T-t)\,\mathrm{d}t-\hat{\mu}(T-t)\,\mathrm{d}t)p_E^G(t).$$

Hence, we have the dynamical system

$$\begin{cases} \frac{\mathrm{d}p_S^G(t)}{\mathrm{d}t} = \hat{\lambda}(T-t) - (\hat{\lambda}(T-t)+\hat{\mu}(T-t))p_S^G(t) \\ \frac{\mathrm{d}p_E^G(t)}{\mathrm{d}t} = \hat{\mu}(T-t) - (\hat{\lambda}(T-t)+\hat{\mu}(T-t))p_E^G(t) \end{cases}$$

which allows us to express the time-dependent birth rate:

$$\hat{\lambda}(t) = \frac{\left(1 - p_E^G(T-t)\right) \frac{\mathrm{d}p_S^G}{\mathrm{d}t}\Big|_{T-t} + p_S^G(T-t) \frac{\mathrm{d}p_E^G}{\mathrm{d}t}\Big|_{T-t}}{p_N^G(T-t)} \tag{1}$$

and death rate:

$$\hat{\mu}(t) = \frac{\left(1 - p_S^G(T-t)\right) \frac{\mathrm{d}p_E^G}{\mathrm{d}t}\Big|_{T-t} + p_E^G(T-t) \frac{\mathrm{d}p_E^G}{\mathrm{d}t}\Big|_{T-t}}{p_N^G(T-t)}. \tag{2}$$

In what follows, we compute the probabilities $p_S^G(t)$ and $p_E^G(t)$ for any time $t \in [0,T]$, which provides us with the equivalent rates.

By assuming that the probabilities of speciation and extinction within an infinitesimal time interval $[t-\mathrm{d}t, t]$ can be written as $\hat{\lambda}(t)\,\mathrm{d}t$ and $\hat{\mu}(t)\,\mathrm{d}t$, we have assumed that these probabilities depend only on time, and not on the history of the lineage considered. This is an approximation, because (under the PBD process) good lineages carry the history of their incipient species. An old lineage is indeed more likely to have a pool of incipient lineages (for instance, when λ_3 is high, their number increases exponentially with time) and therefore less likely to go extinct. The equations we used for the probabilities of speciation $p_S^G(t)$ and extinction $p_E^G(t)$ correspond to the case of a good lineage with no incipient species, and are approximations otherwise. We expect that these approximations will affect mainly the extinction rate, as we ignore the buffering effect on extinction of incipient lineages that exist at the time when the rate is computed. The equivalent extinction rate could therefore overestimate extinction probabilities.

Speciation and extinction probabilities under the PBD

In order to compute $p_S^G(t)$ and $p_E^G(t)$ for any time $t \in [0,T]$, we first need to compute probabilities associated with incipient lineages.

Starting from an incipient lineage, the two exclusive possible ways leading to extinction within a horizon t are (figure 2, upper left panel): (a) the lineage directly goes extinct within a time t or (b) the lineage forms two incipient lineages after a time $u \leq t$ and both incipient lineages go extinct within the remaining time $t-u$. The probability for an incipient lineage and its potential descendants to go extinct within a time t thus satisfies the following equation:

$$\begin{aligned} p_E^I(t) = & \frac{\mu_2}{\Lambda}(1 - e^{-\Lambda t}) && \text{figure 2(a)} \\ & + \int_0^t \lambda_3 e^{-\Lambda u} (p_E^I(t-u))^2 \,\mathrm{d}u && \text{figure 2(b)} \end{aligned} \tag{3}$$

where $\Lambda = \lambda_2 + \lambda_3 + \mu_2$.

The solution to this integral equation is (appendix A1):

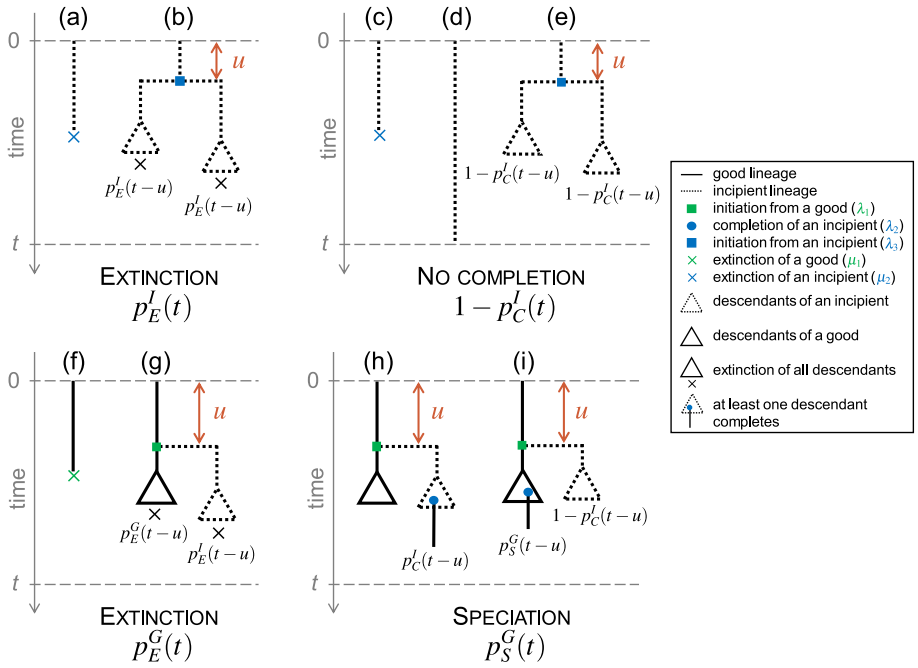

Figure 2. Alternative scenarios by which main outcomes occur in the protracted birth-death model. Upper panels: starting from an incipient lineage at time 0, decomposition of the possible exclusive ways of extinction (left) and no completion (right) before a given time t. Bottom panels: starting from a good lineage at time 0, decomposition of the possible exclusive ways of extinction (left) and speciation (right) before a given time t. Dotted lines represent incipient lineages, solid lines represent good lineages. Triangles summarize the subtrees containing the potential descendants of an ancestor lineage, with the condition that all of them go extinct within the remaining time (indicated by a cross), or that one of the completes speciation (indicated by a blue dot), or that none of these event occur. The probabilities written under the triangles correspond to the probability of the described event.

$$p_E^I(t) = \frac{1}{\lambda_3}\sqrt{\frac{c(\Lambda-k)e^{kt}+\Lambda/k}{(ce^{kt}+1/k)^2} - \frac{\Lambda(k-\Lambda)+2\mu_2\lambda_3}{2}}$$

with $k = \sqrt{\Lambda^2 - 4\mu_2\lambda_3}$ and $c = \frac{2}{k-\Lambda} - \frac{1}{k}$.

Instead of $p_C^I(t)$, it is easier to calculate $1-p_C^I(t)$, the probability of non-completion. Starting from an incipient lineage, the three exclusive ways not to complete speciation are (figure 2, upper right panel): (c) direct extinction of the lineage within a time t, (d) survival of the lineage during a time t without completion, extinction or initiation, or (e) initiation of speciation after a time $u \leq t$ and non-completion of any of the daughter lineages within the remaining time $t-u$. The probability of speciation completion of an incipient lineage (or any of its potential descendant) within a time t thus satisfies the following equation:

$$
\begin{aligned}
1 - p_C^I(t) = & \frac{\mu_2}{\Lambda}(1 - e^{-\Lambda t}) && \text{figure 2(c)} \\
& + e^{-\Lambda t} && \text{figure 2(d)} \\
& + \int_0^t \lambda_3 e^{-\Lambda u}(1 - p_C^I(t-u))^2 \, \mathrm{d}u && \text{figure 2(e)}
\end{aligned} \tag{4}
$$

This equation can be solved numerically by solving an ordinary differential equation (appendix A2).

Starting now from a good lineage, the two exclusive possible ways leading to extinction as first event within a horizon t are (figure 2, bottom left panel): (f) the lineage directly goes extinct within a time t or (g) the lineage forms an incipient lineage and both daughter lineages (one good and one incipient) and their potential descendant die within a time $t-u$ without speciation. The probability for a good lineage and its potential descendants to go extinct within a time t thus satisfies the following equation:

$$
\begin{aligned}
p_E^G(t) = & \frac{\mu_1}{\Theta}(1 - e^{-\Theta t}) && \text{figure 2(f)} \\
& + \int_0^t \lambda_1 e^{-\Theta u} p_E^I(t-u) p_E^G(t-u) \, \mathrm{d}u. && \text{figure 2(g)}
\end{aligned} \tag{5}
$$

where $\Theta = \lambda_1 + \mu_1$.

This equation can be solved numerically, provided that we already have a solution for $p_E^I(t)$ (appendix A3).

Starting from a good lineage, the two exclusive ways in which speciation occurs within a time t are (figure 2, bottom right panel): (h) the lineage initiates speciation after a time $u \leq t$, and the incipient lineage completes speciation within the remaining time $t-u$, or (i) the lineage initiates speciation at a time $u \leq t$, the completion of this incipient lineage fails, and the good lineage speciates within the remaining time $t-u$. The probability that a good lineage fulfils speciation within a time t thus satisfies the following equation:

$$
\begin{aligned}
p_S^G(t) = \int_0^t & \lambda_1 e^{-\Theta u} \times && \text{initiation at time } u \\
& \left[p_C^I(t-u) \right. && \text{figure 2(h)} \\
& \left. + (1 - p_C^I(t-u)) p_S^G(t-u) \right] \mathrm{d}u && \text{figure 2(i)}
\end{aligned} \tag{6}
$$

This equation can be solved numerically (appendix A4).

For all numerical integrations, we used the module `SciPy` in Python (Virtanen et al. 2020).

After solving these equations numerically and using equation 2, we obtain the time-dependent equivalent birth rate $\hat{\lambda}(t)$ and death rate $\hat{\mu}(t)$.

2.3 Equivalent constant BD rates

Defined, as above, as the values such that the probabilities of speciation and extinction in an infinitesimal time interval $[t-\mathrm{d}t, t]$ are given by $\hat{\lambda}(t)\,\mathrm{d}t$ and $\hat{\mu}(t)\,\mathrm{d}t$, for any time t of the PBD process, $\hat{\lambda}(t)$ and $\hat{\mu}(t)$ indeed depend on time, even though the parameters of the PBD are constant. However, as we will show later and in agreement with previous results (Etienne and Rosindell 2012), the time-dependency is particularly manifest towards the present. Intuitively, given enough time, constant rates of initiation, completion and extinction result in constant equivalent speciation and extinction rates. However, towards the present (i.e., towards the end of the PDB process), incipient lineages did not have enough time to complete speciation, resulting in a decline in speciation rates. As one of our main goals here is to understand how speciation initiation, completion and extinction translate into macroevolutionary speciation and extinction rates, we now introduce equivalent constant BD rates, meant to represent the relationship between the parameters of the PBD and BD process far from the present.

We define equivalent constant BD rates, $\tilde{\lambda}$ (birth rate) and $\tilde{\mu}$ (death rate) as the parameters of the standard BD process such that the probability of speciation and the expected time to speciation match those under the PBD process with rates λ_1, λ_2, λ_3, μ_1, and μ_2, starting from a good lineage. The probability of extinction does not bring any additional information since extinction and speciation events are complementary over infinite time. Intuitively, we expect that the equivalent time-dependent birth and death rates tend towards $\tilde{\lambda}$ and $\tilde{\mu}$ in the past; the equivalent time-dependent birth rate then declines toward the present.

We shown that equivalent constant-time rates are given by the following expressions (see details appendix A5):

$$\tilde{\lambda} = (1-\pi)\lambda_1 \quad \text{and} \quad \tilde{\mu} = \mu_1 \tag{7}$$

where $\pi = \frac{\lambda_2+\lambda_3+\mu_2}{2\lambda_3}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{(\lambda_2+\lambda_3+\mu_2)^2}}\right)$ is the probability of non-completion of an incipient lineage.

Equation 7 shows that, far from the present, the equivalent death rate is exactly the rate of extinction of good species, and the equivalent birth rate is directly proportional to the rate of speciation initiation from a good species, with a coefficient of proportionality that represents the probability of completion without time horizon and depends on the rates specific to incipient lineages (initiation, completion and extinction). As in the case of the time-dependent equivalent rates, these rates correspond to the case of a process starting with a good lineage without incipient lineages, and the equivalent extinction rate may thus overestimate actual extinction probabilities under the PBD.

In order to better understand the influence of each parameter of the PBD process on the equivalent constant birth rate $\tilde{\lambda}$, we calculate partial derivatives of this function with respect to the different parameters. A high partial derivative with respect to a given parameter reflects a strong influence of this parameter on the equivalent birth rate,and therefore that the corresponding step of the speciation process may be limiting. We compute the relative influence of a given parameter as the ratio between the absolute

partial derivative with respect to this parameter and the sum of the absolute partial derivatives with respect to all other parameters. We perform these analyses both for the simplified PBD model where good and incipient lineages have the same rates of initiation ($\lambda_1 = \lambda_3$) and extinction ($\mu_1 = \mu_2$) and for the full PBD model. Detailed calculations are provided in appendix A6.

2.4 Simulations under the PBD process and equivalent BD processes

Although the equivalent rates were designed to construct BD processes that approach the PBD process, these processes are not identical. We used simulations to compare reconstructed trees (i.e., trees of extant species) generated by the PBD process and their equivalent BD processes. In each of these simulations, we consider the trees of extant good species, disregarding extinct and incipient lineages. We compared trees simulated under the constant-rate PBD model with trees simulated under the corresponding time constant (BD) and time-varying BD model (varBD). We independently varied each of the 5 parameters of the PBD model, each taking 5 values (the default value and 2 above, 2 below, with amplitude chosen to guarantee computational tractability), resulting in 25 parameter combinations. For each combination, we computed the trajectories of equivalent time-dependent birth and death rates over 15 million years – approximated by piecewise-constant birth and death rates over 200 intervals – and simulated 500 tree replicates using the R library `TreeSim` (Stadler 2011). The values of these rates are given in supplementary figure S4. Each tree simulation was conditioned on the survival of two extant lineages (up to the failure of 100 simulation attempts for each replicate), starting from a single stem branch. We also generated the same number of trees under the PBD model for each combination of parameters using the `PBD` package (Etienne and Rosindell 2012). Finally, we simulated trees under the BD model with equivalent constant rates, using the package `TreeSim`. We expect these simulations to deviate the most from those obtained under the PBD process. We compared the outputs of the simulations under the three models in terms of species richness at present (SR), tree shape and tree topology.

To analyze tree shape, we used the γ statistic (Pybus and Harvey 2000), computed with the package `ape` (Paradis and Schliep 2019). The γ statistics quantifies the relative position of the internal nodes of a tree and compares it to the expectations under a pure-birth (Yule) model. $\gamma > 0$ corresponds to trees where internal nodes are closer to the tips than expected under Yule's model, while $\gamma < 0$ corresponds to trees where internal nodes are closer to the root.

To analyze tree topology, we used the stairs2 balance index (Norström et al. 2012), computed with the package `treestats` (Janzen and Etienne 2024). This statistic measures the mean size ratio between the smaller and larger pending subtree for all vertices. Stairs2 is higher for trees with more balanced subtrees and lower for more imbalanced trees. The strais2 statistic has been shown to perform well (Khurana et al. 2023) and is less sensitive to tree size than other statistics such as Aldous' β (Aldous 2001).

2.5 Ability to recover equivalent constant BD rates by fitting the BD model to truncated trees

If we acknowledge that speciation in nature usually takes time, birth and death rates estimates obtained by fitting a constant-rate BD model to empirical reconstructed trees are hard to interpret. As noted above, under a constant rate PBD model, we expect equivalent birth rates to approach the equivalent constant birth rate $\tilde{\lambda}$ in the past, and to decline closer to the present. Hence, we can expect that speciation rate estimates obtained by fitting a BD model to the entire tree will have intermediate values below $\tilde{\lambda}$. However, fitting a BD model to to older parts of the tree should provide good estimates of the equivalent constant BD rates.

To test this expectation, we truncated the phylogenies simulated under the PBD process in section 2.4 at different time points in the past, and fitted a constant BD model to these truncated phylogenies, using the dedicated function `fit_bd_in_past` (Lewitus et al. 2018; Perez-Lamarque et al. 2022) from the R package `RPANDA` (Morlon, Lewitus, et al. 2016). We fitted a constant-rate birth-death model to both entire trees and truncated trees "sliced" at 17 regularly spaced time points between the present and 4 million years before the present. Finally, we compared the speciation and extinction rates estimates obtained with various truncation times to the analytically-derived equivalent constant BD rates.

3 Results

3.1 Equivalent time-dependent BD rates

We used equations 1 and 2 and numerical solutions of the equations describing the probabilities of speciation, completion and extinction with time to derive equivalent time-dependent birth and death rates (thereafter simply referred to as birth and death rates) for a large range of parameter values.

We find that birth rates decrease close to the present, reaching 0 at present (time $t \to T$, see figure 3). Death rates depend less on time and can be considered almost constant with time if we neglect a small decrease followed by a short increase when $t \to T$.In the past ($t \to 0$), birth and death rates converge to the values predicted by our analytical expression of the equivalent constant BD rates (equation 7). This provides indirect (graphical) evidence that the constant equivalent birth-death rates can be considered as asymptotic rates of the time-dependent equivalent birth-death rates.

Initiation rates have a strong effect on the birth rate (figure 3, panel **A**). In the past, birth rates converge to values scaling with the initiation rate, all other rates being equal, as expected from equation 7. As expected, lower completion rates result in lower birth rates, and an effect of the protracted nature of speciation that extends further into the past (panel **C**). In the limit $\lambda_2 \to \infty$, the model converges to a pure BD model with constant rates, except very close to the present. Indeed, with high completion rates, incipient lineages complete speciation very fast and with high probability, so speciation occurs as soon as an initiation event occurs. Finally, birth rates are lower when

extinction rates are higher, except closer to the present where the effect of extinction diminishes (panel **E**). This effect is entirely due to the extinction of incipient lineages (μ_2), as the extinction rate of good lineages (μ_1) has virtually no effect on the birth rate (supplementary figure S2). The higher extinction rate of incipient lineages renders speciation less likely, as incipient species more often go extinct before completing speciation.

Death rates match closely extinction rates (figure 3, panel **F**), entirely due to the effect of the extinction rate of good lineages, as expected from equation 7; indeed, the extinction rate of incipient lineages has no effect on the death rate (supplementary figure S2).

3.2 Equivalent constant BD rates

As shown by equation 7 and already described above, the equivalent constant death rate equals the extinction rate of good species; the equivalent constant birth rate scales with speciation initiation, it increases with the completion rate and decreases with the extinction rate of incipient lineages (figure 3, dashed lines). We better characterized the influence of each step of the speciation process (i.e., initiation, survival of incipient species, and completion) on $\tilde{\lambda}$, interpreted as the macroevolutionary speciation rate, by computing relative partial derivatives. This allows us to identify the steps of the PBD process that may limit macroevolutionary speciation rates figure 4. We identify that speciation initiation is limiting when its rate is low and the completion rate is high (regions 1 and 2). Speciation completion is limiting when its rate is low compared to the other parameters (regions 3 and 5). An increase in the population extinction rate has most effects when this rate is low (region 6) and when initiation rate is high compared to the completion rate (regions 4 and 5).

3.3 Trees generated by the PBD process and equivalent BD processes

We compared the size, shape and topology of trees generated under the PBD model to those of trees generated under their equivalent time-constant and time-varying BD models (figure 5). As expected, tree size (SR) increases with higher rates of speciation initiation and completion, and decreases with higher rates of extinction of good and incipient lineages (figure 5, top row). These trends are well captured by both the equivalent time-varying and time-constant models, with the exception of the increase in species richness with the rate of initiation from incipient lineages λ_3. Compared to the PBD model, the equivalent BD models produce larger trees when λ_3 is small, and smaller trees when λ_3 is large. λ_3 has little influence on the size of trees generated under equivalent BD models, consistently with the weak influence of λ_3 on the equivalent rates (see supplementary figure S3**C** and **D**).

As expected given unachieved speciation close to the present, the PBD process results in trees with negative γ values (reflecting long terminal branches), unless the completion or extinction rates are high (figure 5, middle row). γ decreases with increasing ratios of speciation initiation to speciation completion rates, i.e. when the protracted

nature of the speciation process is more pronounced). The equivalent variable-rate BD model captures these trends relatively well, while the constant-rate BD model produces trees with positive γ values, indicating nodes closer to the tips. The drop towards 0 in the equivalent time-variable birth rates captures the shortage of speciation events close to the present induced by the protracted process, except when the rate of initiation from an incipient lineage (λ_3) gets large. In this case the PBD process produces trees with increasingly long branch lengths, while the distribution of nodes generated under the equivalent time-dependent BD model remains stable.

Under most PBD parameters, stairs2 values of tree topology are stable, close to 0.65, and well reproduced by both the time-constant and the time-variable equivalent BD models. Higher values of stairs2 (reflecting more balanced trees) are observed in parameter ranges leading to small trees, and likely reflect tree size rather than a true difference in tree balance.

Deviations observed between trees simulated under the PBD process and those simulated under the equivalent time-varying BD model under some λ_3 values likely come from the approximation we made when establishing the link between equivalent rates and the probabilities of speciation and extinction under the PBD (see section 2.2). If, as expected, the equivalent extinction rate overestimate extinction probabilities, this explains why we obtain smaller trees with less negative gamma values, as extinction pushes nodes towards the present (the "pull of the present", Nee et al. 1994). This is especially pronounced if λ_3 is high when the approximation is the strongest, as there are many incipient lineages that are not accounted for (the ones that originated before the time at which the rates are calculated).

3.4 Recovery of equivalent BD rates by fit to truncated PBD trees

Our expectation that fits of a constant rate BD model to the "old" part of trees generated under a PBD process would provide good estimates of the equivalent BD rates was verified (supplementary figures S7a and S7b). When fitting a constant rate BD model to the entire tree (i.e., when the truncation time is zero), the estimated speciation and extinction rates are well below the expected equivalent constant rates. However when truncation time increases, these estimates converge to the expected equivalent rates. In the case of the extinction rate, estimates remain slightly below the expected equivalent rate, supporting our intuition that equivalent extinction rates should overestimate actual extinction events. The convergence occurs with a truncation time relatively close to the present, consistent with the observed time of decline of the equivalent time-dependent BD rates. This recent truncation time appears optimal, as estimates are not as good when more of the tree is truncated, probably due to a loss of statistical power with decreasing data size.

4 Discussion

In this study we derived predictions, under the protracted birth-death (PBD) model of diversification, for "equivalent" birth and death rates, meant to represent macroevolu-

tionary speciation and extinction rates as we may estimate them from reconstructed phylogenies. The protracted birth-death model accounts for the fact that speciation is not an instantaneous process and decorrelates the process of speciation initiation and speciation completion. The first process can be understood by a rate at which a lineage creates incipient lineages and the second process is characterized by a completion rate, i.e. a rate at which these lineages will become distinct species (denoted as *good* species in the model). We showed that the equivalent rates – in particular the birth rate – vary when the process approaches the present, but can be considered constant when far from the present. Our analytical predictions of the rates in the past allowed us to explore the importance of each step of the speciation process (i.e. initiation, survival, completion) in modulating the macroevolutionary speciation rate. In addition, we showed that macroevolutionary speciation rates estimated on truncated reconstructed trees can provide good estimates of the equivalent birth rate. This opens the possibility to relate estimates of macroevolutionary speciation rates, which have major consequences for the build up of diversity over geological time scales, to the microevolutionary processes that modulate speciation initiation, population survival, and speciation completion.

As expected given previous results on the protracted birth-death process (Etienne and Rosindell 2012), we find that equivalent birth rates decline to zero as time approaches the present. Close to the present, speciation is less likely to occur because it requires a delay. The decay in equivalent birth rate starts earlier when the completion rate is lower. Trees simulated under the equivalent time-dependent BD model have similar characteristics to trees simulated under the PBD process, in terms of tree size and shape (figure 5), meaning that the equivalent BD process captures the dynamics of speciation and extinction induced by the protracted model relatively well. It is however important to remember that the PBD and equivalent BD processes are not entirely interchangeable. The PBD process is a process with memory, which induces an age dependency of speciation and extinction rates that is not captured by the equivalent BD process. For example "old" species are more likely to have accumulated incipient lineages, which buffer their extinction risk, but this cannot be captured by an age-independent BD process. More generally, the approximations made in our equivalent BD process lead to an overestimation of extinction events occuring under the PBD process, explaining some of the deviations observed when comparing simulated trees, in particular for high values of the rate of initiation from incipient lineages (λ_3).

Before equivalent speciation rates start to drop, they can be considered virtually constant. In this regime, we derived analytical relationships between the equivalent birth and death rates and the parameters of the PBD process. These relationships show that we can expect macroevolutionary extinction rates to provide a relatively good approximation of the rate at which species go extinct, except if the rate of initiation from incipient lineages is large. They also show that macroevolutionary speciation rates are directly proportional to speciation initiation rates, with a coefficient of proportionality that depends on the rates of completion, initiation and extinction of incipient lineages. Hence, all aspects of the speciation process play a role in modulating macroevolutionary speciation rates, but some play a larger role than others, and which ones will likely depend on the ecology, genetics and biogeography of the species group considered. Indeed, by identifying regimes within which each aspect of the speciation process is

expected to be the most influential (figure 4), we found that the rates of speciation initiation and population extinction often are the limiting factors. Completion rates are only limiting when they are very small and at intermediate values of speciation initiation rates, or high population extinction rates. Hence, a species that accumulates fast reproductive isolation will be characterized by a higher completion rate but this will not necessarily have a strong effect on the speciation rate (figure 3**C**).

Our results have implications for the phylogenetic analysis of diversification. Indeed, ignoring the fact that speciation takes time by fitting standard BD models to empirical phylogenies necessarily leads to model misspecification (e.g. when fitting a constant rate BD model) or misinterpretation (e.g. when interpreting the effect of protraction as a speciation rate decline). In the recent years, there has been a surge in the use of tip-rate speciation estimates, i.e. estimates of speciation rates at present across the tips of a phylogeny, obtained with statistics such as the diversification rate (DR; Jetz et al. 2012) statistic, or models such as Bayesian analysis of macroevolutionary mixtures (BAMM; Rabosky 2014), the cladogenetic diversification shift (ClaDS; Maliet, Hartig, et al. 2019; Maliet and Morlon 2021), or birth-death diffusion (BDD; Quintero et al. 2024) models. Future work investigating how speciation rates estimated with these methods on trees generated under the PBD process compare with equivalent rates will be useful. However the fact that equivalent birth rates converge to zero close to the present suggests that tip rate estimates provide a poor representation of speciation probabilities. Speciation rates estimated on deeper parts of the phylogeny approximate equivalent rates quite well, suggesting that these estimates indeed reflect the macroevolutionary outcome of the combined processes of initiation, survival and completion. Ideally, we would need to account for the protracted nature of speciation in every model used to infer diversification rates from phylogenies. While progress is being made in this direction (Hua et al. 2022), it adds substantial computational complexity. A workaround is to truncate phylogenies; however this requires employing inference methods that account for this truncation. These have been implemented for only a limited set of models, such as the constant-rate, time-dependent and environment-dependent models developed in `RPANDA` (Morlon, Lewitus, et al. 2016). More systematically implementing this truncation option in diversification models would be useful. Our tests indicate that truncating phylogenetic trees at intermediate time points yields the most accurate estimates of equivalent rates. As the optimal truncation time will depend on the extent to which speciation is protracted (i.e., the completion rate), we recommend to apply truncation at various time points and chose the time at which the estimated rates reach a plateau.

Our results also have implications for the ongoing effort to relate macroevolutionary speciation rates to microevolutionary processes (Rolland et al. 2023; Harvey et al. 2019; Rabosky 2016). Indeed, microevolutionary processes act individually on each step of the speciation process (initiation, survival, completion), and our expressions of equivalent rates provide an idea of how these combine to modulate macroevolutionary speciation rates. For example, the apparent absence of correlation between tip rate estimates of speciation and the speed at which reproductive isolation is acquired (Rabosky and Matute 2013; Freeman et al. 2022) is not surprising given our expectation that completion rates have a limited effect on the macroevolutionary speciation rate (and

the limitations of tip rate estimates mentioned before). We however expect to find a correlation between macroevolutionary speciation rates and the rate of population formation, which reflects speciation initiation. This is indeed what was found by Harvey et al. 2019; the lack of correlation in other studies (Singhal, Colli, et al. 2022; Singhal, Huang, et al. 2018; Burbrink et al. 2023) may reflect cases when speciation is modulated by the rate of population extinction rather than by the rate of speciation initiation, as expected if the initiation rate is high and/or the population extinction rate is low.

By accounting for the time it takes to complete speciation, protracted birth-death models provide more biologically realistic models than standard birth-death models, and a theoretical framework for understanding how each step of the speciation processe influences macroevolutionary speciation rates. A more systematic account of the protracted nature of speciation in phylogenetic analyses of diversification would both improve our estimates of diversification rates and our understanding of how microevolutionary processes combine to modulate macroevolutionary speciation rates.

Code availability

The scripts used to make these analyses are available on https://github.com/pierre-veron/PBD_analog.

Aknowledgement

We thank Thibault Juillard, Amandine Véber and Amaury Lambert for mathematical advice; and Benoît Perez-Lamarque who kindly shared his code to estimate the BD rates on truncated phylogenies.

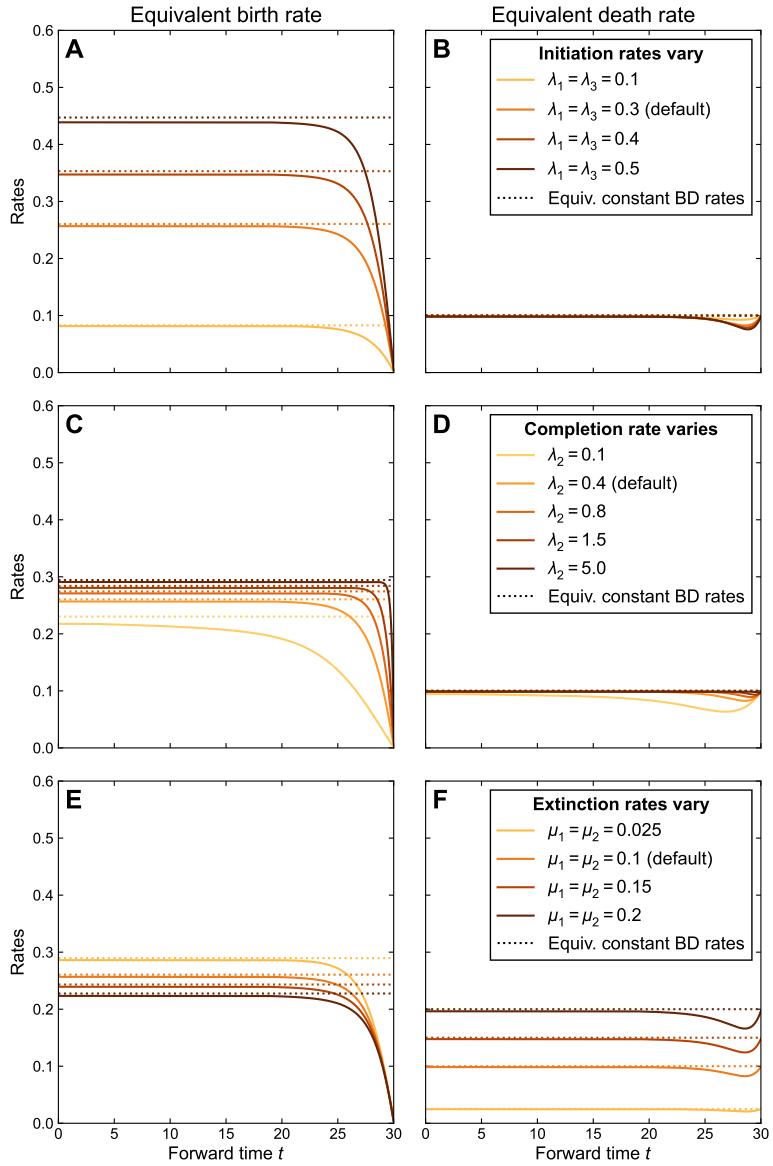

Figure 3. Influence of the parameters of the protracted birth-death (PBD) process on equivalent birth-death (BD) rates. Solid lines represent equivalent BD rates derived from equations 1 and 2 as a function of time for different values of PBD rates. Dotted lines represent constant equivalent BD rates derived analytically from equation 7 for the same PBD parameters. In the top/middle/bottom row the initiation/completion/extinction rates vary, with the other rates constant (default values are $\lambda_1 = \lambda_3 = 0.3, \lambda_2 = 0.4, \mu_1 = \mu_2 = 0.1$). In figures S2 and S3, we calculated the same rates with the 5 parameters varying independently. $t = 30$ corresponds to the present, $t = 0$ to the past.

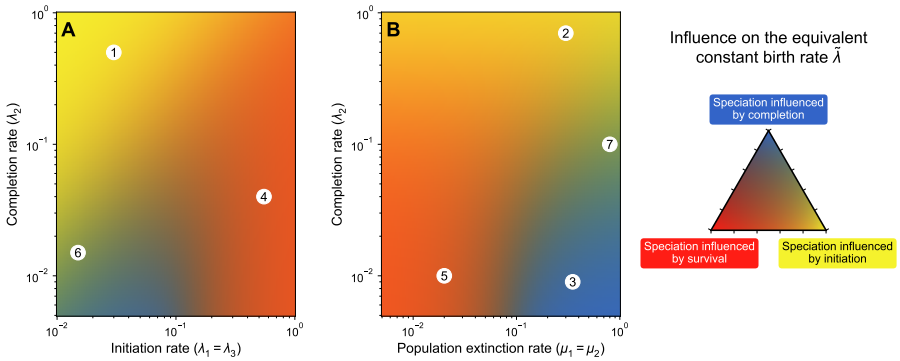

Figure 4. Relative influence of the parameters of the PBD model on the equivalent constant birth rate. Colors indicate which of the PBD process among initiation, completion and population extinction limits the equivalent constant birth rate $\tilde{\lambda}$ most, as a function of (**A**) the initiation and completion rates and (**B**) the population extinction and completion rates, with the color code explained on the triangle in the right. A yellow region (e.g. 1 or 2) indicates a combination of parameters where the most influential parameter on the birth rate is the rate of initiation. A blue region (e.g. 3) indicates a combination of parameters where the most influential parameter is the rate of completion λ_2. A orange region (e.g. 4 or 5) indicates a combination of parameters where both population extinction and speciation initiation are influential. Green regions (e.g. 6 or 7) indicate situations where both speciation initiation and completion have a similar influence on the birth rate. In all cases, the rates of initiation and completion have a positive influence on the birth rate and the rate of population extinction has a negative influence. The detailed methods are explained in appendix A6 and the values of the relative influence are provided in supplementary figure S5. When they do not vary, default values of the parameters are 0.1.

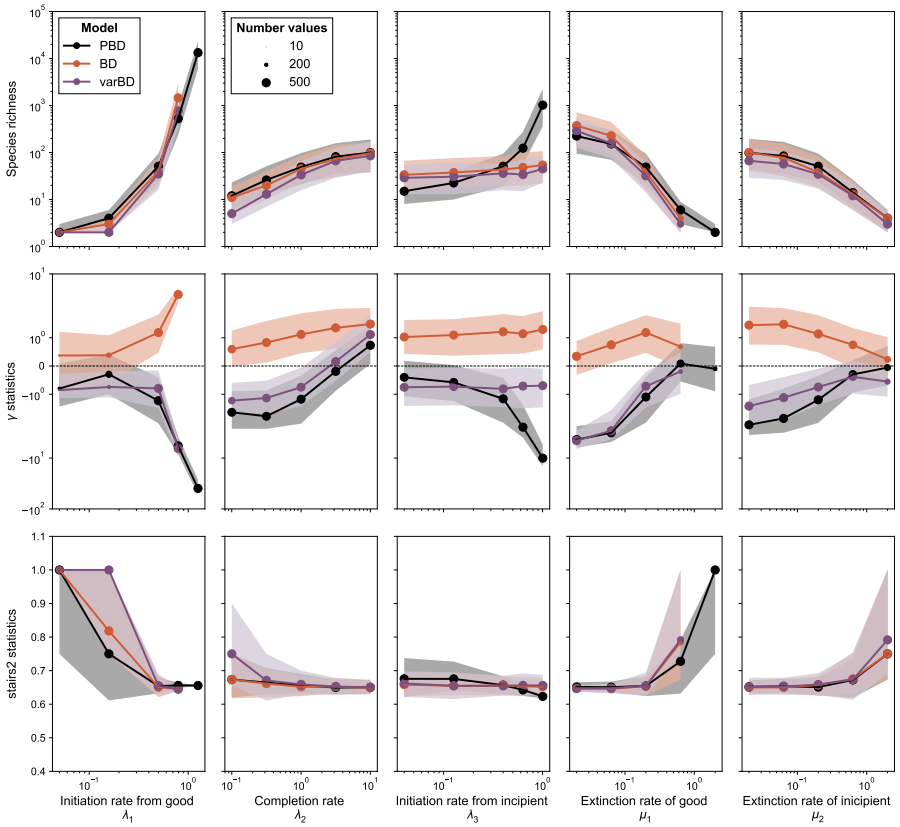

Figure 5. Statistics of trees generated under the protracted birth-death (PBD) process and its equivalent birth-death (BD) processes. By row: species richness SR, γ shape statistic and stairs2 balance index for trees generated under the three models (PBD, equivalent constant rate BD and equivalent time varying rate BD) for different values of the parameters of the PBD model. In each column, only one of the PBD parameters varies, with the others held constant (default $\lambda_1 = 0.5, \lambda_2 = 1.0, \lambda_3 = 0.4, \mu_1 = 0.2, \mu_2 = 0.2$). For each set of parameters of the PBD model, BD and varBD trees were generated under equivalent birth and death rates computed using equations 1, 2 and 7. The line corresponds to the median of statistics across all 200 replicates, the shaded area indicates the first and last quartile of the statistics. The size of the dots indicates the number of valid data for which the statistics could be calculated.

References

Aldous, David J. (2001). "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today". In: *Statistical Science* 16.1, pp. 23–34. DOI: 10.1214/ss/998929474.

Aristide, Leandro and Hélène Morlon (2019). "Understanding the effect of competition during evolutionary radiations: an integrated model of phenotypic and species diversification". In: *Ecology Letters* 22.12, pp. 2006–2017. DOI: 10.1111/ele.13385.

Benton, Michael J. and Paul N. Pearson (2001). "Speciation in the fossil record". In: *Trends in Ecology and Evolution* 16.7, pp. 405–411. DOI: 10.1016/s0169-5347(01)02149-8.

Burbrink, Frank T., Sara Ruane, Nirhy Rabibisoa, Achille P. Raselimanana, Christopher J. Raxworthy, and Arianna Kuhn (2023). "Speciation rates are unrelated to the formation of population structure in Malagasy gemsnakes". In: *Ecology and Evolution* 13.8, e10344. DOI: 10.1002/ece3.10344.

Coyne, Jerry A. and H. Allen Orr (2004). *Speciation*. Sunderland, Massachusetts: Sinauer Associates, Inc. Publ. 545 pp. ISBN: 0878930892.

Dynesius, Mats and Roland Jansson (2014). "Persistence of within-species lineages: A neglected control of speciation rates". In: *Evolution* 68.4, pp. 923–934. DOI: 10.1111/evo.12316.

Etienne, Rampal S., Hélène Morlon, and Amaury Lambert (2014). "Estimating the duration of speciation from phylogenies". In: *Evolution* 68.8, pp. 2430–2440. DOI: 10.1111/evo.12433.

Etienne, Rampal S. and James Rosindell (2012). "Prolonging the Past Counteracts the Pull of the Present: Protracted Speciation Can Explain Observed Slowdowns in Diversification". In: *Systematic Biology* 61.2, pp. 204–213. DOI: 10.1093/sysbio/syr091.

Freeman, Benjamin G., Jonathan Rolland, Graham A. Montgomery, and Dolph Schluter (2022). "Faster evolution of a premating reproductive barrier is not associated with faster speciation rates in New World passerine birds". In: *Proceedings of the Royal Society B* 289.1966, p. 20211514. DOI: 10.1098/rspb.2021.1514.

Harvey, Michael G., Sonal Singhal, and Daniel L. Rabosky (2019). "Beyond Reproductive Isolation: Demographic Controls on the Speciation Process". In: *Annual Review of Ecology, Evolution, and Systematics* 50.1, pp. 75–95. DOI: 10.1146/annurev-ecolsys-110218-024701.

Hua, Xia, Tyara Herdha, and Conrad J. Burden (2022). "Protracted Speciation under the State-Dependent Speciation and Extinction Approach". In: *Systematic Biology* 71 (6), pp. 1362–1377. DOI: 10.1093/sysbio/syac041.

Janzen, Thijs and Rampal S. Etienne (2024). "Phylogenetic tree statistics: a systematic overview using the new R package 'treestats'". In: *bioRxiv*. DOI: 10.1101/2024.01.24.576848.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers (2012). "The global diversity of birds in space and time". In: *Nature* 491.7424, pp. 444–448. DOI: 10.1038/nature11631.

Khurana, Mark P., Neil Scheidwasser-Clow, Matthew J. Penn, Samir Bhatt, and David A. Duchêne (2023). "The Limits of the Constant-rate Birth–Death Prior for Phylo-

genetic Tree Topology Inference". In: *Systematic Biology* 73 (1), pp. 235–246. DOI: 10.1093/sysbio/syad075.

Lewitus, Eric, Lucie Bittner, Shruti Malviya, Chris Bowler, and Hélène Morlon (2018). "Clade-specific diversification dynamics of marine diatoms since the Jurassic". In: *Nature Ecology & Evolution* 2.11, pp. 1715–1723. DOI: 10.1038/s41559-018-0691-3.

Maliet, Odile, Florian Hartig, and Hélène Morlon (2019). "A model with many small shifts for estimating species-specific diversification rates". In: *Nature Ecology & Evolution* 3.7, pp. 1086–1092. DOI: 10.1038/s41559-019-0908-0.

Maliet, Odile and Hélène Morlon (2021). "Fast and Accurate Estimation of Species-Specific Diversification Rates Using Data Augmentation". In: *Systematic Biology* 71.2, pp. 353–366. DOI: 10.1093/sysbio/syab055.

Moen, Daniel and Hélène Morlon (2014). "Why does diversification slow down?" In: *Trends in Ecology and Evolution* 29.4, pp. 190–197. DOI: 10.1016/j.tree.2014.01.010.

Morlon, Hélène, Jérémy Andréoletti, Joëlle Barido-Sottani, Sophia Lambert, Benoît Perez-Lamarque, Ignacio Quintero, Viktor Senderov, and Pierre Veron (2024). "Phylogenetic Insights into Diversification". In: *Annual Review of Ecology, Evolution, and Systematics* 55, pp. 1–21. DOI: 10.1146/annurev-ecolsys-102722-020508.

Morlon, Hélène, Eric Lewitus, Fabien L. Condamine, Marc Manceau, Julien Clavel, and Jonathan Drury (2016). "RPANDA: an R package for macroevolutionary analyses on phylogenetic trees". In: *Methods in Ecology and Evolution* 7.5, pp. 589–597. DOI: 10.1111/2041-210x.12526.

Nee, Sean, Edward C. Holmes, Robert M. May, and Paul H. Harvey (1994). "Extinction rates can be estimated from molecular phylogenies". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1307, pp. 77–82. DOI: 10.1098/rstb.1994.0054.

Norström, Melissa M., Mattia C.F. Prosperi, Rebecca R. Gray, Annika C. Karlsson, and Marco Salemi (2012). "PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences". In: *Evolutionary Bioinformatics* 8, pp. 261–269. DOI: 10.4137/ebo.s9738.

Paradis, Emmanuel and Klaus Schliep (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35.3, pp. 526–528. DOI: 10.1093/bioinformatics/bty633.

Perez-Lamarque, Benoît, Maarja Öpik, Odile Maliet, Ana C. Afonso Silva, Marc-André Selosse, Florent Martos, and Hélène Morlon (2022). "Analysing diversification dynamics using barcoding data: The case of an obligate mycorrhizal symbiont". In: *Molecular Ecology* 31.12, pp. 3496–3512. DOI: 10.1111/mec.16478.

Pybus, Oliver G. and Paul H. Harvey (2000). "Testing macro–evolutionary models using incomplete molecular phylogenies". In: *Proceedings of the Royal Society B* 267.1459, pp. 2267–2272. DOI: 10.1098/rspb.2000.1278.

Quintero, Ignacio, Nicolas Lartillot, and Hélène Morlon (2024). "Imbalanced speciation pulses sustain the radiation of mammals". In: *Science* 384.6699, pp. 1007–1012. DOI: 10.1126/science.adj2793.

Rabosky, Daniel L. (2014). "Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees". In: *PLOS ONE* 9.2, e89543. DOI: 10.1371/journal.pone.0089543.

– (2016). "Reproductive isolation and the causes of speciation rate variation in nature". In: *Biological Journal of the Linnean Society* 118.1, pp. 13–25. DOI: 10.1111/bij.12703.

Rabosky, Daniel L. and Daniel R. Matute (2013). "Macroevolutionary speciation rates are decoupled from the evolution of intrinsic reproductive isolation in Drosophila and birds". In: *Proceedings of the National Academy of Sciences* 110.38, pp. 15354–15359. DOI: 10.1073/pnas.1305529110.

Rolland, Jonathan et al. (2023). "Conceptual and empirical bridges between micro- and macroevolution". In: *Nature Ecology & Evolution* 7 (8), pp. 1181–1193. DOI: 10.1038/s41559-023-02116-7.

Silvestro, Daniele, Jan Schnitzler, Lee Hsiang Liow, Alexandre Antonelli, and Nicolas Salamin (2014). "Bayesian Estimation of Speciation and Extinction from Incomplete Fossil Occurrence Data". In: *Systematic Biology* 63.3, pp. 349–367. DOI: 10.1093/sysbio/syu006.

Singhal, Sonal, Guarino R. Colli, Maggie R. Grundler, Gabriel C. Costa, Ivan Prates, and Daniel L. Rabosky (2022). "No link between population isolation and speciation rate in squamate reptiles". In: *Proceedings of the National Academy of Sciences* 119.4, e2113388119. DOI: 10.1073/pnas.2113388119.

Singhal, Sonal, Huateng Huang, Maggie R. Grundler, María R. Marchán-Rivadeneira, Iris Holmes, Pascal O. Title, Stephen C. Donnellan, and Daniel L. Rabosky (2018). "Does Population Structure Predict the Rate of Speciation? A Comparative Test across Australia's Most Diverse Vertebrate Radiation". In: *The American Naturalist* 192.4, pp. 432–447. DOI: 10.1086/699515.

Stadler, Tanja (2011). "Simulating Trees with a Fixed Number of Extant Species". In: *Systematic Biology* 60.5, pp. 676–684. DOI: 10.1093/sysbio/syr029.

– (2013). "Recovering speciation and extinction dynamics based on phylogenies". In: *Journal of Evolutionary Biology* 26.6, pp. 1203–1219. DOI: 10.1111/jeb.12139.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Appendices

A1 Resolution of the probability of extinction of an incipient lineage

In equation 3, let us do the variable change $v = t - u$ to get rid of the variable t in the integral:

$$\begin{aligned} p_E^I(t) &= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + \int_0^t \lambda_3 e^{-\Lambda(t-v)}(p_E^I(v))^2 \,\mathrm{d}v \\ &= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + \lambda_3 e^{-\Lambda t}\int_0^t e^{\Lambda v}(p_E^I(v))^2 \,\mathrm{d}v \\ &= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + \lambda_3 e^{-\Lambda t} f(t) \end{aligned}$$

with f defined with, for all $t \geq 0$:

$$f(t) := \int_0^t e^{\Lambda v}(p_E^I(v))^2 \,\mathrm{d}v.$$

We note that

$$\begin{aligned} f'(t) &= e^{\Lambda t}(p_E^I(t))^2 \\ &= e^{\Lambda t}\left(\frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + \lambda_3 e^{-\Lambda t} f(t)\right)^2. \end{aligned}$$

This is a non-linear ordinary differential equation (ODE) of first order, with the initial condition $f(0) = 0$. With the help of Wolfram Mathematica we have the solution:

$$f(t) = e^{\Lambda t}\left[\frac{1}{\lambda_3^2\left(ce^{kt} + \frac{1}{k}\right)} - \frac{\Lambda(k-\Lambda) + 2\mu_2\lambda_3}{2\lambda_3^2\Lambda}\right] + \frac{\mu_2}{\lambda_3\Lambda}$$

with $k = \sqrt{\Lambda^2 - 4\lambda_3\mu_2}$ and $c = \frac{2}{k-\Lambda} - \frac{1}{\Lambda}$ a constant to satisfy the initial condition.

The probability of extinction $p_E^I(t)$ can be retrieved:

$$\begin{aligned} p_E^I(t) &= \sqrt{e^{-\Lambda t} f'(t)} \\ &= \frac{1}{\lambda_3}\sqrt{\frac{c(\Lambda-k)e^{kt} + \Lambda/k}{(ce^{kt} + 1/k)^2} - \frac{\Lambda(k-\Lambda) + 2\mu_2\lambda_3}{2}}. \end{aligned}$$

A2 Resolution of the probability of completion of an incipient lineage

To get rid of the variable t in the integral of equation 4, we do the change of variable $v = t - u$. This gives us:

$$\begin{aligned} 1 - p_C^I(t) &= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + e^{-\Lambda t} + \int_0^t \lambda_3 e^{-\Lambda(t-v)}\left(1 - p_C^I(v)\right)^2 \mathrm{d}v \\ &= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + e^{-\Lambda t} + e^{-\Lambda t}\int_0^t \lambda_3 e^{\Lambda v}\left(1 - p_C^I(v)\right)^2 \mathrm{d}v \end{aligned}$$

we now do a logarithmic change of variable in the integral, $z = e^{\Lambda v}$:

$$= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + e^{-\Lambda t} + e^{-\Lambda t}\int_1^{e^{\Lambda t}} \lambda_3 z \left(1 - p_C^I\left(\frac{\log z}{\Lambda}\right)\right)^2 \frac{1}{\Lambda z}\,\mathrm{d}z$$

$$= \frac{\mu_2}{\Lambda}(1-e^{-\Lambda t}) + e^{-\Lambda t}\left(1 + \frac{\lambda_3}{\Lambda} g\left(e^{\Lambda t}\right)\right)$$

with, for $x \geq 1$:

$$g(x) := \int_1^x \left(1 - p_C^I\left(\frac{\log z}{\Lambda}\right)\right)^2 \mathrm{d}z.$$

Noting that $g'(x) = [1 - p_C^I(\log x/\Lambda)]^2$, we have:

$$g'(x) = \left[\frac{\mu_2}{\Lambda}\left(1 - \frac{1}{x}\right) + \frac{1}{x}\left(1 + \frac{\lambda_3}{\Lambda} g(x)\right)\right]^2.$$

This is a non-linear ODE. With the initial condition $g(1) = 0$, it can be solved numerically and we retrieve the probability of completion of an incipient lineage with

$$p_C^I(t) = 1 - \sqrt{g'\left(e^{\Lambda t}\right)}.$$

A3 Resolution of the probability of extinction of a good lineage

From equation 5 we do the same change of variable as in the previous part $v = t - u$:

$$p_E^G(t) = \frac{\mu_1}{\Theta}(1-e^{-\Theta t}) + \lambda_1 e^{-\Theta t}\int_0^t e^{\Theta v} p_E^I(v) p_E^G(v)\,\mathrm{d}v$$

then with $z = e^{\Theta v}$ we have:

$$= \frac{\mu_1}{\Theta}(1-e^{-\Theta t}) + \lambda_1 e^{-\Theta t}\int_1^{e^{\Theta t}} z p_E^I\left(\frac{\log z}{\Theta}\right) p_E^G\left(\frac{\log z}{\Theta}\right)\frac{1}{\Theta z}\,\mathrm{d}z$$

$$= \frac{\mu_1}{\Theta}(1-e^{-\Theta t}) + \frac{\lambda_1}{\Theta} e^{-\Theta t}\int_1^{e^{\Theta t}} p_E^I\left(\frac{\log z}{\Theta}\right) p_E^G\left(\frac{\log z}{\Theta}\right)\mathrm{d}z$$

$$= \frac{\mu_1}{\Theta}(1-e^{-\Theta t}) + \frac{\lambda_1}{\Theta} e^{-\Theta t} h(e^{\Theta t})$$

with, for $x \geq 1$

$$h(x) = \int_1^x p_E^I\left(\frac{\log z}{\Theta}\right) p_E^G\left(\frac{\log z}{\Theta}\right)\mathrm{d}z.$$

We note that

$$h'(x) = p_E^I\left(\frac{\log x}{\Theta}\right) p_E^G\left(\frac{\log x}{\Theta}\right)$$

$$= p_E^I\left(\frac{\log x}{\Theta}\right) \times \left[\frac{\mu_1}{\Theta}\left(1 - \frac{1}{x}\right) + \frac{\lambda_1}{\Theta}\frac{1}{x} h(x)\right]. \tag{8}$$

With the condition $h(1) = 0$, equation 8 gives the ODE satisfied by h. If we solve this equation numerically we can retrieve $p_E^G(t)$:

$$p_E^G(t) = \frac{h'(e^{\Theta t})}{p_E^I(t)}.$$

A4 Resolution of the probability of speciation of a good lineage

From equation 6, we do the same change of variable $v = t - u$, then $z = e^{\Theta v}$:

$$\begin{aligned}
p_S^G(t) &= \lambda_1 e^{-\Theta t} \int_0^t e^{\Theta v} \left[p_C^I(v) + (1 - p_C^I(v)) p_S^G(v) \right] \mathrm{d}v \\
&= \lambda_1 e^{-\Theta t} \int_1^{e^{\Theta t}} z \left[p_C^I\left(\frac{\log z}{\Theta}\right) + \left(1 - p_C^I\left(\frac{\log z}{\Theta}\right)\right) p_S^G\left(\frac{\log z}{\Theta}\right) \right] \frac{1}{\Theta z} \mathrm{d}z \\
&= \frac{\lambda_1}{\Theta} e^{-\Theta t} \int_1^{e^{\Theta t}} \left[p_C^I\left(\frac{\log z}{\Theta}\right) + \left(1 - p_C^I\left(\frac{\log z}{\Theta}\right)\right) p_S^G\left(\frac{\log z}{\Theta}\right) \right] \mathrm{d}z \\
&= \frac{\lambda_1}{\Theta} e^{-\Theta t} m(e^{\Theta t})
\end{aligned}$$

with, for $x \geq 0$

$$m(x) := \int_1^x \left[p_C^I\left(\frac{\log z}{\Theta}\right) + \left(1 - p_C^I\left(\frac{\log z}{\Theta}\right)\right) p_S^G\left(\frac{\log z}{\Theta}\right) \right] \mathrm{d}z.$$

We note that

$$m'(x) = p_C^I\left(\frac{\log x}{\Theta}\right) + \left(1 - p_C^I\left(\frac{\log x}{\Theta}\right)\right) p_S^G\left(\frac{\log x}{\Theta}\right) \tag{9}$$

$$= p_C^I\left(\frac{\log x}{\Theta}\right) + \left(1 - p_C^I\left(\frac{\log x}{\Theta}\right)\right) \frac{\lambda_1}{\Theta} \frac{1}{x} m(x). \tag{10}$$

With the initial condition $m(1) = 0$, the equation 10 gives the ODE satisfied by m. If we solve this equation numerically we can retrieve $p_S^G(t)$ with the derivative of the solution m, from equation 9:

$$p_S^G(t) = \frac{m'(e^{\Theta t}) - p_C^I(t)}{1 - p_C^I(t)}.$$

We show an example of the obtained numerical solutions of p_E^I, p_C^I, p_E^G and p_S^G and empirical probabilities from simulations in supplementary figure S1.

A5 Calculation of the time-constant BD rates under the PBD model

Speciation probability and expected time for speciation under the BD model

Under the BD model with parameters λ and μ, the probability of speciation of a given lineage is the probability that a birth event occurs before a death event, *i.e.*

$$\mathbb{P}(\text{speciation}) = \frac{\lambda}{\lambda + \mu}. \tag{11}$$

Conditionally on speciation, the expected time T it takes for a given lineage to speciate is

$$\mathbb{E}[T|\text{speciation}] = \frac{1}{\lambda + \mu} \tag{12}$$

since T has the same distribution as $\min(X,Y)$ with $X \hookrightarrow \mathscr{E}(\lambda)$ and $Y \hookrightarrow \mathscr{E}(\mu)$ two independent clock variables representing speciation and extinction events. The minimum of two exponential independent processes is distributed exponentially with a parameter equal to the sum of the rates.

Speciation probability and expected time for speciation under the PBD model

Under the PBD model, a good lineage speciates if it generates at least one incipient lineage and one of these incipient lineages or one of the descendants of these lineages completes speciation.

Considering one incipient lineage L that has still no descendent, there are two outcomes: (*i*) the lineage or one of its descendent at least complete speciation (we will denote this event as "complete"), or (*ii*) nor the lineage or any of its descendent complete speciation before dying (we will denote this event as "does not complete").

Let us denote $\pi := \mathbb{P}(L \text{ does not not complete})$ and N the number of direct descendants of the lineage before it dies or completes speciation, and L and $L_1, \ldots, L_N$ those lineages. All the possible events that can happen to L (completion at rate λ_2, new incipient lineage at rate λ_3 or extinction at rate μ_2) are independent point process. The first one to happen is therefore a point process with rate $\lambda_2 + \lambda_3 + \mu_2$ and the probability that it is the formation of an incipient lineage is $\lambda_3/(\lambda_2 + \lambda_3 + \mu_2)$. We can thus decompose:

$$\begin{aligned}\pi &= \sum_{n=0}^{+\infty} \mathbb{P}(L \text{ has } n \text{ children})\mathbb{P}(L \text{ dies}) \prod_{i=1}^{n} \mathbb{P}(L_i \text{ does not complete}) \\ &= \sum_{n=0}^{+\infty} \left(\frac{\lambda_3}{\lambda_2 + \lambda_3 + \mu_2}\right)^n \frac{\mu_2}{\lambda_2 + \lambda_3 + \mu_2} \pi^n.\end{aligned}$$

Therefore

$$\pi = \frac{\mu_2}{\lambda_2 + \lambda_3 + \mu_2} \times \frac{1}{1 - \frac{\pi\lambda_3}{\lambda_2 + \lambda_3 + \mu_2}} \Leftrightarrow \pi^2 - \pi \frac{\lambda_2 + \lambda_3 + \mu_2}{\lambda_3} + \frac{\mu_2}{\lambda_3} = 0.$$

The solution of this equation in $[0,1]$ is

$$\pi = \frac{1}{2}\frac{\lambda_2+\lambda_3+\mu_2}{\lambda_3}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{(\lambda_2+\lambda_3+\mu_2)^2}}\right). \tag{13}$$

Let us now consider a good lineage. Ignoring the proportion π of incipient daughter lineages that will not complete, we can consider the filtered point process with rates of successful initiation $(1-\pi)\lambda_1$ and extinction μ_1. Therefore the probability of speciation is the probability that a successful initiation occurs before the extinction, given by:

$$\mathbb{P}(\text{speciation}) = \frac{(1-\pi)\lambda_1}{\mu_1+(1-\pi)\lambda_1}. \tag{14}$$

The speciation time is the branching time of the first successful incipient lineage to speciate. Therefore the speciation time is distributed as the result of the filtered point process and has the expected value:

$$\mathbb{E}[T|\text{speciation}] = \frac{1}{(1-\pi)\lambda_1+\mu_1}. \tag{15}$$

Expression of the equivalent rates of speciation and extinction

The equivalent rates of BD are the rates such that $\mathbb{P}(\text{speciation})$ and $\mathbb{E}[T|\text{speciation}]$ are equal in both models. To do so, we set $\tilde{\lambda}$ and $\tilde{\mu}$ such that the pairs of expressions in equation 11/14 and 12/15 are equal. This gives us:

$$\begin{aligned}\tilde{\lambda} &= \frac{\mathbb{P}(\text{speciation})}{\mathbb{E}[T|\text{speciation}]}\\ &= (1-\pi)\lambda_1\end{aligned} \tag{16}$$

and

$$\begin{aligned}\tilde{\mu} &= \frac{1-\mathbb{P}(\text{speciation})}{\mathbb{E}[T|\text{speciation}]}\\ &= \mu_1.\end{aligned}$$

A6 Analysis of parameters limiting the equivalent birth rate

We measure the influence of variation in each of the parameters of the PBD model on the equivalent birth rate $\tilde{\lambda}$. Parameters that influence $\tilde{\lambda}$ the most can be considered potentially limiting, as they can drive $\tilde{\lambda}$ down depending on their values. We calculated the partial derivatives of $\tilde{\lambda}$, given by the expression equation 16, with respect to the PBD parameters $\lambda_1, \lambda_2, \lambda_3$ and μ_2. μ_1, denoting the rate of extinction of a good lineage, has no influence on the equivalent birth rate.

$$\frac{\partial\hat{\lambda}}{\partial\lambda_1} = (1-\pi), \quad \frac{\partial\hat{\lambda}}{\partial\lambda_2} = -\lambda_1\frac{\partial\pi}{\partial\lambda_2}, \quad \frac{\partial\hat{\lambda}}{\partial\lambda_3} = -\lambda_1\frac{\partial\pi}{\partial\lambda_3},$$
$$\text{and} \quad \frac{\partial\hat{\lambda}}{\partial\mu_2} = -\lambda_1\frac{\partial\pi}{\partial\mu_2}.$$

In what follow, we explicit the expressions (*i*, *ii* and *iii*) of the three partial derivatives of π with respect to the three parameters of the incipient lineages. From equation 13 we remind $\pi = \Lambda\left(1-\sqrt{1-4\lambda_3\mu_2/\Lambda^2}\right)/2\lambda_3$ with $\Lambda = \lambda_2+\lambda_3+\mu_2$.

$$(i) \quad \frac{\partial\pi}{\partial\lambda_2} = \frac{1}{2\lambda_3}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}\right) + \frac{\Lambda}{2\lambda_3}\left(\frac{-\frac{8\lambda_3\mu_2}{\Lambda^3}}{2\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}\right)$$
$$= \frac{\pi}{\Lambda} - \frac{2\mu_2}{\Lambda\sqrt{\Lambda^2-4\lambda_3\mu_2}},$$
$$(ii) \quad \frac{\partial\pi}{\partial\lambda_3} = \frac{1}{2}\frac{\lambda_3-\Lambda}{\lambda_3^2}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}\right) + \frac{\Lambda}{2\lambda_3}\times\left(-\frac{\partial u}{\partial\lambda_3}\frac{1}{2\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}\right)$$

with u defined as $1-4\frac{\lambda_3\mu_2}{\Lambda^2}$, so $\frac{\partial u}{\partial\lambda_3} = 4\mu_2\frac{\lambda_3-\lambda_2-\mu_2}{\Lambda^3}$, hence

$$\frac{\partial\pi}{\partial\lambda_3} = -\frac{\lambda_2+\mu_2}{2\lambda_3^2}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}\right) + \frac{\Lambda}{4\lambda_3}\frac{4\mu_2\Lambda}{\Lambda^3\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}$$
$$= -\frac{\lambda_2+\mu_2}{\Lambda\lambda_3}\pi + \frac{\mu_2}{\lambda_3\Lambda^2}\frac{\lambda_2+\mu_2-\lambda_3}{\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}, \quad \text{and}$$
$$(iii) \quad \frac{\partial\pi}{\partial\mu_2} = \frac{1}{2\lambda_3}\left(1-\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}\right) - \frac{\Lambda}{2\lambda_3}\frac{\partial u}{\partial\mu_2}\frac{1}{2\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}$$

with $\frac{\partial u}{\partial\mu_2} = -4\lambda_3\frac{\lambda_2+\lambda_3-\mu_2}{\Lambda^3}$, so

$$\frac{\partial\pi}{\partial\mu_2} = \frac{\pi}{\Lambda} + \frac{\lambda_2+\lambda_3-\mu_2}{\Lambda^2\sqrt{1-4\frac{\lambda_3\mu_2}{\Lambda^2}}}.$$

We then use the simplified framework of the PBD model where all rates of initiation are equal ($b := \lambda_1 = \lambda_3$) and all rates of population extinction are equal ($e := \mu_1 = \mu_2$). By the chain rule:

$$\frac{\partial\tilde{\lambda}}{\partial b} = \frac{\partial\tilde{\lambda}}{\partial\lambda_1} + \frac{\partial\tilde{\lambda}}{\partial\lambda_3} \quad \text{and} \quad \frac{\partial\tilde{\lambda}}{\partial e} = \frac{\partial\tilde{\lambda}}{\partial\mu_1} + \frac{\partial\tilde{\lambda}}{\partial\mu_2}.$$

We then evaluate the relative influence of a parameter as the ratio between the absolute partial derivatives. For instance the relative influence of the rate of initiation b is:

$$\text{relative influence}(b) = \frac{\left|\frac{\partial \hat{\lambda}}{\partial b}\right|}{\left|\frac{\partial \hat{\lambda}}{\partial b}\right| + \left|\frac{\partial \hat{\lambda}}{\partial \lambda_2}\right| + \left|\frac{\partial \hat{\lambda}}{\partial e}\right|}.$$

We provided a detailed plot of the values of this relative influence of the 3 simplified parameters in supplementary figure S5. The summary of these relative influences is provided on figure 4. A similar analysis with the detailed model (with $\lambda_1 \neq \lambda_3$ and $\mu_1 \neq \mu 2$ a priori) was also done (supplementary figure S6).