Syst. Biol. 71(6):1362–1377, 2022

https://doi.org/10.1093/sysbio/syac041
Advance Access publication June 14, 2022

Protracted Speciation under the State-Dependent Speciation and Extinction Approach

XIA HUA*, TYARA HERDHA, AND CONRAD J. BURDEN

Mathematical Sciences Institute, Australian National University, Canberra ACT 0200 Australia
**Correspondence to be sent to: Mathematical Sciences Institute, Australian National University, Canberra ACT 0200, Australia; E-mail: xia.hua@anu.edu.au.*

Received 16 July 2021; reviews returned 16 May 2022; accepted 7 June 2022
Associate Editor: Jennifer Mandel

Abstract.—How long does speciation take? The answer to this important question in evolutionary biology lies in the genetic difference not only among species, but also among lineages within each species. With the advance of genome sequencing in non-model organisms and the statistical tools to improve accuracy in inferring evolutionary histories among recently diverged lineages, we now have the lineage-level trees to answer these questions. However, we do not yet have an analytical tool for inferring speciation processes from these trees. What is needed is a model of speciation processes that generates both the trees and species identities of extant lineages. The model should allow calculation of the probability that certain lineages belong to certain species and have an evolutionary history consistent with the tree. Here, we propose such a model and test the model performance on both simulated data and real data. We show that maximum-likelihood estimates of the model are highly accurate and give estimates from real data that generate patterns consistent with observations. We discuss how to extend the model to account for different rates and types of speciation processes across lineages in a species group. By linking evolutionary processes on lineage level to species level, the model provides a new phylogenetic approach to study not just when speciation happened, but how speciation happened. [Micro–macro evolution; Protracted birth–death process; speciation completion rate; SSE approach.]

Speciation takes time (Avise 1999), yet most macroevolutionary models for speciation and extinction assume that speciation takes place instantaneously (Maddison et al. 2007; Ricklefs 2007). A more realistic picture recognizes the continual branching and extinction of new lineages within a species, and each lineage within the species has the potential to evolve into a new independent species. This process of a lineage of one species becoming a new independent species is called protracted speciation, first proposed by Rosindell et al. (2010). Using the terminology in speciation biology, a new lineage can be thought of as a population newly isolated from the other populations of the same species, and the lineage becoming a new species can be thought of the isolated population accumulating reproductive incompatibilities to the other populations. In this study, we propose a new protracted speciation and extinction model applicable to both lineage-level and species-level trees. The model hinges on the innovation of labeling one arbitrarily chosen lineage within a species at a given time to be in "representative" state, indicating that the lineage represents the species in a tree, and tracing the probabilities of the lineage in "representative" state backwards in time through the tree. The model includes calculation of an exact likelihood of the tree and the tip species identities via the state-dependent speciation and extinction approach (SSE; Maddison et al. 2007), and is readily extended to account for variation in speciation processes across lineages (see Discussion section). Importantly, the likelihood is shown to be independent of the choice of representatives, whose only role is to act as a mathematical device for keeping track of the number of distinct species at a given time in the tree. We call the new model protracted SSE or ProSSE.

The modeling of protracted speciation has a complex history, and in justice to earlier contributions that have inspired the current study, we summarize a number of relevant existing papers. Of particular significance is the protracted birth and death model proposed by Etienne and Rosindell (2012), hereafter referred to as PBD. In PBD, a speciation initiation event generates a new lineage, that is, regarded as in an "incipient" state. An incipient lineage then undergoes a speciation completion event that turns the lineage into a "good" new species. PBD is the first diversification model that accounts for different stages of speciation. Accounting for different stages of speciation in a diversification process is critical for linking microevolution on a lineage level to macroevolution on a species level, for two main reasons. First, it provides an important prior on the time window when gene flow or hybridization is possible between two lineages in a phylogenetic network (Huson et al. 2010). For example, PBD was recently applied to delimit species (Sukumaran et al. 2021). Second, it provides an alternative methodology for studying speciation, moving from studying sister species that are usually model organisms, to studying shared patterns in speciation processes across lineages in a species group (Marie Curie SPECIATION Network 2012).

Lambert et al. (2015) derived an approximated likelihood function for PBD on a species-level tree that includes one representative lineage for each species. Fitting PBD to the tree allows us to estimate how often a new lineage arises (speciation initiation rate), how fast it develops reproductive isolation (speciation completion rate), and how often lineages go extinct (extinction rate). Because each edge of a tree has to be a lineage of a species, the likelihood function defines the lineage that leads to a "good" species as the "good" lineage of the species and all the other lineages of the species as "incipient" lineages. The likelihood function is an approximation because it assumes that the representative lineage of

a species in the tree is either the good lineage of the species, or the first descendant of the good but now extinct lineage of the species (Etienne et al. 2014). This approximation does not cause large bias in the estimates (Simonet et al. 2018).

However, fitting the model to phylogenies that only include one lineage of a species does not give accurate estimates of speciation completion rate from small phylogenies (<400 species; Fig. S4 in Simonet et al. 2018), and the estimates of speciation initiation rate and extinction rate are unidentifiable (Simonet et al. 2018). One way to improve the accuracy of PBD is to fit the model to lineage-level trees that include multiple extant lineages of the same species in the tree. This is because lineage-level trees have more lineages near the present that carry most information on the speciation rate independent of extinction rate (parameter λ_0 in Louca and Pennell 2020), and evolutionary histories among incipient lineages of the same species carry important information on the speciation completion rate.

The likelihood function by Lambert et al. (2015) cannot be applied to lineage-level trees for two reasons. First, in the likelihood function, a good lineage and an incipient lineage of a species have different evolutionary dynamics in that a good lineage can have a different speciation initiation rate and extinction rate from an incipient lineage and a good lineage cannot undergo a speciation completion event (otherwise the species has no good lineage on the tree). But in reality, all lineages of a species have the potential to become a distinct species from the other lineages of the species, so all lineages are in the incipient state by definition. In other words, there is no way that we can tell a "good" population from the other "incipient" populations in an extant species. Because a lineage's evolutionary dynamics depends on its state, good or incipient, the likelihood function cannot be applied to lineage-level trees when we cannot tell the state of each extant lineage. Second, the likelihood function assumes trees as coalescent point processes, where node depths are a sequence of independent, identically distributed random variables, and so all trees have equal probability. When we have multiple extant lineages of a species in the tree, the tree is no longer a coalescent point process, because speciation processes depend on the states (proved in Lambert and Stadler 2013). As a simple example, speciation completion events cannot occur along an extant incipient lineage, because it will turn the lineage to a different species at present. This makes a tree with two species paraphyletic to each other impossible, so not all the trees have equal probability. For the same reason, the likelihood function cannot be extended to allow variation in the types or the rates of speciation processes across lineages.

Therefore, we develop ProSSE, a new model to describe the protracted speciation and extinction process. In ProSSE, we still assign states to extant lineages in order to trace the number of distinct species along the tree. But how we assign states to extant lineages does not change the likelihood of the tree and the tip species identity, so we can apply the model to lineage-level trees to improve estimation of speciation initiation rate, extinction rate, and speciation completion rate. Because ProSSE uses SSE approach, it is flexible enough to allow variation in speciation processes across lineages. Below, we begin with a description of how ProSSE differs from PBD, followed by details of the ProSSE algorithms for lineage-level and species-level trees. We then assess the performance of ProSSE on simulated trees and compare the performance of ProSSE and PBD. We further demonstrate ProSSE using Australian rainbow skinks as a case study.

Materials and Methods

In ProSSE, the analogue of the "good" state in PBD is a "representative" state, a label which is attached to one lineage chosen to represent each extant species in the tree. Each species has one representative lineage and multiple incipient lineages if any (Fig. 1a). Unlike PBD's good state, a representative lineage in ProSSE has the same evolutionary dynamics as an incipient lineage, that is, they have the same speciation initiation rate (b), speciation completion rate (λ), and extinction rate (μ). The representative state is included only to indicate how many distinct species there are at a time slice in the tree. As we show below, no matter which lineage we pick as the representative of a species, the likelihood does not change. For example, in Figure 1a, any of the three lineages of species A can be its representative lineage. This allows us to use the SSE approach to calculate the likelihood of the protracted speciation and extinction model, or the joint probability of a tree and its tip species identities under the protracted speciation and extinction model.

The main difference between the "representative" state in ProSSE and the "good" state in PBD is that a speciation completion event is not allowed to happen on a good lineage in PBD, but it is allowed to happen on a representative lineage as often as on an incipient lineage in ProSSE. Biologically speaking, this difference not only gives all lineages of the same species equal chance to become a new independent species, but also makes anagenesis possible in ProSSE. For example, under the Bateson–Dobzhansky–Muller incompatibility model, the speciation completion rate is determined by how fast a population accumulates nearly neutral substitutions on loci that may cause incompatibility (Gavrilets 2014). When there is only one population of a species, the population can still accumulate substitutions on these loci and become a distinct species to its ancestral species, except that its speciation probability is one half of that between two populations, as either of the two populations can become a distinct species relative to the ancestral species.

The likelihood function by Lambert et al. (2015) also uses the terminology "representative," but this differs from that in ProSSE. A representative lineage in PBD is assumed to be either the good lineage of the species, or the first descendant of the good but now extinct lineage

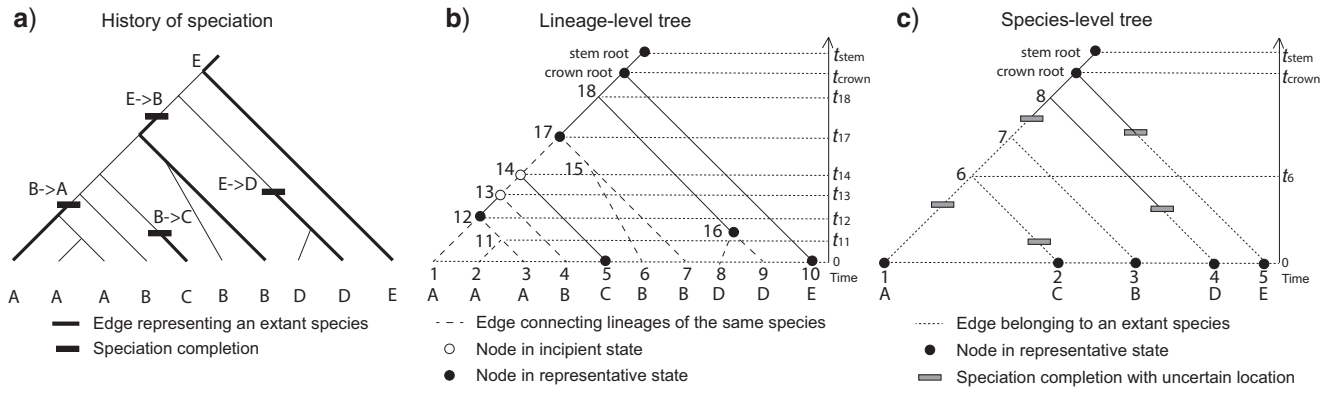

FIGURE 1. Illustration of ProSSE algorithms. a) Shows the true lineage-level tree and true history of speciation completion events on the tree, with each letter representing a distinct species. Each extant lineage is labeled by its species identity. The letter on the LHS and the RHS of the arrow for each speciation completion event indicates the LHS species becomes the RHS species. Assuming no error in tree estimation, we can get some information on the history of speciation (a) from both lineage-level tree (b) and species-level tree (c), even though we do not know exactly where speciation completion events occurred on the tree. From lineage-level tree (b), we know which lineages belong to the same species (dashed edges) and what state certain nodes have (circles). We are also sure that the speciation completion event that leads to an extant species must occur along an edge with ancestral node in incipient state and descendent node in representative state. For species-level tree (c), we only sample one lineage of each extant species from lineage-level tree (b). Lineages 2 and 3 of species A, lineages 4 and 6 of species B, lineage 8 of species D in tree (b) are not in tree (c). Without these lineages, we do not know how many lineages each extant species has. We also do not know that, for example, along the edge connecting Nodes 1 and 6 in tree (c), there are unsampled descendants that lead to two extant lineages of species A and one extant lineage of species B in tree (b). As a result, we cannot be sure if Node 6 in tree (c) is species A, B, C, or some species that does not belong to any extant species. The only thing we know for sure is that, given the location of the speciation completion event of an extant species, edges after the speciation completion event should not leave any lineage of any other extant species in the tree, otherwise, the lineage-level tree will include at least two extant species that are paraphyletic to each, which cannot be generated from protracted speciation and extinction process alone (e.g., it would require gene flow after the completion of speciation). Because the location of the speciation completion event on the tree is uncertain, equations (5–8) are used to account for this uncertainty.

of the species (Etienne et al. 2014), so the representative lineage still has different evolutionary dynamics to incipient lineages. In ProSSE, a representative lineage has the same evolutionary dynamics to incipient lineages, so any lineage of a species can be its representative lineage and the choice of the representative lineage does not change the likelihood.

ProSSE Algorithm for Lineage-Level Trees

In common with other SSE models, ProSSE starts with states of the extant lineages in the tree, then, for each tip edge of the tree, calculates the probability that the edge stays as a single edge along the tree by integrating over all possible events (speciation initiation, speciation completion, and extinction) at any time point along the edge, calculates the joint probability of observing two sister tip edges at the node that connects them, and repeats the process along the internal edges and nodes until reaching the root to get the probability of the tree. What is different from other SSE models is that ProSSE uses different differential equations along edges connecting lineages of the same species (dashed edges in Fig. 1b) and edges connecting lineages of different species (solid edges in Fig. 1b). Along edges connecting lineages of the same species, for example, the dashed edges connecting the three A extant lineages in Figure 1b, any speciation completion event would result in zero probability of observing these three A lineages in the tree. In contrast, suppose we have N distinct species at present, then at least $N-1$ speciation completion events must have occurred along edges connecting lineages of different species (Fig. 1a).

Let us define $D_R(t)$ and $D_I(t)$ as the probabilities that an edge in representative state, state R, and in incipient state, state I, at time t leads to extant descendants as observed in the tree. Here, t is measured backwards in time from the present at $t=0$. Alternatively, we can think of $D_R(t)$ [resp. $D_I(t)$] as the probability that an edge in state R (resp. state I) at time t stays as a single edge until its descendent node on the tree, and the boundary condition at the descendent node is determined by the probability of observing the part of the tree descending from that node. Also define $E(t)$ as the probability that an edge at time t does not leave any descendant at present. Because R and I lineages have the same evolutionary dynamics, $E(t)$ is the same for both. So, for all the edges in the tree, $E(t)$ is calculated under the standard birth–death process (Stadler 2010), with boundary condition $E(0)=0$ and solution $E(t)=\frac{\mu-\mu e^{(b-\mu)t}}{\mu-be^{(b-\mu)t}}$.

Along edges that connect lineages of the same species (dashed edges in Fig. 1b), $D_R(t)$ and $D_I(t)$ can be derived from

$$D_R'(t)=-(\lambda+b+\mu)D_R(t)+2bD_R(t)E(t)+\lambda\cdot 0$$
$$D_I'(t)=-(\lambda+b+\mu)D_I(t)+2bD_I(t)E(t)+\lambda\cdot 0.$$

These equations are similar to the equations in Binary SSE model or BiSSE (Maddison et al. 2007), except that the last term on the RHS equals zero. This is because

if a speciation completion event happens, then not all the lineages belong to the same species. As a result, a speciation completion event only reduces the probability of observing these edges connecting lineages of the same species, in both $D_R(t)$ and $D_I(t)$, as reflected in the first term. Note that these equations satisfy the same differential equation, but the initial condition is either $D_I(0)=0, D_R(0)=1$ for an R lineage, in which case $D_I(t)\equiv 0$; or $D_I(0)=1, D_R(0)=0$ for an I lineage, in which case $D_R(t)\equiv 0$. So to simplify the notation, define $D(t)=D_R(t)+D_I(t)$, then

$$D'(t)=-(\lambda+b+\mu)D(t)+2bD(t)E(t), \quad (1)$$

with the initial condition $D(0)=1$. As a result, which lineage we choose to represent the species does not affect the likelihood. Actually, it is not necessary to assign a state to any edge that connects lineages of the same species, because we only need to calculate $D(t)$, rather than calculate $D_R(t)$ and $D_I(t)$ separately. The solution at the ancestral node of an edge at time $t+s$ is:

$$D(t+s)=D(t)e^{-\lambda s}A(t,s),$$

where $D(t)$ is at the descendant node of the edge at time t, s is the finite edge length, and

$$A(t,s)=\frac{(b-\mu)^2 e^{(b-\mu)s}}{\left[be^{(b-\mu)s}-\mu+bE(t)(1-e^{(b-\mu)s})\right]^2}$$

which is also the solution to the SSE equations for constant-rate birth–death model (FitzJohn 2009).

Along edges connecting lineages of different species (solid edges in Fig. 1b), R and I lineages have different dynamics, because a speciation completion event resulting in an R state may occur in both cases. $D_R(t)$ and $D_I(t)$ can be derived from

$$D'_R(t)=-(b+\mu)D_R(t)+2bD_R(t)E(t) \quad (2)$$

$$D'_I(t)=-(\lambda+b+\mu)D_I(t)+2bD_I(t)E(t)+\lambda D_R(t). \quad (3)$$

A speciation completion event has no effect on $D_R(t)$ in the equation because although an R lineage becomes a distinct species, it is still in R state, albeit representing a different species, that is, $\lambda D_R(t)$ term is cancelled out in equation (2) from $D'_R(t)=-(\lambda+b+\mu)D_R(t)+2bD_R(t)E(t)+\lambda D_R(t)$. The solutions at the ancestral node of an edge at time $t+s$ are

$$D_R(t+s)=D_R(t)A(t,s)$$

$$D_I(t+s)=D_R(t+s)+[D_I(t)-D_R(t)]A(t,s)e^{-\lambda s},$$

where $D_R(t)$ and $D_I(t)$ are at the descendant node of the edge at time t and s is the finite edge length. For an extant species with only one lineage, such as species C in Figure 1b, its tip edge is also considered as connecting lineages of different species, so we apply equations 2 and 3 to the edge with initial conditions $D_R(0)=1$ and $D_I(0)=0$.

Consider the tree in Figure 1b, where there are 10 extant lineages belonging to 5 distinct extant species. ProSSE first calculates from tip lineages that belong to the same species toward their common ancestors on the tree. For example, tip edges 1, 2, 3 belong to species A (Fig. 1b), so we calculate $D(t)$ along these tip edges using equation 1. Then, for each node along the edges connecting these tips, such as node 11, we calculate

$$D_{11}(t_{11})=bD_2(t_{11})D_3(t_{11}).$$

Here, we use the subscript to indicate which edge we are working on, for example, $D_{11}(t_{11})$ is the boundary condition at the descendant node of the edge linking Node 11 and Node 12 and the time at Node 11 is t_{11}. Repeating these calculations till Node 12, we have $D_{12}(t_{12})$.

Next, ProSSE works on the edges connecting lineages of different species, which include the edges connecting the common ancestors of different extant species (e.g., the edge connecting Nodes 17 and 18), the edges linking the common ancestor of some extant species to a lineage of a different extant species that is paraphyletic (e.g., the edge connecting Node 12 and Node 13; species B is paraphyletic), as well as the tip edge of a species that has only one extant lineage (e.g., tip edge 5). Along these edges, we calculate $D_R(t)$ and $D_I(t)$ using equations 2 and 3.

For edges that connect the common ancestors of different extant species, the boundary conditions at each common ancestor, for example, Node 17, are $D_{R,17}(t_{17})=D_{17}(t_{17})$, $D_{I,17}(t_{17})=0$, because the edge linking Nodes 17 and 18 are the only lineage representing species B at t_{17}, so Node 17 must be in R state (Fig. 1b). Then, we can calculate along these edges toward the root of the tree. For example, at Node 18, we have

$$D_{R,18}(t_{18})=bD_{R,17}(t_{18})D_{I,16}(t_{18})+bD_{I,17}(t_{18})D_{R,16}(t_{18})$$

$$D_{I,18}(t_{18})=bD_{I,17}(t_{18})D_{I,16}(t_{18})$$

Note that these equations are different from BiSSE, because the new lineage generated from each speciation initiation event is always in the I state.

Edges connecting the common ancestor of some extant species to a lineage of a different extant species that is paraphyletic, for example, Node 13 in Figure 1b, are calculated as follows. One edge splitting from Node 13 belongs to species B and the other edge leads to species A. Because the edge that leads to species A cannot represent species B, the state of the edge linking Nodes 12 and 13 at t_{13} must be in I state, so

$$D_{R,13}(t_{13})=bD_{R,4}(t_{13})D_{I,12}(t_{13})$$

$$D_{I,13}(t_{13})=bD_{I,4}(t_{13})D_{I,12}(t_{13}).$$

Because tip edge 4 belongs to species B, either $D_{R,4}(t_{13})=D_4(t_{13})$ and $D_{I,4}(t_{13})=0$, or $D_{R,4}(t_{13})=0$ and $D_{I,4}(t_{13})=D_4(t_{13})$, depending on the choice of R lineage for species B. Then, we calculate along the edge linking Nodes 13 and 14, which is a dash edge (Fig. 1b), so the calculation

uses equation 1 and the boundary condition at t_{13} is $D_{13}(t_{13}) = D_{R,13}(t_{13}) + D_{I,13}(t_{13})$.

Tip edges of a species that has only one extant lineage, for example, tip edge 5 in Figure 1b, are calculated as follows. The calculation of Node 14 is similar to Node 13, so

$$D_{R,14}(t_{14}) = bD_{R,13}(t_{14})D_{I,5}(t_{14})$$
$$D_{I,14}(t_{14}) = bD_{I,13}(t_{14})D_{I,5}(t_{14}),$$

where $D_{I,5}(t_{14})$ is calculated using equations 2 and 3 with $D_{R,5}(0) = 1$, $D_{I,5}(0) = 0$. In this particular example, the edge linking Nodes 14 and 17 is also a dashed edge belonging to species B, so $D_{14}(t_{14}) = D_{R,14}(t_{14}) + D_{I,14}(t_{14}) = bD_{13}(t_{14})D_{I,5}(t_{14})$.

Finally, we are at the crown root. Because there must be only one lineage representing the root species, the crown root is in R state and the likelihood for the tree is $D_{R,crown}(t_{crown})$. If the tree has a root edge as in Figure 1b, then the likelihood is further calculated along the root edge using equation 2 with boundary condition $D_{R,crown}(t_{crown})$. To condition the final likelihood on the survival of the tree that has at least two distinct extant species, the likelihood is divided by $b[1-E(t_{stem})][1-E_R(t_{crown})]$, where $t_{stem} = t_{crown}$ if the tree does not include a root edge and $E_R(t)$ is the probability that a lineage of a certain species at time t (root species at t_{crown} in this case) left no extant species that is different from its species identity. $E_R(t)$ is derived, as in Lambert et al. (2015), from

$$E_R'(t) = -(\lambda + b + \mu)E_R(t) + 2bE_R(t)^2 + \mu + \lambda E(t) \quad (4)$$

with boundary condition $E_R(0) = 1$. The last term on the RHS models a speciation completion event that leads to a distinct species but goes extinct before the present.

ProSSE Algorithm for Species-Level Trees

The above ProSSE algorithm can be adapted to species-level trees that PBD uses, that is, trees including one representative lineage of each extant species (see Fig. 1c). By representative sampling, we do not know how many unsampled lineages there are in each extant species or how these lineages are placed in the tree. As a result, we cannot be sure of the species identity of any edge on the tree (Fig. 1c). Species identity is particularly important in this case. Let us demonstrate this using the tree in Figure 1c and assume that we know where the speciation completion event that leads to each extant species occurred on the tree, so that we know which part of each edge belong to the same species as an extant species. Along a part of an edge that does not belong to any extant species (the solid part of the edges in Fig. 1c), we are certain that any lineage splitting from the part of the edge (except for the descendent node) went extinct before the present, so we can use $E(t)$ to describe this probability as in equations 2 and 3. However, along a part of an edge that belongs to an extant species (the dotted part of the edges in Fig. 1c), we are only certain that any lineage splitting from the part of the edge (except for the descendent node) did not leave an extant species, that is, different from the species identity of the edge, so we need to use $E_R(t)$, instead of $E(t)$, to describe this probability. For example, the edge linking Node 6 and 7 in Figure 1c can only leave extant lineages of species B.

Because we do not know where speciation completion events occurred on the tree, we need to use different equations for edges belonging to an extant species and edges not belonging to any extant species, and use speciation completion rate as the transition rate between them. Whether an edge belongs to an extant species is denoted by the subscript after R or I state, with 1 for an edge belonging to an extant species and 0 for an edge not belonging to any extant species. For example, $D_{R1}(t)$ is the joint probability of the part of tree descended from an edge in state R at time t and the edge's species identity being an extant species. These equations are

$$D_{R1}'(t) = -(\lambda + b + \mu)D_{R1}(t) + bD_{R1}(t)E_R(t) \quad (5)$$
$$D_{R0}'(t) = -(b + \mu)D_{R0}(t) + 2bD_{R0}(t)E(t) + \lambda D_{R1}(t) \quad (6)$$
$$D_{I1}'(t) = -(\lambda + b + \mu)D_{I1}(t) + 2bD_{I1}(t)E_R(t) + \lambda D_{R1}(t) + \lambda D_{R0}(t) \quad (7)$$
$$D_{I0}'(t) = -(\lambda + b + \mu)D_{I0}(t) + 2bD_{I0}(t)E(t) + \lambda D_{R0}(t) + \lambda D_{R1}(t), \quad (8)$$

with the boundary condition for each tip edge at present as $D_{R1}(0) = 1$, $D_{R0}(0) = 0$, $D_{I1}(0) = 0$, $D_{I0}(0) = 0$. Note that $D_R(t) = D_{R1}(t) + D_{R0}(t)$ and $D_I(t) = D_{I1}(t) + D_{I0}(t)$, such that these equations equate to equations 2 and 3 when all lineages are sampled.

Equations 6 and 8 are for edges not belonging to any extant species. They are similar to equations 2 and 3, except that we need $\lambda D_{R1}(t)$ in both equations to describe the event where an edge in R or I state becomes an R lineage of an extant species. Equations 5 and 7 are for edges belonging to an extant species. Equation 5 is similar to equation 1, because a speciation completion event along an R lineage of an extant species will change its species identity and so will only reduce the probability of observing the lineage. Equation 5 differs from equation 1 in that 1) $E(t)$ is replaced by $E_R(t)$ to account for any unsampled lineages of the same species; 2) the coefficient of the last term on the RHS changes from 2 to 1. The coefficient is 2 in equation 1 because the tree does not change when either edge splitting from a speciation initiation event is the observed edge, as the other edge must be extinct. This is no longer true in equation 5, because both edges may leave some extant lineages of the species and so the observed edge has to be the one that led to the sampled lineage of the species, otherwise the tree will be different. Similarly, we need to replace $E(t)$ by $E_R(t)$ in equation 7 and add the term $\lambda D_{R0}(t)$ to describe the event where an I lineage of an

extant species becomes a new species that does not leave any extant lineage of its species identity with probability $D_{R0}(t)$.

Equations (5–8) have no easily determined analytical solution, so we numerically integrate the equations along each edge. At each internal node, such as Node 6 in Figure 1c, we have

$$D_{R1,6}(t_6) = bD_{R1,1}(t_6)D_{I1,2}(t_6) + bD_{I1,1}(t_5)D_{R1,2}(t_6)$$
$$D_{R0,6}(t_6) = bD_{R0,1}(t_6)D_{I0,2}(t_6) + bD_{I0,1}(t_5)D_{R0,2}(t_6)$$
$$D_{I1,6}(t_6) = bD_{I1,1}(t_6)D_{I1,2}(t_6)$$
$$D_{I0,6}(t_6) = bD_{I0,1}(t_6)D_{I0,2}(t_6).$$

Repeating the calculation along each edge from tips to the crown root, or to the stem root if the tree has a root edge. The root must be in an R state, so the final likelihood is $D_{R1}(t_{stem}) + D_{R0}(t_{stem})$, where $t_{stem} = t_{crown}$ if the tree has no root edge. To condition the likelihood on the survival of the tree, $D_{R1}(t_{stem})$ is divided by $b[1-E(t_{stem})][1-E_R(t_{crown})]$, but $D_{R0}(t_{stem})$ is divided by $b[1-E_R(t_{crown})][1-E_R(t_{crown})]$. This is because if the root does not belong to any existing species, then both edges split from the crown root should leave some extant species that are distinct to the root species.

Assessing ProSSE Performance

We assess the performance of ProSSE by first simulating lineage-level trees under our protracted speciation and extinction model, then searching for the maximum-likelihood (ML) estimates of each parameter in the model using the ProSSE algorithm for lineage-level trees, and last reporting the deviation in the ML estimates to the true parameter values used in the simulation (hereafter referred to as the "error"). Because ProSSE assumes all lineages have the same evolutionary dynamics, it should give very similar ML estimates of speciation initiation rate (b) and extinction rate (μ) to the constant-rate birth–death model. To test this, we use the likelihood function in Nee et al. (1994) for constant-rate birth–death model to get the ML estimates of b and μ, and compare the error in these ML estimates with that of ProSSE. For each simulated tree, we also construct its species-level tree by randomly picking a representative lineage of each extant species in the tree and discarding the other lineages. We then search for the ML estimates of each parameter in the model using the ProSSE algorithm for species-level trees, report the error in the ML estimates, and compare the error to that reported in Simonet et al. (2018) for PBD. In addition to the three parameters b, μ, and λ, we also report the error in the estimated value for the diversification rate of lineages, which is calculated as $b-\mu$, as well as the error in the estimated value for the expected duration of speciation, which is the average amount of time that any lineage of a species takes to become a distinct species. The formula for the expected duration of speciation is

$$\tau_{mean} = \frac{2}{D-\lambda+b-\mu}\log\left(\frac{2}{1+\frac{\lambda-b+\mu}{D}}\right)$$

where $D = \sqrt{(\lambda+b-\mu)^2+4\lambda\mu}$, as given by Etienne and Rosindell (2012). This formula still applies to ProSSE because it was derived from the master equation of PBD for the number of species and lineages over time (Etienne and Rosindell 2012), which does not distinguish between good and incipient lineages of a species (see Discussion).

We simulate trees using the same parameter sets as used in Simonet et al. (2018). In brief, we simulate 1000 trees under our protracted speciation and extinction model for various combinations of parameter values ($b = 0.3, 0.4, 0.5, 0.6, 0.7$; $\mu = 0, 0.1, 0.2$; $\lambda = 0.1, 0.3, 1$). But our simulation is different from the PBD simulation in Simonet et al. (2018) in three ways. First, we do not track the good or incipient state of each lineage over time but track each lineage's species identity. Second, all the lineages in our simulation can go through speciation initiation, speciation completion, and extinction events. Third, Simonet et al. (2018) fixed the crown age of each simulated tree to 15, but this will exclude many possible trees we may observe in nature under high extinction rate. Instead, we fix the stem age of each simulated tree to 15. This difference will result in smaller and less informative trees in our simulation with high extinction rate, compared with the PBD simulation.

We implement the ProSSE algorithms and the simulation in the R package for SSE models: "diversitree" (FitzJohn 2012). The implementation includes new R and C functions, new Rd files, and updated namespace file to the package. These are available at github.com/huaxia1985/ProSSE. Users can copy these files to the package source code, compile, and install the modified package in R. The main functions are make.prosse for the algorithm for lineage-level tree, make.prosse.sp for the algorithm for species-level tree, and make.tree.prosse for simulating trees under our protracted speciation model. These functions can be used in the same ways as the other SSE models in the package, for example, estimating parameters using both ML and Bayesian approaches.

Case Study: Australian Rainbow Skinks

To demonstrate the applicability of ProSSE on real species groups, we use Australian rainbow skinks as the case study. Rainbow skinks include three recognized genera: *Carlia*, *Lygisaurus*, and *Liburnascincus*, which together contain 41 named species in Australia. Many of these species have clear phylogeographic lineages. Bragg et al. (2018) published a preliminary multispecies coalescent tree of the species group and Bragg et al. (unpublished data) updated the tree with a complete sample of all recognized species and all phylogeographic lineages within each species using StarBEAST2

(Ogilvie et al. 2017). Singhal and Moritz (2013) found absence of introgression in the hybrid zone between *Carlia rubrigularis* N and S lineages in the group. This provides key evidence of reproductive isolation between the two lineages and *C. rubrigularis* N lineage is now elevated to a distinct species *Carlia crypta* (Singhal et al. 2018). Singhal et al. (2018) suggested that the divergence time between *C. crypta* and *C. rubrigularis* (t_c) is a conservative estimate of the amount of time to complete speciation in the group, because two of the four pairs of sister species in two typical species groups (*C. rubrigularis* group and *Lampropholis coggeri* group) of the subfamily that contain the rainbow skinks have shorter divergence time than t_c. The completeness of the lineage-level tree and independent evidence on the time to complete reproductive isolation make Australian rainbow skinks a good case study to test the performance of ProSSE. If ProSSE can reliably estimate speciation initiation rate b, extinction rate μ, and speciation completion rate λ in real species groups, then the estimated parameter values for Australian rainbow skinks should have high probability to result in two out of four pairs of sister species having shorter divergence time than t_c, assuming that lineages in the species group have similar speciation processes (see Discussion section).

Using ProSSE for lineage-level trees, we get the ML estimates of the three parameters from each of the 1800 posterior samples of the multispecies coalescent tree of the species group (Bragg et al. unpublished data), in order to account for phylogenetic uncertainty. For each sample of the tree with root depth t_{root} and the corresponding ML estimates of the three parameters, we derive the probability density distribution of divergence time between sister species in the species group and use this distribution to calculate the proportion of divergence time between sister species shorter than t_c. To derive the probability density distribution, we apply ProSSE for species-level trees to calculate the probability of a tree (unconditional on the survival of the tree) that consists of one pair of sister species linked by a crown root with tip edge length t_d. Because the tree is just a pair of sister species with divergence time t_d, the tree probability unconditional on tree survival gives the probability of observing a pair of sister species with divergence time t_d. Integrating the tree probability over t_d from 0 to t_{root} gives the overall probability of observing a pair of sister species. Then, the tree probability divided by the integral gives the probability density of t_d.

Results

High Accuracy of ProSSE Estimates for Lineage-Level Trees

For speciation initiation rate b and extinction rate μ, the absolute medians of error in ProSSE ML estimates are close to zero (b: 0.01 ± 0.005; μ: 0.03 ± 0.028) over all simulation scenarios (Figs. 2a–e and 3a–e). We report the absolute medians of error, instead of the common mean squared error, because all estimates are constrained to be non-negative, so they cannot be mean-unbiased. The interquartile ranges of error decrease rapidly when the number of lineages increases (Figs. 2f and 3f), from b: −0.16 to 0.46 and μ: −0.16 to 0.84 for trees with 5 to 50 tip lineages, to b: −0.07 to 0.11 and μ: −0.10 to 0.19 for trees with over 50 tip lineages. These results suggest that ProSSE ML estimates of speciation initiation rate and extinction rate for completely sampled tree are median-unbiased consistent estimators, with high accuracy for trees on lineage level that typically consist of over 50 tip lineages. The errors for both parameters, as well as the lineage diversification rate ($b-\mu$), are very similar between ProSSE and constant-rate birth–death model (Figs. 2 and 3; Fig. S1 of the Supplementary material available on Dryad at https://doi.org/10.5061/dryad.59zw3r27p), confirming that ProSSE is the exact likelihood function for protracted speciation and extinction process where all lineages have the same evolutionary dynamics.

For speciation completion rate λ, the absolute medians of error in ProSSE ML estimates are close to zero (0.01 ± 0.017) when $\lambda \leq b-\mu$, where $b-\mu$ is lineage diversification rate (Fig. 4a–e). When $\lambda > b-\mu$, ProSSE overestimates λ, particularly for trees with all extant species having a single lineage, that is, no extant lineage is in incipient state (compare boxplots in white and gray in Fig. 4a–e). These are the trees with little information on the upper bound of λ, because speciation completion only needs to happen fast enough to make all extant lineages become distinct species. In contrast, incipient lineages give information on the lower bound of λ, because speciation completion needs to happen sufficiently slowly, so that it does not happen along any incipient lineage. As a result, the proportion of lineages representing a single-lineage species in a tree determines whether ProSSE overestimates speciation completion rate.

The threshold for this proportion above which ProSSE overestimates λ is about 0.5 (Fig. 4f): when tree size increases, errors in the ProSSE estimate for trees with over half lineages representing single-lineage species converges to 0.33, whereas errors for trees with under half lineages representing single-lineage species converge to zero. These results suggest that ProSSE ML estimate is an unbiased consistent estimator for speciation completion rate when trees have under half their lineages representing single-lineage species. The interquartile range of error decreases rapidly from −0.11 to 0.26 for trees with 5 to 50 tip lineages, to −0.02 to 0.08 for trees with over 50 tip lineages. When trees have over half their lineages representing a single-lineage species, the ProSSE ML estimate is an inconsistent estimator with bias 0.33 and the interquartile range of error reduced with tree size, from −0.11 to 49 for trees with 5 to 50 tip lineages to 0.26 to 0.52 for trees over 50 tip lineages. Overall, ProSSE gives reasonably good estimate for speciation completion rate for lineage-level trees that typically consist of over 50 tip lineages.

The above bias in the estimates of speciation completion rate (λ) has little influence on the estimates of

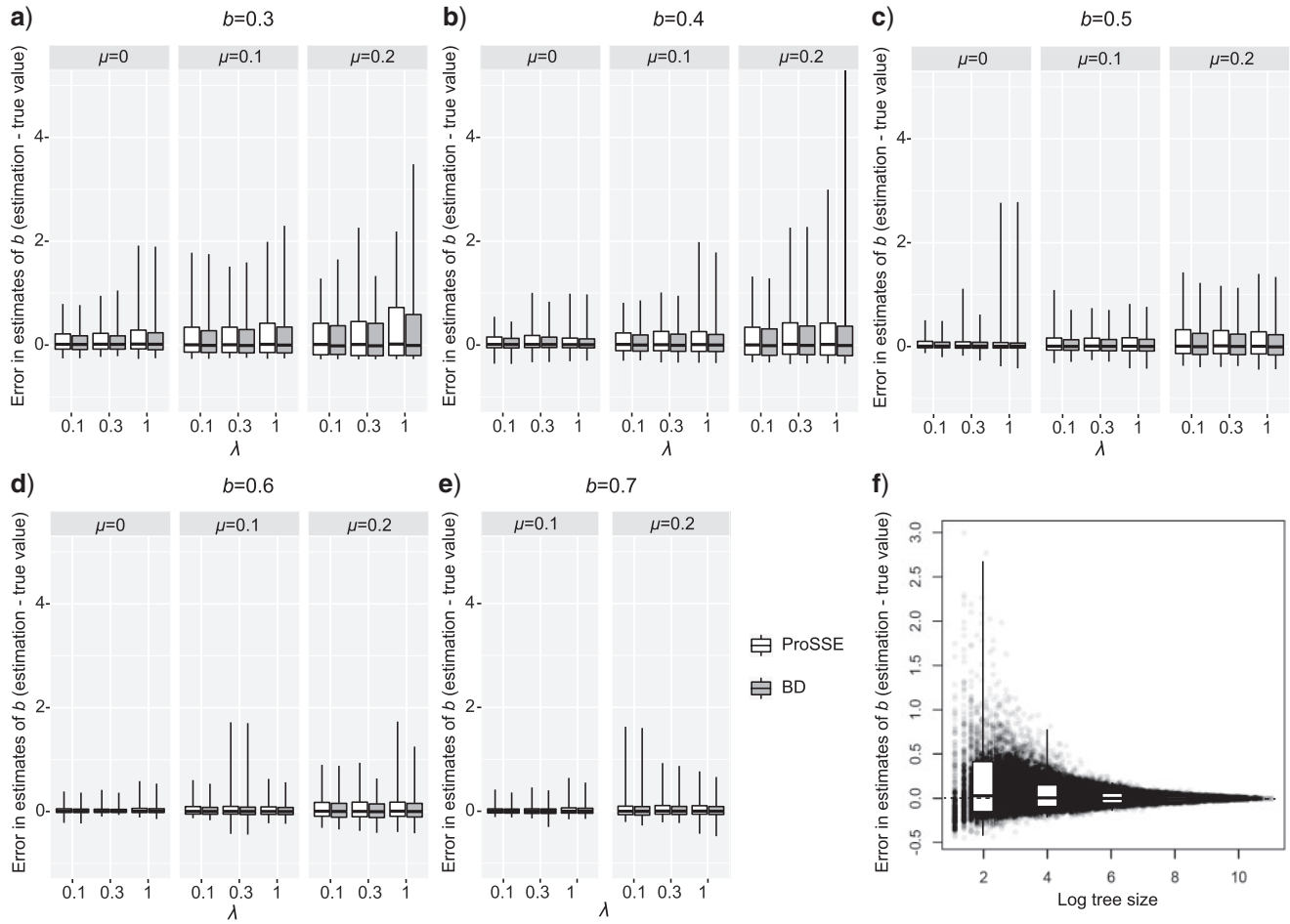

FIGURE 2. Error in ML estimates of the speciation initiation rate b for each simulation scenario using lineage-level trees. Each plot of a–e) represents a value of b used in the simulation. Each facet represents a value of μ used in the simulation. Each tick on x-axis represents a value of λ used in the simulation. Each boxplot represents the distribution of errors, the difference between the estimated and the true value of b over 1000 simulated trees, showing the minimum, the maximum, the median, and the first and third quartiles of the distribution. White boxplots are errors of ProSSE estimate on lineage-level trees. Gray boxplots are errors of constant-rate birth–death model estimate on lineage-level trees. f) Plots errors of ProSSE estimate of b on lineage-level tree over all simulation scenarios against log tree size (the number of lineages in the simulated tree on log scale), with each datapoint corresponding to a simulated tree. White boxplots in f) show the distribution of errors for log tree size between 2 and 4, between 4 and 6, between 6 and 8, 8 and 10, and over 10.

the expected duration of speciation (τ_{mean}) (Fig. S2f of the Supplementary material available on Dryad). The absolute medians of error in ProSSE estimates of τ_{mean} are close to zero (0.00 ± 0.001) over all simulation scenarios (white boxplots in Fig. S2a–e of the Supplementary material available on Dryad). The interquartile ranges of error decrease rapidly when the number of lineages increases (Fig. S2f of the Supplementary material available on Dryad), from −2.14 to 1.61 for trees with 5 to 50 tip lineages, to −0.28 to 0.19 for trees with over 50 tip lineages. The large proportion of lineages representing single-lineage species in the trees, which causes bias in the estimates of λ, only influences the range of error, rather than the median of error in the estimates of τ_{mean} (Fig. S2f of the Supplementary material available on Dryad). These results suggest that ProSSE ML estimates of the expected duration of speciation for completely sampled tree are median-unbiased consistent estimators with high accuracy for trees on lineage level of typical tree size.

ProSSE Outperforms PBD for Species-Level Trees

For species-level trees, that is, trees with only one representative lineage sampled for each species, both ProSSE and PBD give biased estimators for the three parameters (Figs. 5–7). Over all simulation scenarios, ProSSE ML estimates have similar biases (the absolute medians of error for b: 0.12 ± 0.090; μ: 0.11 ± 0.081; λ: 0.21 ± 0.168) to PBD ML estimates (b: 0.16 ± 0.222; μ: 0.17 ± 0.226; λ: 0.09 ± 0.107). However, ProSSE ML estimates have narrower interquartile range of errors (b: 0.17 ± 0.116; μ: 0.08 ± 0.138; λ:0.27 ± 0.171) than PBD (b: 16 ± 31.5; μ: 16 ± 31.5; λ:1.58 ± 2.190). In particular, speciation initiation rate b and extinction rate μ are

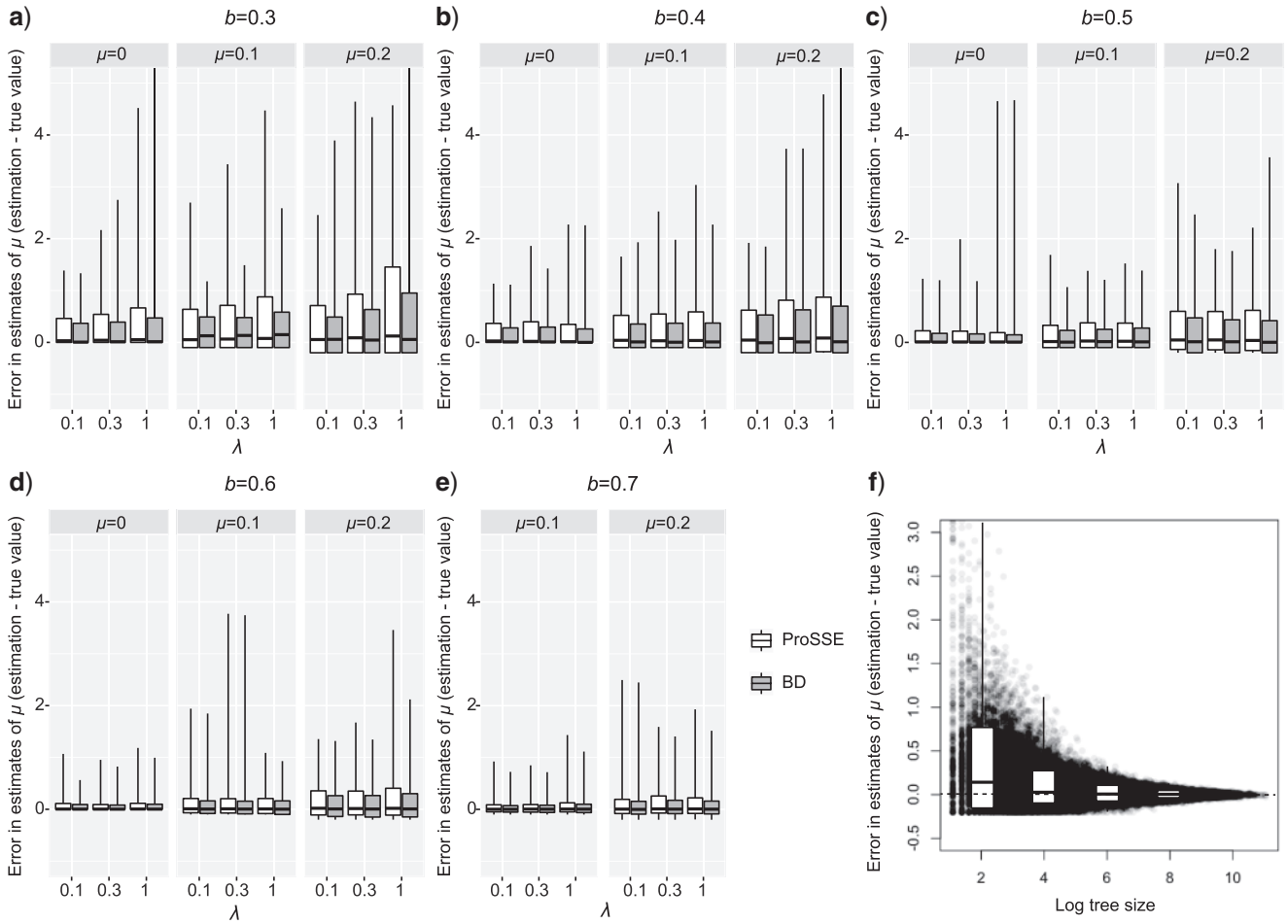

FIGURE 3. Error in ML estimates of the extinction rate μ for each simulation scenario using lineage-level trees (a–e) and the decrease of error bounds over tree size for ProSSE estimate of μ on lineage-level trees (f). See figure details in the legend of Figure 2.

generally unidentifiable in PBD (Simonet et al. 2018). This unidentifiability issue can be seen from the PBD likelihood function (see equations [2–3] in Etienne et al. 2014), where likelihood mainly depends on b and $b-\mu$, so likelihood tends to increase with b. In ProSSE, likelihood is calculated from D_{R1} and D_{R0}, both of which mainly depend on b and μ. As a result, b and μ are identifiable in ProSSE (Figs. 5 and 6).

However, ProSSE tends to underestimate b when the speciation completion rate λ is lower. This is because more incipient lineages are unsampled in trees generated under low λ, which results in loss of information on recent speciation initiation events and so lower estimation of b. The underestimation of μ is consistent across simulation scenarios. This is a common problem due to incomplete sampling in the SSE approach (FitzJohn et al. 2009), because incomplete sampling reduces the effect of μ on the extinction probability $E(t)$, for example, in ProSSE, $E(t)$ becomes $E_R(t)$, so the extinction rate affects tree likelihood mainly via the $-\mu$ term in equations 5 and 7, which makes the ML estimate of μ lower than the true extinction rate.

As a result of underestimation in the speciation initiation rate b under a lower speciation completion rate λ and consistent underestimation of the extinction rate μ, ProSSE underestimates the lineage diversification rate ($b-\mu$) under low λ (λ = 0.1 in Fig. S1 of the Supplementary material available on Dryad) and overestimates it under high λ (λ = 1 in Fig. S1 of the Supplementary material available on Dryad). This leads to overestimation of λ when λ is small (under λ = 0.1 in Fig. S1 of the Supplementary material available on Dryad) and underestimation when λ is high (under λ = 1 in Fig. S1 of the Supplementary material available on Dryad), because an underestimated lineage diversification rate increases the probability that a tree has no unsampled incipient lineages, and so favors a high speciation completion rate, and vice versa. This also leads to underestimation of the expected duration of speciation (τ_{mean}) when λ is small (under λ = 0.1 in Fig. S2 of the Supplementary material available on Dryad) and overestimation when λ is high (under λ = 1 in Fig. S2 of the Supplementary material available on Dryad), because τ_{mean} largely depends on $b-\mu$. Compared with ProSSE, although the PBD likelihood function cannot

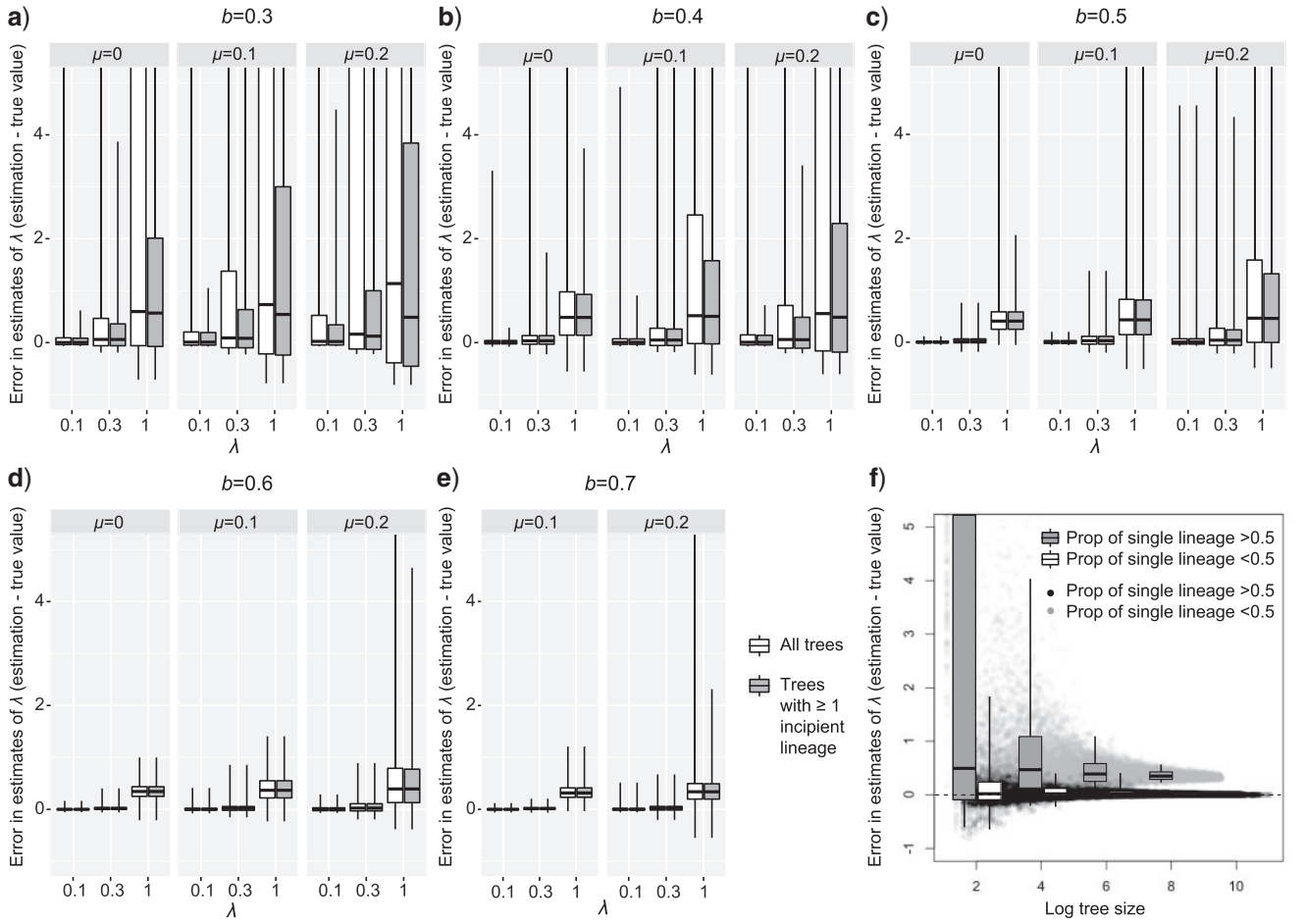

FIGURE 4. Error in ML estimates of the speciation completion rate λ for each simulation scenario using lineage-level trees (a–e) and the decrease of error bounds over tree size for ProSSE estimate of λ on lineage-level trees (f). ProSSE overestimates λ, particularly for trees with no extant lineage in incipient state, so a–e) gives boxplots for all the simulated trees (white) and for trees with at least one extant incipient lineage (gray). In cases of extreme values, boxplots are cut off for graphical readability. Facet f) show the distribution of errors in λ for log tree size between 2 and 4, between 4 and 6, between 6 and 8, 8 and 10, and over 10. ProSSE estimate of λ is biased for trees with more than half lineages representing single-lineage species, so f) plots trees with under and above half lineages representing single-lineage species separately, with gray dots and gray boxplots for trees with more than half lineages being single-lineage species and black dots and white boxplots for trees with less than half lineages being single-lineage species.

separately identify b and μ, it gives unbiased estimates of lineage diversification rate (Fig. S1 of the Supplementary material available on Dryad). Similar to ProSSE, PBD also underestimates τ_{mean} when λ is small (under $\lambda = 0.1$ and $\lambda = 0.3$ in Fig. S2 of the Supplementary material available on Dryad). Both methods give similar range of errors in the estimates of lineage diversification rate and τ_{mean} (Figs. S1 and S2 of the Supplementary material available on Dryad). In general, ProSSE and PBD have similar performance for species-level trees, except that speciation initiation rate and extinction rate are unidentifiable in PBD.

ProSSE Gives Reliable Estimates from Real Data

For Australian rainbow skinks, ProSSE estimates, on average, 0.27 speciation initiation rate (Fig. 8b), nearly zero extinction rate (Fig. 8c), and 0.31 speciation completion rate (Fig. 8d). These estimates mean that, along a given lineage of Australian rainbow skinks, the interval between two successive speciation initiation events along the lineage follows an exponential distribution with mean ~3.7 Myr. After a speciation initiation event splits the lineage into two, the time for either of the two lineages to become a new species follows an exponential distribution with mean ~3.2 Myr. These estimates lead to nearly half extant sister species in the species group having shorter divergence time than that between *C. crypta* and *C. rubrigularis*, which gives about 0.35 probability that two out of four pairs of extant sister species have shorter divergence time than that between *C. crypta* and *C. rubrigularis* (Fig. 8e). This result agrees with Singhal et al. (2018) that the divergence time between *C. crypta* and *C. rubrigularis* is a conservative cutoff to delimit distinct species in Australian rainbow skinks, suggesting that ProSSE is able to give reasonable estimates of parameters from real data. The near zero extinction rate suggests that

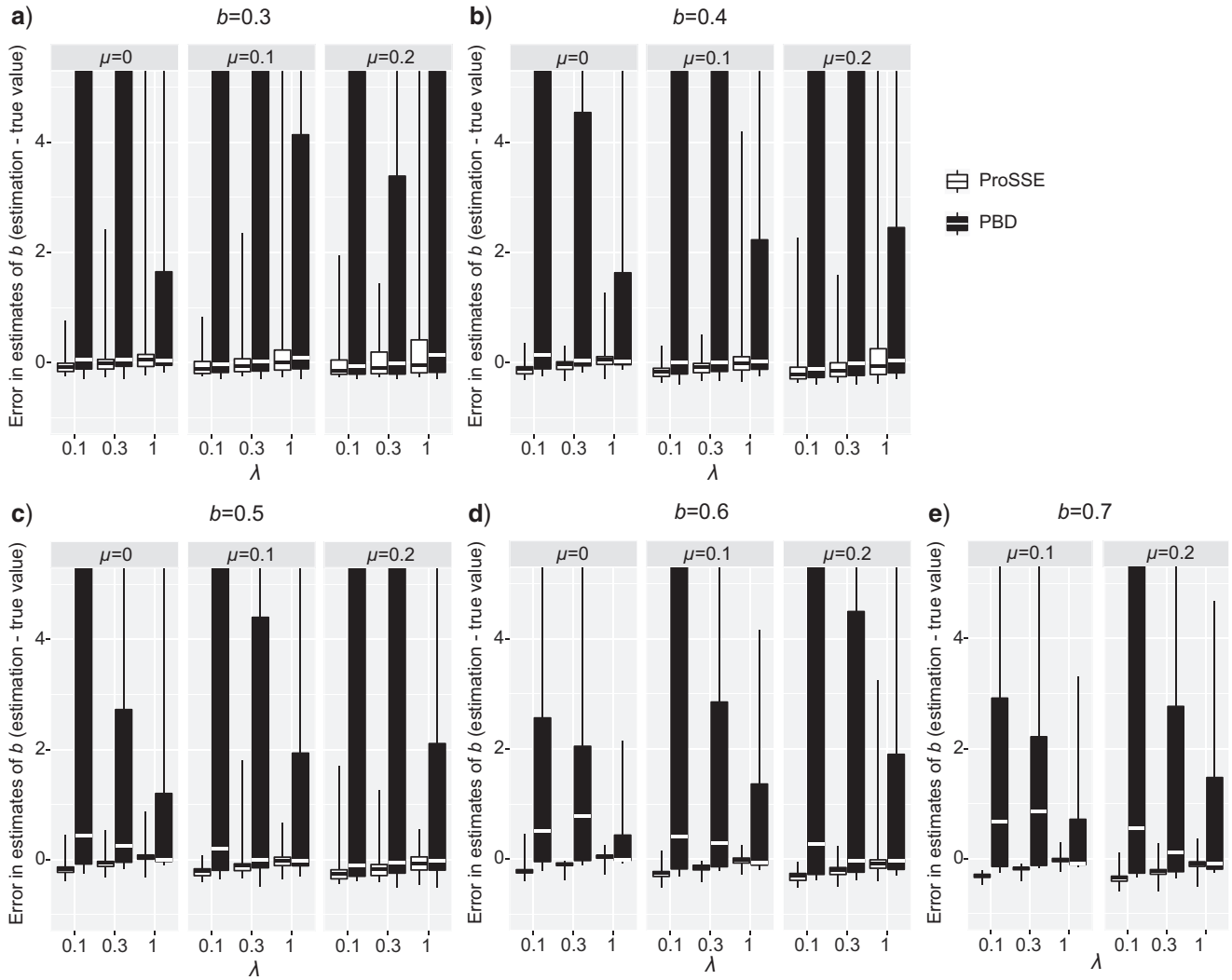

FIGURE 5. Error in ML estimates of the speciation initiation rate b for each simulation scenario using species-level trees. Each plot of a–e) represents a value of b used in the simulation. Each facet represents a value of μ used in the simulation. Each tick on x-axis represents a value of λ used in the simulation. Each boxplot represents the distribution of errors, the difference between the estimated and the true value of b over 1000 simulated species-level trees, showing the minimum, the maximum, the median, and the first and third quartiles of the distribution. White boxplots are errors of ProSSE. Black boxplots are errors of PBD estimate using the likelihood function by Lambert et al. (2015). The data for PBD are from Simonet et al. (2018). In cases of extreme values, boxplots are cutoff for graphical readability.

the diversification history of Australian rainbow skinks may not be sufficiently modeled by the constant-rate birth–death model, that is, constant speciation initiation rate and extinction rate over time and lineages, because the maximum-likelihood estimate of the extinction rate may be negative if it is not constrained to be non-negative (Louca and Pennell 2021). However, this should not affect the estimation of speciation completion rate, because speciation completion events are independent of birth–death process in ProSSE.

DISCUSSION

This study provides an exact likelihood function for the protracted speciation and extinction model, using state-dependent speciation extinction approach (ProSSE). We show that, for completely sampled trees on lineage level, ProSSE is able to give accurate estimates for trees of typical size in macroevolutionary analyses. When applied to real data, ProSSE is able to give reliable estimates that are consistent with independent evidence on species divergence time. For species-level trees with representative sampling, ProSSE ML estimates have similar biases but consistently narrower bounds of errors than PBD ML estimates for the three parameters. In summary, ProSSE gives good estimates for the three parameters over a wide range of conditions.

Biological Meaning of the "Good" State

ProSSE improves the estimation of speciation initiation rate, extinction rate, and speciation completion rate in the protracted speciation and extinction process by removing the difference in evolutionary dynamics

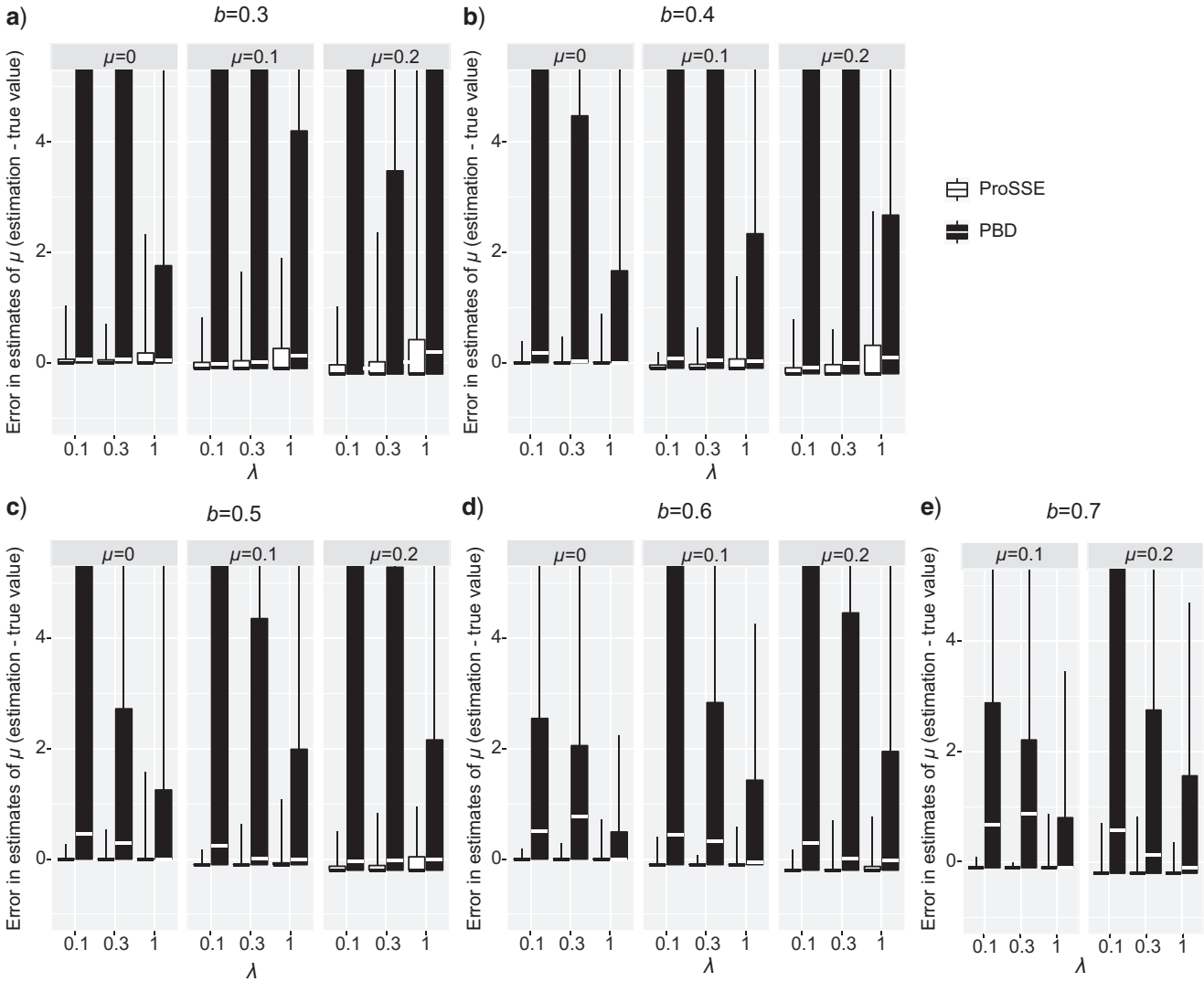

FIGURE 6. Error in ML estimates of the extinction rate μ for each simulation scenario using species-level trees. See figure details in the legend of Figure 5.

between the "good" lineage and the "incipient" lineage in PBD, so that we can model the process along lineage-level trees without the need to identify the "good" lineage of each extant species. Here, we discuss the biological meaning of the "good" state, which should not be the state of a single lineage, but an indicator to distinguish lineages of one species from lineages of another species. In ProSSE, we call it the "representative" state to avoid confusion.

When Etienne and Rosindell (2012) first described the observed protraction in speciation by PBD, they used a master equation to model changes in the number of incipient lineages and the number "good" species over time. Below is the master equation with modified notations.

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}[N_g,N_i;t]}{\mathrm{d}t}&=b_gN_g\mathbb{P}[N_g,N_i-1;t]+b_i(N_i-1)\mathbb{P}[N_g,N_i-1;t]\\&+\lambda(N_i+1)\mathbb{P}[N_g-1,N_i+1;t]\\&+\mu_g(N_g+1)\mathbb{P}[N_g+1,N_i;t]+\mu_i(N_i+1)\mathbb{P}[N_g,N_i+1;t]\\&-[(b_g+\mu_g)N_g+(b_i+\mu_i+\lambda)N_i]\mathbb{P}[N_g,N_i;t]\end{aligned}$$

where N_g and N_i are the numbers of good species and incipient lineages; b_g and b_i are the speciation initiation rates of a good species and an incipient lineage, that is, the rate of generating a new incipient lineage from an existing good species and an existing incipient lineage; μ_g and μ_i are the extinction rate of a good species and an incipient lineage; λ is the speciation completion rate, that is, the rate of an existing incipient lineage becoming a good species. Note that this equation does not assume the existence of a "good" lineage in a species, instead it simply defines a "good" species as the species that has completed speciation and so is distinct from the other species. Actually, the concept of "good" lineage was not introduced until Lambert et al. (2015) derived

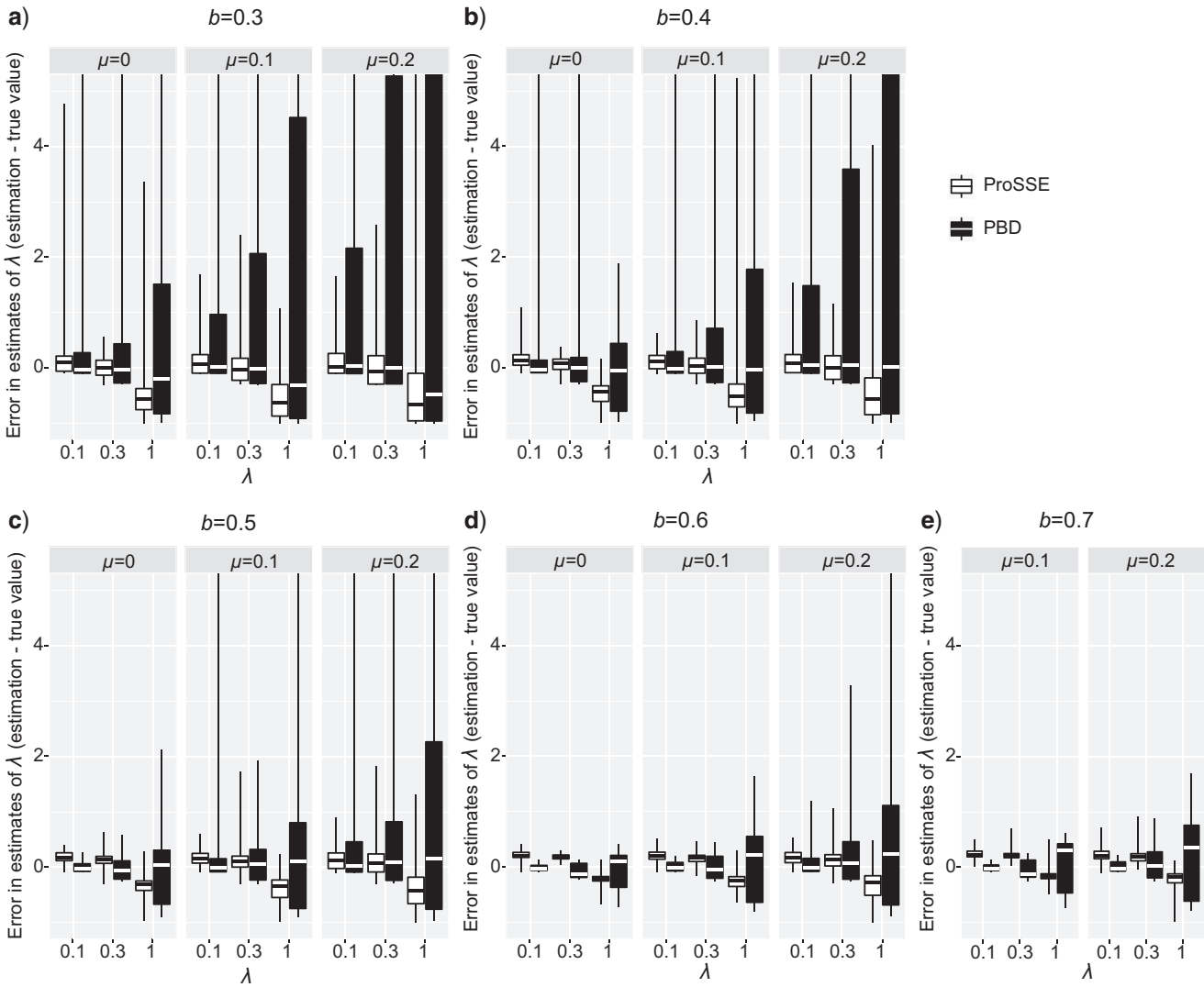

FIGURE 7. Error in ML estimates of the speciation completion rate λ for each simulation scenario using species-level trees. See figure details in the legend of Figure 5.

the likelihood function of PBD given a tree, because each edge of a tree has to be a lineage of a species.

Given that N_g is the number of good species, not good lineages, it makes sense why a good species has different speciation initiation rate and extinction rate to an incipient lineage. This is simply because a good species can have multiple lineages, so each lineage of the species can generate new incipient lineages and the species does not go extinct until all its lineages go extinct. Also, the equation does not assume that a "good" species cannot become a distinct species to itself. It does not include the term to describe a speciation completion event in a good species, because such an event does not change N_g or N_i. Therefore, assuming that all lineages have the same evolutionary dynamics does not make ProSSE incompatible to the original model of PBD. By explicitly modeling the evolutionary dynamics along each lineage of a species, ProSSE accounts for different speciation initiation rate and extinction rate between a good species and an incipient lineage in a more biologically realistic way, because the two rates of a good species should depend on the number of lineages in that species, rather than sharing the same values with other species in the master equation of PBD.

Assumptions of ProSSE

ProSSE has three main assumptions. The first assumption, complete sampling for lineage-level trees, is relatively minor as the assumption can easily be relaxed in a similar way to how we account for representative sampling in species-level trees. The probability that a lineage of species i is not included in the tree can be described by $E_i(t)$ using the same equation as for $E(t)$, except that the initial condition is not zero but the fraction of unsampled lineage of species i (FitzJohn et al. 2009). For species with different sampling

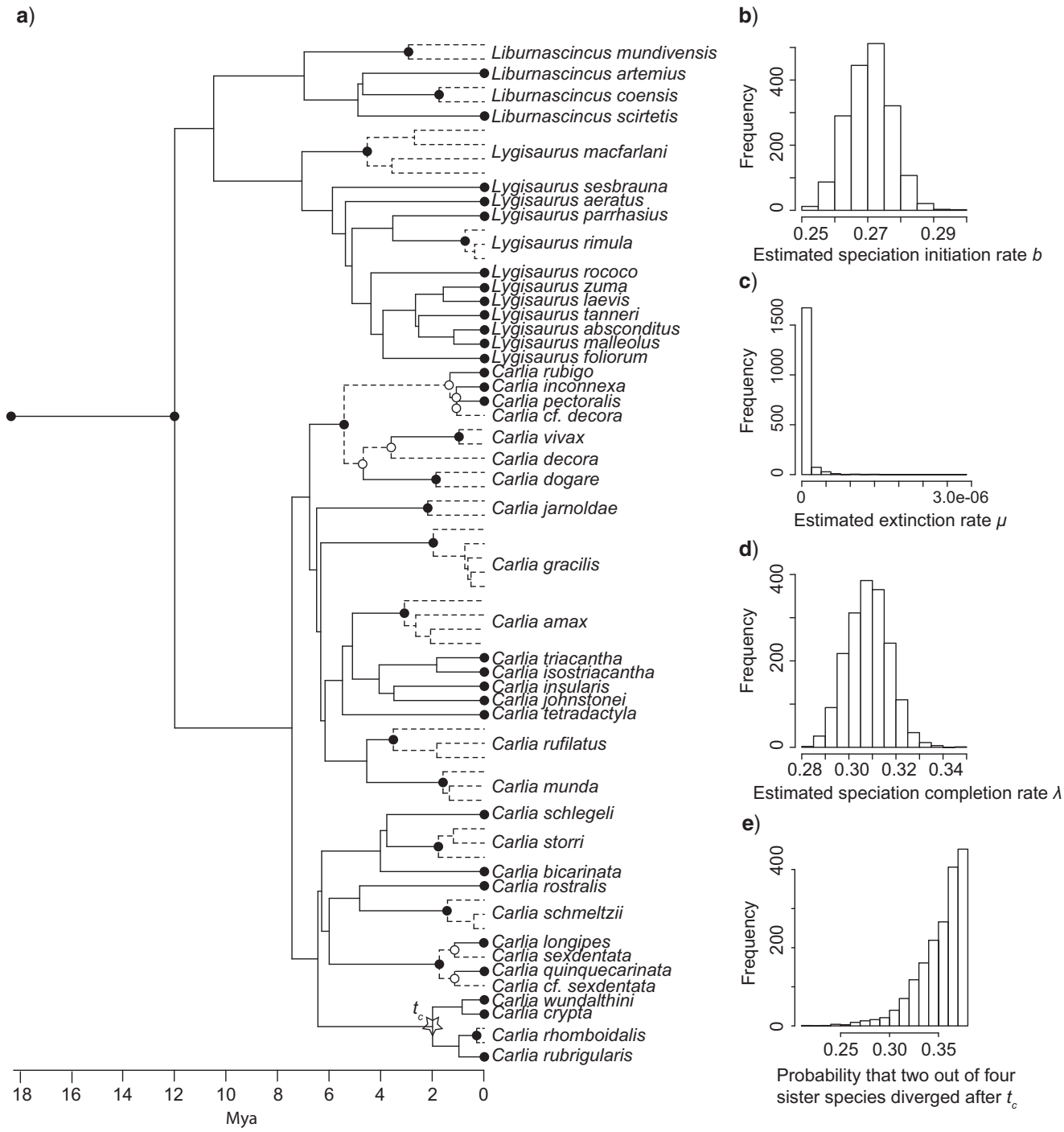

FIGURE 8. Maximum clade credibility tree of Australian rainbow skinks on lineage level (a) and the results of ProSSE analysis on the tree (b–d). The information we know from the tree is marked in (a), with dashed edges belong to the same species, nodes with black circles in representative state, and nodes with white circles in incipient state. Plots b–d) show the distribution of ProSSE ML estimates for speciation initiation rate (b), extinction rate (c), and speciation completion rate (d) over 1800 posterior samples of the tree. Under these sets of parameter values, e) plots the probability that two out of four randomly sampled pairs of sister species have divergence time lower than that between *C. crypta* and *C. rubrigularis* (t_c, the node height indicated by a star in a).

fractions, we need to use separate equations for lineages belonging to these different species, similar to what we have done in equations (5–8) to treat lineages not belonging to any extant species and lineages belonging to an extant species separately. So at the root of the tree, we will have the joint probabilities of the tree and the root belonging to a certain extant species, in addition to the joint probability of the tree and the root not belonging to any extant species, as we did for species-level trees.

The two remaining assumptions of ProSSE are 1) that the species identity of each extant lineage is known; and 2) that all lineages in the species group have similar rates and types of speciation processes. These two assumptions do not hold for many species groups. But because ProSSE treats each edge in a tree separately, we are able to relax these assumptions under the same model framework. Below, we discuss how to extend ProSSE to relax these assumptions.

Simultaneous Inference for Species identity and Speciation Process

ProSSE assumes that the species identity of each extant lineage is known (a related problem in species tree inference is discussed in O'Meara 2010). We can relax this assumption by simultaneously inferring species identities and parameters in ProSSE. This can be done by introducing a random variable for each extant lineage with unknown species identity, which has the set of all possible species identities for the lineage, including all extant species and a unique species identity for each lineage with unknown species identity. Then, a Markov chain Monte Carlo algorithm can be applied to approximate the joint probability distribution of the ProSSE parameters and these random variables. Because the probability of each of these random variables, conditional on the other random variables and the ProSSE parameters, can be calculated analytically from Bayes' theorem and the ProSSE likelihood function. Gibbs sampling can be applied to each random variable sequentially, whereas ProSSE parameters are sampled by a Metropolis–Hastings algorithm. As a result, the proportion of posterior samples of a random variable that takes a specific species identity gives the probability that the corresponding extant lineage has the species identity. This will allow usage of lineages of known species identity to inform ProSSE parameters and to delimit species in the tree for lineages with uncertain species identities. This approach is different from the approach taken in Sukumaran et al. (2021). Our approach will jointly infer species identities and parameters in ProSSE using information from the whole tree because species identities and speciation completion rate depend on each other. In contrast, the approach by Sukumaran et al. (2021) first infers speciation completion rate from a subset of the tree that only consists of lineages with known species identities, and then infers species identities conditional on the inferred speciation completion rate. To infer speciation completion rate from this subset of lineage-level tree, they developed an algorithm that requires listing all possible histories of speciation completion events along the tree, which is much more time-consuming than the analytical solution of ProSSE.

Accounting for Different Speciation Processes in the Tree

Another critical assumption of ProSSE is that all lineages in the species group have similar rates and types of speciation processes, so that the same set of ProSSE parameter values is fitted to all lineages. ProSSE can be extended to relax this assumption for two reasons. First, ProSSE models the branching events on the tree and the speciation completion events as independent processes. This can be seen from the solution to equations (1–3), where $A(t,s)$ is exactly the solution to constant-rate birth–death model. Factoring out $A(t,s)$ in the calculation along edges, we can write the final ProSSE likelihood as the product of the probability of the tree under a birth–death model and the probability of observing the tip species identities given the tree. This indicates that ProSSE can be easily adapted to any birth–death model with lineage-independent rates by replacing the likelihood of the constant-rate birth–death model with the likelihood of the chosen model, including the likelihood function for congruent classes of birth–death models proposed by Louca and Pennell (2020).

Second, ProSSE uses the SSE approach, which is designed to account for variation in rates among lineages (Maddison et al. 2007), so ProSSE can be easily extended to account for variation in speciation completion rates over lineages. For example, although the majority of Australian rainbow skinks are generalists and distributed in the Australian tropical savanna, there are still some species distributed in rainforests, potentially having undergone allopatric speciation in glacial refugia (Graham et al. 2006), and some other adapted to rock habitat that potentially have undergone ecological speciation (Blom et al. 2016). Because the relative prevalence of different speciation processes is largely associated with habitats, we can use the habitat of extant lineages to inform variation in speciation processes among lineages.

For example, there are a large number of cryptic lizard species in the Australian rainforests (Moritz et al. 2009), suggesting that most speciation in the area is completed by the accumulation of incompatible genes in allopatry. If this is true, then the speciation completion rate in the rainforests should be slower than the other habitats. We can account for this variation by fitting different speciation completion rates to different habitats, as Goldberg et al. (2011) did to associate geographic range evolution and diversification in their geographic state speciation and extinction model.

For another example, Blom et al. (2016) found that, in *Cryptoblepharus*, a genus closely related to the rainbow skinks in Australia, adaptation from arboreal to rock habitat repeatedly promoted adaptive diversification, whereas speciation within either arboreal or rock habitat resulted in species with similar morphology. This suggests that adaptation to rock habitat drives ecological speciation, which sometimes happens so rapidly that shifts to rock habitat co-occur with speciation completion events. To account for ecological speciation, we can introduce a new parameter for how often shift to rock habitat co-occur with speciation completion event and modify ProSSE in the same way as Magnuson-Ford and Otto (2012) did to link trait changes to speciation event in their BiSSE-node enhanced state shift model (BiSSE-ness).

In summary, by using the SSE approach to model protracted speciation and extinction process, ProSSE gives accurate estimates of speciation initiation rate, extinction rate, and speciation completion rate. It is able to be extended to account for variation in speciation processes across lineages in a species group and to use not only genomic data, but other data types, such as ecological information of each tip lineage, to delimit species in a probabilistic way (Sukumaran and Knowles 2017). These properties of ProSSE make it a promising analytical tool for biologists to study speciation processes under a phylogenetic framework.

Data Availability

The source code of ProSSE and all the codes to generate the results are available at github.com/huaxia1985/ProSSE.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.59zw3r27p.

Funding

This work was supported by the Australia Research Council [DE190100491].

Acknowledgments

We thank Craig Moritz's lab for providing the unpublished tree of the Australian rainbow skinks. We thank Rampal S. Etienne for providing the performance data for PBD. We thank Craig Mortiz, Lindell Bromham, Jason Bragg, and Sally Potter for valuable comments on the manuscript.

Author Contribution

All the authors contributed critically to the method development and the drafts and gave final approval for publication. X.H. conceived the project, led method development, programmed the method, and led the writing of the manuscript. T.H. and C.B. participated in the method development and the writing. An earlier version of ProSSE for lineage-level trees is developed in T.H.'s master thesis, supervised by X.H. and C.B.

References

Avise J.C. 1999. Phylogeography: the history and formation of species. Cambridge (MA): Harvard University Press.

Blom M.P.K., Horner P., Moritz C. 2016. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. Proc. R. Soc. B 283:20160181.

Bragg J.G., Potter S., Afonso Silva A.C., Hoskin C.J., Bai B.Y.H., Mortiz C. 2018. Phylogenomics of a rapid radiation: the Australian rainbow skinks. BMC Evol. Biol. 18:15–26.

Etienne R.S., Morlon H., Lambert A. 2014. Estimating the duration of speciation from phylogenies. Evolution 68:2430–2440.

Etienne R.S., Rosindell J. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61:204–213.

FitzJohn R.G. 2009. Quantitative traits and diversification. Syst. Biol. 59:619—633.

FitzJohn R.G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3:1084–1092.

FitzJohn R.G., Maddison W.P., Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

Gavrilets S. 2014. Models of speciation: where are we now? J. Hered. 105:743–755.

Goldberg E.E., Lancaster L.T., Ree R.H. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Syst. Biol. 60:451–465.

Graham C.H., Moritz C., Williams S.E. 2006. Habitat history improves prediction of biodiversity in rainforest fauna. Proc. Natl. Acad. Sci. USA 103:632–636.

Huson D.H., Rupp R., Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms and applications. Cambridge (UK): Cambridge University Press.

Lambert A., Morlon H., Etienne R.S. 2015. The reconstructed tree in the lineage-based model of protracted speciation. J. Math. Biol. 70:367–397.

Lambert A., Stadler T. 2013. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. Theor. Popul. Biol. 90:113–128.

Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580:502–505.

Louca S., Pennell M.W. 2021. Why extinction estimates from extant phylogenies are so often zero. Curr. Biol. 31:3168–3173.

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

Magnuson-Ford K., Otto S.P. 2012. Linking the investigations of character evolution and species diversification. Am. Nat. 180:225–245.

Marie Curie SPECIATION Network. 2012. What do we need to know about speciation? Trends Ecol. Evol. 27:27–39.

Moritz C., Hoskin C.J., MacKenzie J.B., Phillips B.L., Tonione M., Silva N., VanDerWal J., Williams S.E., Graham C.H. 2009. Identification and dynamics of a cryptic suture zone in tropical rainforest. Proc. R. Soc. B. 276:1235–1244.

Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344:305–311.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59–73.

Ricklefs R.E. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22:601–610.

Rosindell J., Cornell S.J., Hubbell S.P., Etienne R.S. 2010. Protracted speciation revitalizes the neutral theory of biodiversity. Ecol. Lett. 13:716–727.

Simonet C., Scherrer R., Rego-Costa A., Etienne R.S. 2018. Robustness of the approximate likelihood of the protracted speciation model. J. Evol. Biol. 31:469–479.

Singhal S., Hoskin C.J., Couper P., Potter S., Moritz C. 2018. A framework for resolving cryptic species: a case study from the lizards of the Australian Wet Tropics. Syst. Biol. 67:1061–1075.

Singhal S., Moritz C. 2013. Reproductive isolation between phylogeographic lineages scales with divergence. Proc. R. Soc. B. 280: 20132246.

Stadler T. 2010. Sampling-through-time in birth–death trees. J. Theor. Biol. 267:396–404.

Sukumaran J., Holder M.T., Knowles L.L. 2021. Incorporating the speciation process into species delimitation. PLoS Comput. Biol. 19:e1008924.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. USA 114: 1607–1612.