Syst. Biol. 74(5):723–739, 2025

https://doi.org/10.1093/sysbio/syaf004
Advance Access Publication January 26, 2025

A Phylogenetic Approach to Delimitate Species in a Probabilistic Way

Xia Hua[1,*] and Craig Moritz[2]

[1]*Mathematical Sciences Institute, 145 Science Road, The Australian National University, Canberra, ACT 2601, Australia*
[2]*Division of Ecology & Evolution, Research School of Biology, 46 Sullivans Creek Road, The Australian National University, Canberra, ACT 2601, Australia*
[*]*Correspondence to be sent to: Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia; E-mail: xia.hua@anu.edu.au.*

Received 11 September 2023; accepted 23 January 2025
Associate Editor: Bryan Carstens

Abstract.—Different species concepts and their associated criteria have been used to delimit species boundaries, such as the absence of gene flow for the biological species concept and the presence of morphological distinction for the morphological species concept. The need for different delimitation criteria largely reflects the fact that species are generated under various speciation mechanisms. A key question is how to make species delimitation consistent in a species group, especially when we want to delimit the species boundaries over many newly discovered evolutionary lineages and add these new lineages into a comparative analysis. Instead of forcing a single definition of "species," we can acknowledge different delimitation criteria by modeling how fast lineages in a species group evolve to meet these criteria along a phylogenetic tree. This study presents such a new model and a new delimitation approach that calculates the probability of each possible species identity of a lineage. We use simulations to show that our likelihood function gives accurate estimates of parameters in the model and our approach has high power to correctly identify species identities. We apply the approach to lineages in 2 real species groups that already have genomic and morphological evidence for their species identities. Our approach gives consistent inference of species identities with these existing pieces of evidence. We also demonstrate how to use our model to test a popular hypothesis about speciation process across all lineages in a species group and discuss further extension of the model to study speciation. [ProSSE; protracted birth-death process; speciation completion rate; speciation-based delimitation; The Bateson-Dobzhansky-Muller model.]

Species delimitation is about when and how to identify a given evolutionary lineage as a species distinct from other such lineages. This is the critical first step for any study at the species level, and crossing scales from micro to macroevolution. Yet, species delimitation is a notoriously hard problem, because different criteria on different data types are used to delimit species boundaries under different species concepts (de Queiroz 2007). For example, genomic data are used to delimit species boundaries under the general lineage concept of species (de Queiroz 1998) when the observed genetic variation among individuals is more likely to be generated from multiple genetic populations than from a single one, known as multispecies coalescent methods (Yang and Rannala 2010; Zhang et al. 2011; Jackson et al. 2017). Genetic composition and clines of individuals sampled from contact zones are used to test the absence of hybridization or gene flow as evidence for reproductive isolation, which delimits species boundaries under the biological species concept (Barton and Gale 1993; Pritchard et al. 2000; Anderson and Thompson 2002). Phenotypic data are used to delimit species boundaries under the morphological species concept when there are qualitative traits that differ among groups of sampled individuals but are fixed within each group (Wiens and Servedio 2010) or when there are quantitative traits whose observed variation among sampled individuals follows different normal distributions, rather than one normal distribution as expected for polygenic traits under random mating (reviewed by Cadena et al. 2018). However, not all distinct species diverge in their phenotypic traits; the phenomenon of cryptic species, is broadly defined as species that are morphologically and ecologically similar but genetically divergent in Singhal et al. (2018). Similarly, while ecological information such as differences in habitat use or in life histories is used to delimit species boundaries under the ecological species concept (Van Valen 1976), many cryptic species may not differ at all in these aspects.

Often, we include species in a comparative analysis that are defined by different delimitation criteria. For example, we may include in the same phylogenetic tree previously recognized species based on their morphological and/or ecological distinctiveness, together with newly elevated, and largely cryptic, species recognized by their genomic data. Such a practice presents us with 2 challenging problems to solve. First, what happens when we have newly discovered lineages? How do we delimit the species boundaries across these new lineages in order to keep our approach to delimitation consistent before and after these new lineages are added to the analysis? Second, we often use such a phylogenetic tree to study speciation dynamics of a species group, for example, to ask what affects speciation rate across the tree. How do we interpret variations in speciation rate when tip species are identified by different delimitation

criteria? What happens when the effects of a factor on speciation rate depend on the delimitation criteria?

In this study, we develop a phylogenetic modeling approach to solve both challenges. The basic idea is to model how fast lineages in a species group evolve to meet different delimitation criteria along a phylogenetic tree. For example, cryptic species are usually identified by the presence of genetic divergence under the general lineage concept of species and/or genetic evidence for reproductive isolation under biological species concept (Fišer et al. 2018). So cryptic species, regardless of their actual speciation mechanisms, likely meet these delimitation criteria through a gradual process, either by accumulating genetic differences or intrinsic genomic incompatibilities between populations. In contrast, species identified by their ecological or morphological distinctiveness likely meet the delimitation criteria through habitat shifts or trait evolution. If we can model how species evolve to meet different delimitation criteria along the phylogeny, then we can use the model to mathematically describe how we delimit the species boundaries in the species group and to predict the probability that a new lineage is a distinct species. This solves the first challenge. This solution is particularly useful for detecting cryptic species, which often requires extensive sequencing of many samples over the distribution range of the new lineages, ideally including contact zones. When we do not have this opportunity to get the data, the modeling approach we present here offers a quick way to detect cryptic species in the new lineages, which is based on how fast the other lineages in the species group have evolved to meet the delimitation criteria for cryptic species.

Modeling how fast lineages evolve to meet different delimitation criteria also helps us solve the second challenge, because we now have a more specific definition of speciation rate, which is the rate at which a lineage evolves to meet a particular delimitation criterion. We can then test the effect of a factor on this rate for different delimitation criteria separately. In this study, we demonstrate this idea by testing the Bateson–Dobzhansky–Muller model (BDM model; Bateson 1909; Dobzhansky 1937; Muller 1942). The BDM model is the most popular model for the accumulation of intrinsic incompatibilities and has a clear prediction that how fast a lineage evolves intrinsic reproductive isolation is proportional to its mutation rate. This prediction is made by mathematically modeling the accumulation of BDM incompatibilities between populations, conditional on the initiation of speciation, and predicting how long it takes for the number of incompatibilities to reach a threshold for reproductive isolation (Gavrilets 2003). Because we have modeled how fast lineages evolve to meet different delimitation criteria including intrinsic reproductive isolation and we know which extant species are cryptic species identified by the criterion of reproductive isolation, we can test the prediction of the BDM model by allowing the rate of evolving intrinsic reproductive isolation vary with mutation rate in our model and fitting the model to the phylogenetic tree and the taxonomic information of the species group.

To understand how fast lineages evolve to meet different delimitation criteria, we need lineages that have met the criteria, that is, interspecific lineages, as well as lineages that have not met the criteria, that is, intraspecific lineages, and we need a model to describe how intraspecific lineages becomes interspecific. A useful model is the protracted birth-death model (Etienne and Rosindell 2012). In the model, a newly arising lineage takes time to become a distinct species to its ancestral species. The new lineage can be thought of as a population newly isolated from the other populations of the same species. This event is called a speciation initiation event (a node in Fig. 1a). The lineage is regarded as in an "incipient" state before a speciation completion event (marked by an arrow from the ancestral species identity to a new species identity in Fig. 1a) turns the lineage into a 'good' species (shaded in Fig. 1a) that is distinct to its ancestral species. Regardless of how a new lineage arises, the model uses a speciation initiation rate b to describe how often a speciation initiation event occurs. Similarly, regardless of what speciation mechanisms are involved, the model uses a speciation completion rate λ to describe the time duration between a speciation initiation event and the following speciation completion event. Meanwhile, both "incipient" lineages and "good" species may go extinct but at different extinction rates (μ_I and μ_G in Fig. 1d).

Hua et al. (2022) argued that in order to fit the protracted birth death model to a phylogenetic tree with both intraspecific and interspecific lineages (Fig. 1b), unless we have additional information, all intraspecific lineages of a species should be treated equally, that is, they have equal speciation initiation rate b, speciation completion rate λ, and extinction rate μ (Fig. 1e). This is in contrast with fitting the model to a phylogenetic tree with only interspecific lineages (Fig. 1a after removing the gray edges), which requires us to define the lineage that leads to a new species in the tree as a "good" lineage (Lambert et al. 2015; edges in solid lines in Fig.1a), of which extinction means the extinction of all the lineages of the species and so its extinction rate μ_G is different from that of "incipient" lineages μ_I (Fig. 1d; "incipient" lineages are edges in dashed lines in Fig. 1a). The exact likelihood function for the protracted birth death model on a tree with both intraspecific and interspecific lineages is given by the ProSSE model (Hua et al. 2022) when all the 3 rates are constant across the tree. In this study, we extend ProSSE to allow lineages of different species to have different speciation completion rates, that is, different rates to evolve to meet different delimitation criteria in the context of this study. We demonstrate how to extend ProSSE to delimit species boundaries and test the BDM model.

We note that the idea of modeling protracted speciation for species delimitation is not new. Sukumaran et al. (2021) used the speciation completion rate estimated from lineages with known species identities to calculate

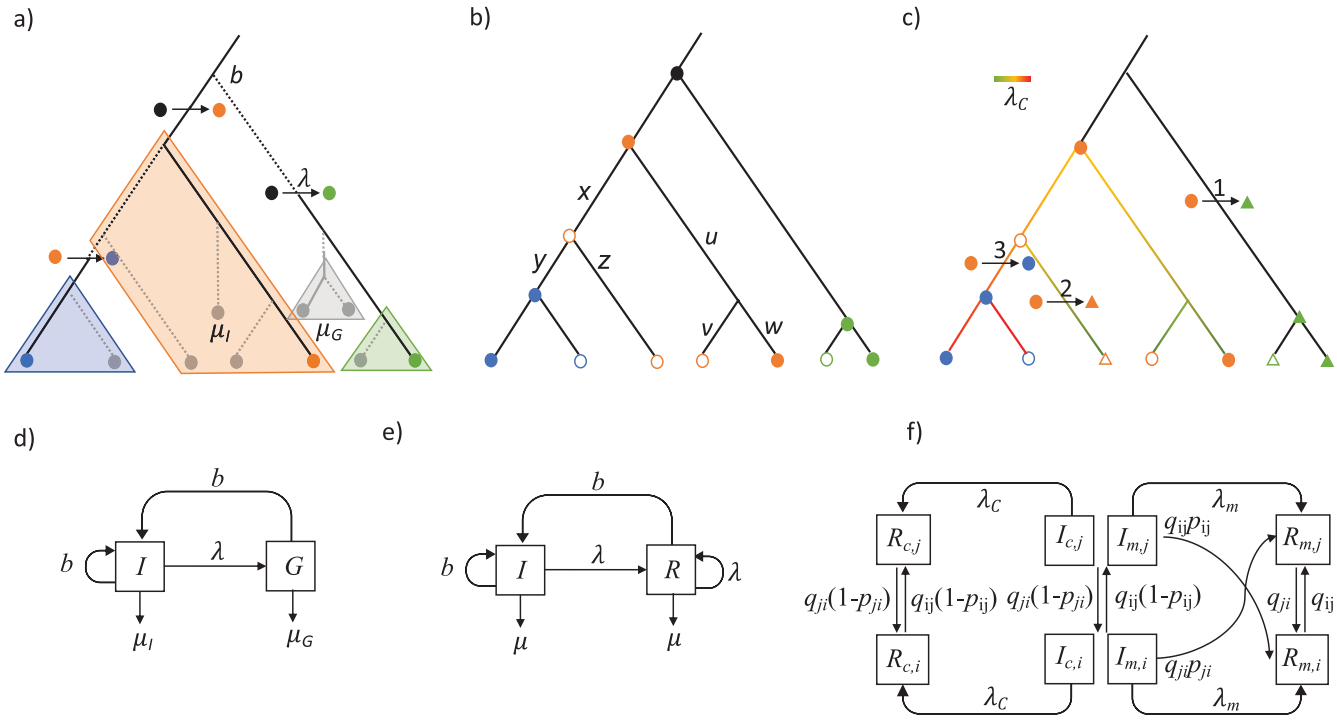

FIGURE 1. Illustration of protracted birth-death model (a, d), ProSSE algorithm (b, e), and extended ProSSE algorithm (c, f). In (a), there are three extant species and an extinct species. Lineages of these four species are enclosed by four polygons. A species went extinct at rate μ_G, which is different from the extinction rate of a lineage μ_I. A speciation process starts with the rise of an incipient lineage at rate b and completes at a speciation completion event with rate λ (marked by an arrow from the ancestral species to the new species). When we fit the protracted birth-death model to a tree where only one lineage of each species is included (coloured tip nodes), we need to define a good (G) lineage (solid edge) separately from an incipient (I) lineage (dashed edge) as in Lambert et al. (2015), such that the good lineage is a whole species and thus have extinction rate μ_G rather than μ_I. The dynamics between incipient (I) and good (G) states are described in (d). In (b), we have included all the lineages of each species and so all lineages now have the same extinction rate μ in (e). ProSSE randomly pick an extant lineage of each species (tip nodes with solid colors) as the representative (R) to its species. All the other lineages of each species are incipient (I). Different from G state, R state does not indicate different evolutionary dynamics (e). It is used only to track the number of distinct species at a time along the tree. In (c), ProSSE is extended to account for different delimitation criteria. For example, the tip lineages use different habitats or have different trait states (triangle and circle), which can be used as a delimitation criterion for ecological species, such as the green species. Suppose that the left two extant species in (a) have no morphological difference and their species boundaries are delimited by genomic isolation in (c), then these 2 species are cryptic species, so state I_c and state R_c are modeled along edges connecting lineages in the 2 species. The transition between the 2 states is governed by speciation completion rate λ_c (left diagram in f), which, according the BDM model, is correlated with mutation rate. To test BDM model, we allow λ_c to vary continuously across the tree (edges with color gradients). The remaining edges on the tree connect morphological species, so state I_m and state R_m are modeled by the right diagram in (f).

the probability that lineages with unknown species identities have completed speciation. They introduced this method, DELINEATE, as the first method of a new class of delimitation: speciation-based delimitation. However, DELINEATE has 3 key limitations. First, it requires enumeration of all possible configurations of speciation completion events across branches and/or all possible ways to partition lineages with unknown species identities into different species, so the method is computationally inefficient, especially in terms of memory requirements. Second, when there are lineages of known species identities, DELINEATE estimates the speciation completion rate only from these lineages. However, the estimation of the speciation completion rate is not independent of the lineages with unknown species identities. For example, if a lineage with an unknown species identity belongs to the same species as a lineage with a known species identity, then the speciation completion rate needs to be low enough to account for no speciation completion event since the split between the 2 lineages. Last and most important, DELINEATE assumes a constant speciation completion rate across lineages, an assumption that is not easy to relax under its methodological framework.

Our extended ProSSE does not have these limitations. It uses analytical solutions and so is efficient to apply to big phylogenies. It simultaneously fits the protracted speciation model and estimates the species identity for each lineage of interest. It allows different lineages to have different speciation completion rates. Importantly, rather than claiming to model the speciation process per se, we explicitly define a speciation completion event as the point when a lineage meets a delimitation criterion. With this new definition, we consider speciation-based delimitation as an approach that summarizes our current practice of species delimitation to lineages in a species group by estimating how fast these lineages have evolved to meet the delimitation criteria.

The estimated rates allow us to predict how likely it is that the remaining lineages in the species group have met the delimitation criteria. As a result, the constant speciation completion rate in DELINEATE is equivalent to assuming that all the lineages in a species group are delimited by the same criterion. Our extended ProSSE improves speciation-based delimitation by allowing different criteria applied to different lineages in a species group and estimating how fast a lineage evolves to each of these delimitation criteria. We use simulated data to quantify how much this improvement increases our power to correctly delimit species boundaries.

We demonstrate the applicability of extended ProSSE to real species groups, using 2 case studies: North American *Lionepha* ground beetles in the family Carabidae and Australian *Carlia* skinks in the subfamily Eugongylinae. We use *Lionepha* because Sukumaran et al. (2021) used this species group to demonstrate the applicability of DELINEATE, which allows us to compare the performance between DELINEATE and extended ProSSE on real species groups. We use *Carlia* because, similar to *Lionepha*, the multispecies coalescent tree is available for all phylogeographic lineages of the species group, with 46 tip lineages and 26 recognized species. Lineages in 4 out of 11 species complexes have been tested for some delimitation criteria. This allows us to ask if extended ProSSE can correctly predict the results of these tests. Different from *Lionepha*, *Carlia* species are thought to have undergone different speciation mechanisms and so multiple delimitation criteria have been applied to the species group. This allows us to demonstrate the flexibility of extended ProSSE to estimate how fast a lineage evolves to each of the delimitation criteria. We also use *Carlia* to test the BDM model by allowing the rate of evolving intrinsic reproductive isolation vary with mutation rate as predicted by theory (Gavrilets 2003). This is possible because 1) cryptic species in *Carlia* have been delimited by the criterion of absence of gene flow, so these cryptic species have likely evolved intrinsic reproductive isolation; 2) in *Carlia* and related genera, mutation rate varies with temperature among species (Ivan et al. 2021), so we have prior information on the mutation rate at each tip lineage in *Carlia*.

MATERIALS AND METHODS

We will first introduce the general framework of the method and then describe in detail how to model protracted speciation with different delimitation criteria, how to account for uncertainties in lineages' species identities, and how to test the BDM model. The method takes a Bayesian approach to infer the posterior distribution of model parameters, summarized in θ, and the species identities of lineages of interest, summarized in ξ'. The input is the time tree connecting all the lineages in the species group τ, the known species identities of some lineages in the tree ξ, and any information that indicates the delimitation criteria used on these lineages, summarized in X. The Bayesian framework is

$$f(\theta,\xi' \mid \tau,\xi,X) \propto f(\tau,\xi \mid \xi',\theta,X)\, f(\xi' \mid \theta,X)\, f(\theta \mid X)$$
$$= f(\tau,\xi,\xi' \mid \theta,X)\, f(\theta \mid X),$$

where $f(\tau,\xi,\xi' \mid \theta,X)$ is the core of our method, which uses the ProSSE model (Hua et al. 2022) and extends it to allow lineages to have different rates to evolve to meet different delimitation criteria. In this paper, we call our new model "extended ProSSE," to distinguish it from the ProSSE model in Hua et al. (2022). But the ProSSE model is just a special case of extended ProSSE, so we consider all cases of extended ProSSE under the "ProSSE" method.

Extended ProSSE

ProSSE (Hua et al. 2022) calculates the probability that a tree and the species identities of the extant lineages in the tree are generated under the protracted birth–death model (Etienne and Rosindell 2012). The main idea behind ProSSE is to track 2 states: representative (state R) and incipient (state I), along each edge of a tree from the present to the root. State R means the edge represents a species at a time, so, by definition, the most recent common ancestor (color-filled nodes in Fig. 1b) of an extant species is in state R. State I means the edge belongs to the same species as other edges in the tree at a time and one of these other edges is in state R. For example, in Fig. 1b, orange species have 3 extant lineages, 2 in state I (unfilled tip nodes) and one in state R (orange-filled tip node) at the present. Note that state R is introduced to track the number of distinct species at a time along the tree rather than indicate changes in the evolutionary dynamic of a lineage, so a lineage, either in state R or state I, has the same chance to complete speciation, which is modeled as a transition into state R with rate λ (Fig. 1e). Along an edge, ProSSE calculates $D_R(t)$ and $D_I(t)$, which are the probabilities that the edge in state R (resp. state I) at time t leads to extant descendants as observed in the tree. $D_R(t)$ and $D_I(t)$ are derived from equations (1), (3), (4) in Appendix 1. At a node where 2 edges join, $D_R(t)$ and $D_I(t)$ are updated by equations (2), (5), (6) in Appendix 1. The calculation continues until we reach the root. Since the root must be in state R (black-filled node in Fig. 1b), $D_R(t)$ at the root gives the joint probability of observing the tree τ and the species identity of each extant lineage ξ under the protracted birth death model, that is, $f(\tau,\xi \mid \theta)$ in the general framework, where θ consists of λ, b, and μ.

Because ProSSE considers each edge independently, it can be extended to allow different lineages to have different ways to transit into State R, which represents a speciation completion event, or the point when a lineage meets a delimitation criterion in the context of this study. Below, let us consider each type of delimitation criteria separately. The mathematical details of this extended ProSSE model are in Appendix 2.

When we use qualitative traits as diagnostic characters to distinguish species, a lineage turns into a distinct "morphological species" to its ancestral species when it transits from the ancestral trait state to a new trait state, which can

be modeled as co-occurrence between transition in trait states and speciation completion event (event 1 in Fig. 1c). To model how fast a lineage evolves to meet this delimitation criterion, we can introduce parameter q_{ij} for the rate of transition from trait state i to trait state j and use the trait states of the extant lineages to inform how large q_{ij} is. This is similar to a Markov model for discrete trait evolution (Pagel 1994; Fig. 1f). Note that when 2 sister lineages transit into the same trait state that is different from the ancestral trait state, our model will consider these 2 lineages as distinct species, but in practice, we will not, because we use the difference in morphological traits between sister lineages as the delimitation criterion. However, we usually use many diagnostic characters to distinguish species and 2 sister lineages will rarely transit to the same trait states for all the traits. So our model is still a good approximation to how we delimit morphological species using qualitative traits.

We can apply the same model to situations where we consider lineages occupying different habitats as distinct "ecological species," except now the trait states are replaced by habitats. However, habitat information alone is rarely used to delimit species boundaries, because habitat shifts in lineages within a species can happen as a result of habitat expansion (event 2 in Fig. 1c). To account for this, we include additional parameter p_{ij} for the probability that the transition from habitat i to habitat j co-occurs with speciation completion event (Fig. 1f). Using this parameter, we can separately model habitat shifts that co-occur with speciation completion events from those that do not co-occur with speciation completion events.

When we use quantitative traits as diagnostic characters to distinguish species, a lineage gradually evolves its trait value away from its ancestral trait value and turns into a distinct "morphological species" to its ancestral species when difference in their trait values are over a threshold, for example, when the difference is sufficiently larger than intraspecific variation in the trait values that the current trait values of the lineage belong to a distinct normal distribution to its ancestral trait values. How fast this happens can be estimated by fitting the speciation completion rate in the protracted birth death model to the phylogenetic relationship between species whose boundaries are delimited by the quantitative traits, which we denote as λ_m. The rate that 2 sister lineages become morphologically diverged from each other doubles λ_m, because both lineages can evolve away from their ancestral trait values.

Similarly, when we use the degree of genetic divergence or genetic evidence for intrinsic reproductive isolation to identify cryptic species, a lineage gradually accumulates genetic substitutions until it becomes genetically diverged from or genetically incompatible with its ancestral species. How fast this happens can also be estimated by fitting the speciation completion rate in the protracted birth death model to the phylogenetic relationship among cryptic species, which we denote as λ_c. The rate that 2 sister lineages become genetically diverged from each other doubles λ_c, because both lineages accumulate substitutions over time.

To separately estimate λ_c and λ_m, we denote state I and state R with a subscript c for cryptic species and with a subscript m for morphological species. Transition from state I_c to state R_c is through λ_c (Fig. 1f). Transition from state I_m to state R_m is through λ_m and/or transition in qualitative traits with rate q_{ij} (Fig. 1f). When habitat shifts are not associated with obvious morphological changes (e.g., in cryptic species), we can still model habitat shifts within state I_c and within state R_c (Fig. 1f). To calculate $D_R(t)$ and $D_I(t)$ along the tree, we follow state I_c and state R_c, using equations 7, 9, 11 in Appendix 2, for edges that link lineages in the same species complex (a group of cryptic species hidden under one species name). For the remaining edges on the tree, we follow state I_m to state R_m, using equation 7, with $D_c(t)$ and λ_c replaced by $D_m(t)$ and λ_m, and equations 13 and 15 in Appendix 2. This is because a species complex is often morphologically distinct from other species complexes and from other morphological species and so the most recent common ancestor of a species complex is a morphological species.

Accounting for Uncertainty in Species Identity

So far, we have used extended ProSSE to calculate $f(\tau, \xi \mid \theta, X)$, the probability of the tree that connects all the extant lineages of certain species identities ξ. When we are not certain about the species identity of some extant lineages ξ', $f(\tau, \xi \mid \theta, X)$ becomes $f(\tau, \xi, \xi' \mid \theta, X)$ and we want to use $f(\tau, \xi, \xi' \mid \theta, X)$ to calculate the probability of each possible species identity of these lineages. We use Gibbs sampling to approximate this probability distribution, in which the species identity of lineage l with uncertain species identity ξ'_l is sampled in sequence based on the probability $P(\xi'_l \mid \tau, \xi, \xi'_{-l}, \theta, X)$, where ξ'_{-l} are the species identities previously sampled for all the lineages with uncertain species identities except for lineage l. Using the definition of conditional probability,

$$P(\xi'_l \mid \tau, \xi, \xi'_{-l}, \theta, X) = \frac{f(\tau, \xi, \xi'_{-l}, \xi'_l \mid \theta, X)}{\sum_{\xi'_l \in S_l} f(\tau, \xi, \xi'_{-l}, \xi'_l \mid \theta, X)},$$

where S_l is the set of all possible species identities of lineage l, which can be predefined based on prior knowledge, for example, the lineage is split from a recognized species. When there is no prior knowledge, S_l includes any of the distinct species identities in ξ and ξ'_{-l}. In both cases, S_l also includes a species identity that is unique to lineage l, that is, lineage l belongs to a species that is different from any other lineages. For each ξ'_l, $f(\tau, \xi, \xi'_{-l}, \xi'_l \mid \theta, X)$ is calculated in the same way as $f(\tau, \xi \mid \theta, X)$ that we described in the previous sections. From the resulting Gibbs samples, we can calculate the probability that a lineage belongs to a certain extant species by counting how often the lineage belongs to the species in the samples.

Testing BDM Model

Under the BDM model, the rate of evolving intrinsic reproductive isolation is determined by the rate of

accumulating genetic incompatibilities, which depends on mutation rate (Gavrilets 2003). Specifically, the average waiting time for speciation to complete is proportional to the inverse of mutation rate (Gavrilets 2003), under both allopatric and parapatric speciation (assuming that the rate of gene flow is much larger than mutation rate). So, we expect the rate of evolving intrinsic reproductive isolation proportional to mutation rate if the BDM model is the genetic mechanism of intrinsic reproductive isolation.

To test the BDM model, we need to estimate the rate of evolving intrinsic reproductive isolation. In our extended ProSSE, parameter λ_c is used to model the rate of a lineage to evolve to meet the delimitation criteria for cryptic species. If nonallopatric cryptic species in a species complex are identified by the absence of gene flow among each other, then these cryptic species have likely evolved intrinsic reproductive isolation. We can use the phylogenetic relationship among these cryptic species to estimate parameter λ_c, which is now the rate of a lineage to evolve intrinsic reproductive isolation. To test the BDM model, we can incorporate the prediction of the BDM model into extended ProSSE by modeling λ_c as a function of mutation rate. We use $\lambda_c(r) = \beta e^r$, $\beta > 0$ as the function, where r is mutation rate on log scale. When mutation rate does not differ among cryptic species within a species complex, we can simplify the problem by assuming a constant mutation rate for each species complex and estimating β from variation in mutation rate and variation in λ_c among species complexes. Otherwise, we assume r evolves under Brownian Motion with drift term $\phi = 0$, and diffusion term $\sigma^2(r,t) = \sigma^2$. This model for mutation rate is also known as autocorrelated-rates model in the literature of molecular dating (Thorne et al. 1998) to account for varying substitution rates across a tree. We use this model because mutation rate is strictly positive and it is efficient to numerically integrate Brownian motion across the tree (FitzJohn 2009). The mathematical details of the model are in Appendix 3. Similar to λ c, parameter β and σ^2 are estimated from the phylogenetic relationships among cryptic species in each species complex. The mutation rate of each tip lineage, if available, improves our estimation of parameter σ^2. Alternatively, we can combine molecular dating and extended ProSSE to estimate σ^2 directly from genetic sequences (see Discussion).

Software Implementation

We implement the ProSSE method as an add-on to the R package "diversitree" (FitzJohn 2012). The source code and the instruction to use the ProSSE method are available at github.com/huaxia1985/ProSSE. Specifically, make.prosse function defines the model, with different types of inputs indicating different delimitation criteria. The base model is the ProSSE model, which has the same assumption as DELINEATE that speciation completion is governed by a single speciation completion rate. This can be used, for example, when all the species are morphological species delimitated by some quantitative traits. Building upon the base model, delimitation criterion by qualitative traits is added when users provide "traits" input, which gives the states of the qualitative traits for each tip lineage. The delimitation criterion for cryptic species is added when users provide "types" input, which gives the name of the species complex that each cryptic tip lineage belongs to. The speciation completion rate for cryptic species is a function of a continuous trait, when users provide "states" input, which gives the value of the continuous trait of each tip lineage. Uncertainty in the value of the continuous trait is accounted for when users provide "states.sd" input, which gives the standard deviation in the continuous trait value of each tip lineage. After using make.prosse function to define our model of delimitation criteria, mcmc.prosse function runs the MCMC analysis to use the model to infer species identities while estimating the parameters in the model.

Simulation Study

We test the performance of ProSSE method on simulated trees in 2 aspects: 1) Does the likelihood function give accurate estimates of parameters in the extended ProSSE model? 2) Does ProSSE method increase our power to correctly delimit species boundaries, compared to DELINEATE? We simulate trees under protracted birth–death model where speciation completion is defined by multiple delimitation criteria. Specifically, speciation completion events can co-occur with transitions between 2 states of a discrete trait and speciation completion rate depends on a continuous trait that evolves under a Brownian motion. This is designed to reflect how extended ProSSE models how fast lineages evolve to differ in discrete diagnostic characters or to occupy different habitats, how fast lineages accumulate a sufficient amount of morphological difference or genetic difference to be considered distinct species, and how the rate of evolving intrinsic reproductive isolation depends on mutation rate.

The simulation algorithm is based on "tree.quasse" function in the R package "diversitree" (Fitzjohn 2012), which simulates trees under a process where speciation rate and extinction rate can depend on a continuous trait that evolves under a Brownian motion (Fitzjohn 2009). The algorithm in "tree.quasse" scales time so that there are on average 500 time steps for the continuous trait to evolve between any 2 discrete events. While the discrete events in "tree.quasse" are speciation events and extinction events, the discrete events in our simulation are speciation initiation events with rate b, extinction events with rate μ, speciation completion events with rate $\lambda(r)$, and transition events between 2 states in a discrete trait with rate q_{12} and q_{21}. Speciation completion events also happen with probability p_{12} when the discrete trait transits from state 1 to state 2, and with probability p_{21} when the discrete trait transits from state 2 to state 1. We use $\lambda(r) = \beta e^r$, $\beta > 0$, the same function we used to test BDM model, where r is

assumed to evolve under Brownian Motion with $\phi = 0$, and $\sigma^2(r,t) = \sigma^2$.

To assess the accuracy of extended ProSSE model in parameter estimation, we fit extended ProSSE model to each simulated tree by searching for the maximum likelihood (ML) estimates of each parameter in the model using the find.mle function with default setting in "diversitree," and report the deviation in the ML estimates to the true parameter values used in the simulation (hereinafter referred to as the "error"). We simulate 100 trees under each of 30 combinations of parameter values, including 3 sets of speciation initiation rate and extinction rate: $(b,\ \mu) = (0.3,\ 0)$ or $(0.4,\ 0.1)$ or $(0.5,\ 0.1)$; 3 sets of trait transition rates: $(q_{12},\ q_{21}) = (0.1,\ 0.03)$ or $(0.1,\ 0.1)$ or $(0.03,\ 0.1)$; 2 probabilities that trait transition events co-occur with speciation completion events: $(p_{12},\ p_{21}) = (0,\ 1)$ or $(0.5,\ 0.5)$; 2 coefficient values between speciation completion rate and the continuous trait value: $\beta = 0.1$ or 1, with the diffusion rate of the trait σ^2 fixed to 0.1.

We decide the order of magnitude for speciation initiation rate based on the average speciation rate 0.15 per Myr estimated for major bird clades by Maliet et al. (2019). The values for trait transition rates relative to speciation initiation rate are chosen from the range of values that had been used by various studies (e.g., Maddison et al. 2007; Simpson et al. 2018) to test the effects of a binary state trait on speciation rate and extinction rate. For the coefficient between speciation completion rate and the continuous trait value β, its value equals the speciation completion rate at the root, because the continuous trait has value 0 at the root. So, we use previously estimated speciation completion rate to decide the order of magnitude for β. Using ProSSE, we estimated the speciation completion rate of *Carlia* skinks about 0.3/Myr (Hua et al. 2022). Using DELIENATE, Sukumaran et al. (2021) estimated the speciation completion rate of *Lionepha* beetles about 178.3 per branch length unit. Assuming that the 2 genera have the same origin date, the scaled speciation completion rate of *Lionepha* is about 0.28/Myr. So, −1 order of magnitude is biologically realistic for speciation completion rate. We choose the value for the diffusion rate of the continuous trait based on the variance of the substitution rates on the log scale on the third codon position estimated by Nabholz et al. (2016) across extant bird species because substitution rates on the third codon position approximates mutation rate according to the neutral theory of molecular evolution (Kimura 1983). Since birds originated at least 150 Ma, this variance is at least 150 times larger than the diffusion rate of mutation rate in birds if the mutation rate evolves as a Brownian motion. We use this unrealistically high value for diffusion rate because inferring correct species identities is harder with more variable speciation completion rates.

We assess the power of ProSSE method to correctly infer species identities by asking if ProSSE infers the same species identities of tip lineages as their "true" species identities in the simulation. To do this, we randomly pick 10% of tip lineages in each of the simulated trees as lineages with unknown species identities and no prior knowledge on the set of possible species identities, then apply ProSSE to estimate the probability that each of these tip lineages belong to its simulated species identity. We run this on 10 trees, each having less than 100 tip lineages, under each of the 30 simulation schemes. To compare the power of ProSSE with that of DELINEATE, we run DELINEATE on each of these simulated trees using default settings. The extended ProSSE model takes a relatively long computing time to account for varying speciation completion rates, because it requires numerical techniques to integrate Brownian motion. So, we only tested 10% of lineages with unknown species identities under varying speciation completion rates.

We examine the impact of the percentage of lineages with unknown species identities using lineage-level trees simulated in Hua et al. (2022), where speciation completion is at a constant rate across the tree, that is, a single delimitation criterion. These simulated trees comply with the assumption of DELINEATE. We use 100 trees, each having less than 100 tip, under each of 6 simulation schemes: combinations of the 3 sets of speciation initiation rate and extinction rate: $(b,\ \mu) = (0.3,\ 0)$ or $(0.4,\ 0.1)$ or $(0.5,\ 0.1)$, and 2 values of speciation completion rate $\lambda = 0.1$ or 1, which are comparable to $\beta = 0.1$ or 1. In each simulated tree, we randomly pick 10% or 30% tip lineages as lineages with unknown species identities, then apply both ProSSE and DELINEATE to infer their species identities. Note that we cannot run DELINEATE on trees with more than 50 species and over 10% lineages with unknown species identities, because it requires over 500 Gb memory, which is over the limit of our best computing facility. As a result, the reported power of DELINEATE is based on trees with less than 50 species, which potentially contain more intraspecific lineages, given that tree size is limited to 100. More intraspecific lineages improve estimation of speciation completion rate (Hua et al. 2022), so the reported power of DELINEATE may be overestimated.

For each simulated tree, slice sampling is used to sample parameter values from priors. We use a broad exponential prior (with rate 0.5) for each rate parameter (all nonnegative), including b, μ, σ^2, β, and all q_{ij}. We use a Jeffreys prior for each probability parameter (between 0 and 1), including all p_{ij}. Each MCMC is run for 5030 iterations with the first 30 iterations to determine the tuning parameter of the slice sampler, which is set to the distance between the 5% and 95% quantiles of the distribution of samples of each parameter in the first 30 iterations (FitzJohn 2012). Each chain converges quickly in the first 30 iterations and has low autocorrelation in the next 5000 posterior samples, with effective sample size for any parameter at least 1000.

Case study: Australian Carlia Skinks

Same as DELINEATE, our ProSSE method is built upon preliminary analyses including identifying discrete phylogeographic lineages, inferring their 'time'

tree, and establishing the species identities of some of these lineages. For Australian *Carlia* skinks, discrete phylogeographic lineages within taxonomically recognized species were defined based on the presence of divergent, geographically coherent mtDNA lineages, supported by evidence of strong divergence in genomic DNA from sequencing of 1000's of exons (Potter et al. 2016, 2018; Afonso Silva et al. 2017; Bragg et al. 2024). In total, the species group has 46 phylogeographic lineages and 26 recognized species. Bragg et al. (2018) published a preliminary multispecies coalescent tree of the species group and that tree to is now updated to include all the recognized species and all the phylogeographic lineages using StarBEAST2 (Ogilvie et al. 2017) by Bragg et al. (2024). Of the 46 lineages, 2 lineages in *C. gracilis* complex, 2 lineages in *C. rufilatus* complex, and 2 lineages in *C. munda* complex have evidence for hybridization and gene flow (Potter et al. 2018; Supplementary Table S1), so these lineages are considered as intraspecific lineages in their corresponding species complex, rather than lineages of uncertain species identities.

In the 26 recognized species, 23 are morphologically distinct from each other and 3 are essentially cryptic species, including *C. isostriacantha* recognized by Afonso Silva et al. (2017) from *C. triacantha*, *C. insularis* recognized by Afonso Silva et al. (2017) from *C. johnstonei*, and *C. crypta* recognized by Singhal et al. (2018) from *C. rubrigularis.* In addition to the 3 cryptic species, there is published preliminary evidence for the absence of gene flow among nonallopatric lineages in *C. gracilis* complex and *C. amax* complex (Supplementary Table S1). The evidence is preliminary because sampling is limited at parapatric boundaries of phylogeographic lineages in these species complexes. But it gives us some prior belief that these lineages are likely to be new cryptic species. So, we use ProSSE to estimate the species identities of these lineages and check if ProSSE can identify these lineages as new distinct species. We also use ProSSE to assess the completeness of speciation in island-only (allopatric) lineages, as well as lineages in species complexes (including *C. vivax, C. jarnoldae, C. storri, C. schmeltzii, C. rhomboidalis*) that have no evidence for or against the absence of gene flow. In addition, there are 2 specimens with uncertain species identities *C.* cf. *dogare* and *C.* cf. *sexdentata.* We use ProSSE to estimate their species identities too. In total, we use ProSSE to estimate species identities of 17 lineages, which is about 37% lineages in Australian *Carlia*.

Among the 23 morphological species, the most significant morphological difference is male breeding color and specifically blueish male throat color vs. all the other male throat colors (reddish color and achromatic color), because, in *Carlia*, blue females prefer blue males, but red and achromatic females show no preference (Dolman 2008). On the *Carlia* tree, there are at least 6 independent transitions between nonblue and blue male throat (Supplementary Figure S1), after which, blue species are further morphologically distinct in other aspects of color patterns and in a suite of additional characters. The same is true for nonblue species. Given that blue male throat indicates assortative mating, we model the transition from nonblue male throat color to blue male throat color to co-occur with speciation completion events. We model divergence in all the other morphological traits as a gradual process governed by λ_m.

In addition to morphological traits, we model the transition between rock habitat and grassland or forest-floor litter habitats to co-occur with speciation completion events, because habitat shift between rocks and grasslands could result in extrinsic reproductive isolation. This parallels the case in a related genus *Cryptoblepharus*, where species in rock habitats belong to distinct ecotypes to species in arboreal habitats (Blom et al. 2016). In *Carlia*, there are 2 distantly related lineages in rock habitat and each one is a distinct recognized species from its closest relative which are grass or litter dwellers (Supplementary Figure S1), indicating that there are at least 2 independent shifts to rock habitat and both result in species that may be delimitated under ecological species concept.

To model transitions in rock vs. grass/litter habitat and blue vs. nonblue male throat, we group the 2 traits into a single trait with 4 states: rock blue, rock nonblue, grass/litter blue, and grass/litter nonblue. We have not observed any species with some lineages only in rock habitats and others in grass/litter habitats, nor have we observed any species with some lineages having blue male throat and others having nonblue male throat, so we assume $p_{ij} = 1$, meaning that shift between rock and grass/litter habitats and evolving blue male throat always co-occur with speciation completion in Australian *Carlia*. We also assume that habitat shift is an independent event to change in male throat color, so double transitions, for example, from grass/litter nonblue to rock blue, are considered unlikely in a very small time interval compared to a single transition. Denoting transition rate from rock to grass/litter as q_{rg}, from grass/litter to rock as q_{gr}, from blue to nonblue as q_{bn}, from nonblue to blue as q_{nb}, off-diagonals in Q_R in Appendix 2 are:

	rock blue	rock nonblue	grass/litter blue	grass/litter nonblue
rock blue	--	q_{bn}	q_{rg}	0
rock nonblue	q_{nb}	--	0	q_{rg}
grass/litter blue	q_{gr}	0	--	q_{bn}
grass/litter nonblue	0	q_{gr}	q_{nb}	--

To test BDM model, we have prior knowledge on the mutation rate of each tip lineage from Ivan et al. (2021). The study used sister pair comparison to show that about half of the variation in the mutation rate (measured by synonymous substitution rate) across tip species in the Australian Eugongyline skinks can be explained by the regression model $r \sim r_0 + 0.56 \ln(T) + \mathcal{N}(0, 0.30)$, where r is mutation rate on log scale, T is the average annual mean temperature over a species' distribution range, and r_0 is the mean mutation rate on log scale. Although r_0 could not be estimated from the sister pair comparison, substituting the regression model to $\lambda_c(r) = \beta e^r$, e^{r_0} becomes a part of the coefficient β to estimate in our case study. MCMC analysis is applied to estimate these parameters, using the same settings as in the simulation study. To test if speciation completion rate in cryptic species λ_c depends on mutation rate, as predicted by the BDM model in the form of $\lambda_c(r) = \beta e^r$, we repeat the MCMC analysis with constant speciation completion rate $\lambda_c = \beta$, and compare their marginal likelihoods. The MCMC analysis on each of the 2 models was repeated on each of 100 random posterior samples of the multispecies coalescent tree of Australian *Carlia* skinks (Bragg et al. 2018;, 2024), to account for phylogenetic uncertainty. The joint posterior distribution of parameters in each model over the 100 trees is approximated by the posterior samples from the 100 MCMC analyses. The marginal likelihood of the model is estimated from these posterior samples through an ellipsoid to avoid the instability issue of the harmonic mean estimator (Reichl 2020), where the sampled tree is considered as a hidden variable.

Case study: North American Lionepha Beetles

We use *Lionepha* to compare the performance between DELINEATE and ProSSE on real species groups, because the *Carlia* dataset is too large for DELINEATE to run. Sukumaran et al. (2021) used *Lionepha* to demonstrate the applicability of DELINEATE, so we use their lineage level tree and apply ProSSE on the same lineages with uncertain species identities as they did. In total, there are 9 recognized species and 13 out of 56 lineages with uncertain species identities. Sukumaran et al. (2021) picked these 13 lineages as lineages with uncertain species identities because the inference of their species boundaries is ambiguous from their multispecies coalescent analysis. The alpha taxonomy of *Lionepha* recently reviewed by Maddison and Sproul (2020) gives a clear understanding of the species identities for the 13 lineages, which includes 2 recognized species. This allows Sukumaran et al. (2021) to check if DELINEATE gives the correct species identities to lineages in a real species group. Maddison and Sproul (2020) used the same species delimitation criteria to all the recognized species, including morphological divergence in a suit of diagnostic characters and genetic distinctiveness supported by STACEY analysis (Jones 2017). So, following Sukumaran et al. (2021), we assume a constant speciation completion rate across the tree. Under this assumption, DELINEATE and ProSSE are similar mathematical models implemented in different algorithms, so we expect ProSSE to give similar inference of species identities to DELINEATE.

RESULTS

Extended ProSSE Model Gives Accurate Estimates of Parameters

Figure 2 and Supplementary Fig. S2 show the errors in estimates of parameters in extended ProSSE. Under low speciation completion rate ($\beta = 0.1$), extended ProSSE gives median-unbiased consistent estimators of all the parameters (i.e., both the median and the range of errors diminish to close to zero with tree size; Supplementary Fig. S3). These include the coefficient of speciation completion rate for continuous traits β and the co-occurrence probability between speciation completion and transition in discrete traits p_{12}, p_{21} (Fig. 2), as well as the rate of evolution in both discrete and continuous traits (Supplementary Fig. S2). Under high speciation completion rate ($\beta = 1$), extended ProSSE still gives median-unbiased consistent estimators of the rate of evolution in both discrete and continuous traits (Supplementary Fig. S4), but the coefficient of speciation completion rate for a continuous trait is consistently overestimated (i.e., the median of errors diminishes to a nonzero value with tree size in Supplementary Fig. S4). This result is consistent with the performance of ProSSE (Hua et al. 2022) in that speciation completion rate becomes overestimated when the true speciation completion rate is higher, relative to the difference between speciation initiation rate and extinction rate $b - \mu$, as this leads to fewer extant species having intraspecific lineages and so less information on the upper bound of speciation completion rate.

Similarly, the co-occurrence probabilities between speciation completion and transition in discrete trait p_{12}, p_{21} are overestimated under high speciation completion rate ($\beta = 1$; Fig. 2), although the mean and the range of errors tend to reduce with tree size (Supplementary Fig. S4). For these parameters, the information on the upper bound is mostly from incipient lineages with the same trait state. As a result, the estimates of the co-occurrence probabilities are worst under the scenario where the trait transition event that co-occurs with speciation completion happens less frequently than the other trait transition event ($q_{12} = 0.03$, $q_{21} = 0.1, p_{12} = 1$, $p_{21} = 0$; dark green bars in Fig. 2), because we tend to observe more lineages with trait transitions that first co-occur with speciation completion and then back to the same state as its sister lineage. For the same reason, the estimates of the co-occurrence probabilities have larger error ranges under the scenarios where state transitions in traits do not always co-occur with speciation completion (light green and yellow bars in Fig. 2).

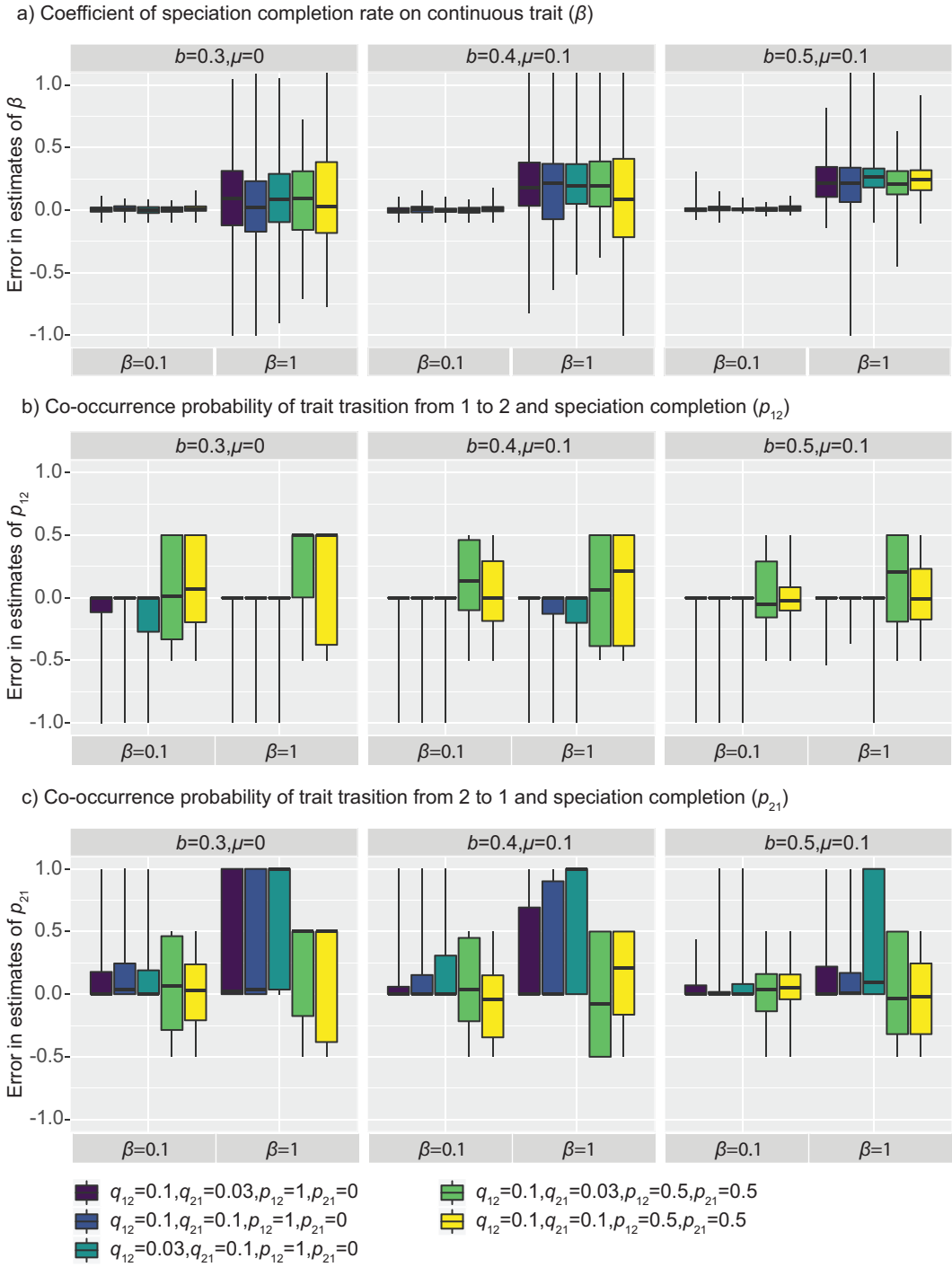

FIGURE 2. Error in ML estimates of the parameters related to speciation completion for each simulation scenario. These parameters are a) β in $\lambda(r) = \beta e^{r}$ that describes the dependence of speciation completion rate λ on the continuous trait r; b) p_{12} and c) p_{21}, the probability that speciation completion co-occurs with transition in trait states, from state 1 to 2 and from state 2 to 1. Each facet represents a group of simulation scenarios with the same speciation initiation rate b, extinction rate μ, and β. Within each facet, the 5 colored boxplots represent the 5 combinations of parameters used to model speciation completion associated with state transition in a discrete trait. Each boxplot represents the distribution of errors, the difference between the estimated and the true value over 100 simulated trees, showing the minimum, the maximum, the median, the first, and third quartiles of the distribution.

Speciation initiation rate b and extinction rate μ are effectively estimated from constant birth death process (Hua et al. 2022), because extended ProSSE models speciation initiation and extinction as independent events to speciation completion. So we do not discuss them here, but we can model speciation initiation

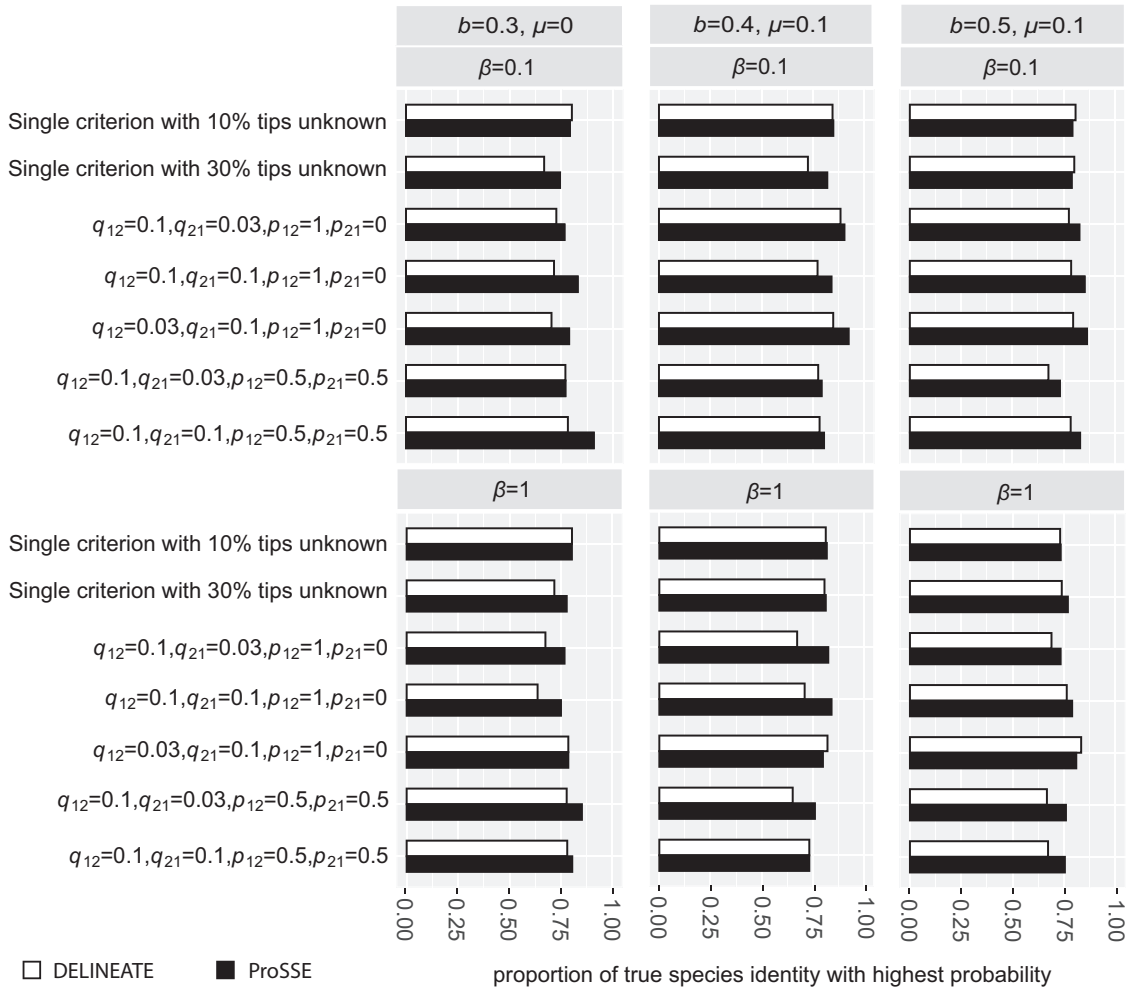

FIGURE 3. Power of extended ProSSE and DELINEATE to correctly identify species identities under each simulation scheme. Each facet represents a group of simulation scenarios with the same speciation initiation rate b, extinction rate μ, and β. Within each facet, the 7 sets of 2 bars represent the 7 simulation schemes, with the white bar showing the power of DELINEATE and the black bar showing the power of extended ProSSE.

rate and extinction rate as functions of discrete and continuous traits as the other SSE models have done, as long as we have good reasons why these rates are trait dependent.

ProSSE Method Outperforms DELINEATE

Figure 3 shows the power of ProSSE and DELINEATE to correctly identify species identities in simulations. Power is defined as the proportion of lineages, for which ProSSE or DELINEATE identifies their true species identities as the most probable species identity. For ProSSE, the most probable species identity is the one with the highest posterior probability. For DELINEATE, it is the one with highest marginal likelihood (Sukumaran et al. 2021). Of the 10% tip lineages with unknown species identities in each of the 300 trees simulated under multiple delimitation criteria, ProSSE has on average 81% power, 6% higher than DELINEATE (Fig. 3). In general, ProSSE method outperforms DELINEATE more under simulation schemes where extended ProSSE model has more accurate parameter estimation (compare Figs. 2 and 3).

For trees simulated under a single delimitation criterion that follows DELIENATE assumption, both methods have on average 80% power for 10% tip lineages with unknown species identities. For 30% tip lineages with unknown species identities, the average power of ProSSE remains at 80%, but the average power of DELINEATE drops to 74%. The drop in power of DELINEATE relative to ProSSE is most significant when speciation initiation rate b and speciation completion rate β is low, which generates smaller trees with fewer species. This is because DELINEATE estimates the speciation completion rate only from lineages of known species identities. Removing the lineages of unknown species identities from small trees with few species, leaves insufficient information to accurately estimate speciation completion rate.

TABLE 1. Delimitation results of ProSSE and DELINEATE for *Lionepha* beetles. The most recently reviewed taxonomy is given by Maddison and Sproul (2020), which recognized 2 new species *L.kavanaughi* and *L.tuulukwa*. Probability of exclusive conspecificity is the marginal likelihood in DELINEATE and the posterior probability in ProSSE that all the corresponding lineages are conspecific to the species suggested by Maddison and Sproul (2020) and no other lineages of uncertain species identities belong to the species. For lineages of *L.kavanaughi* and *L.tuulukwa*, DELINEATE infers a higher probability that they are conspecific to some other existing species. For the UT lineage of *L.probata*, ProSSE infers a higher probability that the lineage is a new species. The probabilities of these delimitations are given in parentheses for both DELINEATE and ProSSE.

Lineage name	Taxonomy by Maddison and Sproul (2020)	Probability of exclusive conspecificity	
		DELINEATE	ProSSE
coalescentpop008	*L.kavanaughi*	0.36	0.49
coalescentpop009		(0.38)	(0.27)
L_lindrothi_CA_Emerald_Lake	*L.lindrothi*	0.77	0.71
coalescentpop011			
L_lindrothi_CA_South_Fork_Bishop_Creek			
L_probata_UT_Shingle_Creek	*L.probata*	0.60	0.46
	(warrant further study)	(0.28)	(0.48)
L_disjuncta_OR_Mt_Hood	*L.disjuncta*	0.75	0.62
L_disjuncta_CA_Lily_Lake			
coalescentpop006			
L_tuulukwa_CA_Trinity_Alps	*L.tuulukwa*	0.36	0.6
L_tuulukwa_OR_Knowles_Creek		(0.44)	(0.27)
L_tuulukwa_OR_Marys_Peak			

Table 1 compares the delimitation results of ProSSE and DELINEATE for *Lionepha* beetles. The ProSSE results are more consistent with the most recently reviewed taxonomy of the species group (Maddison and Sproul 2020). Specifically, ProSSE strongly supports the taxonomy of Maddison and Sproul (2020) that recognizes lineages coalescentpop008 and coalescentpop009 as new species *L.kavanaughi* and lineages L_tuulukwa_CA_Trinity_Alps, L_tuulukwa_OR_Knowles_Creek, and L_tuulukwa_OR_Marys_Peak as new species *L. tuulukwa*. In contrast, DELINEATE infers slightly higher marginal likelihoods of these lineages to be conspecific to existing species other than the new species. In addition, ProSSE infers a slightly higher probability of lineage L_probata_UT_Shingle_Creek to be a new species than to be conspecific to *L. probata*, while DELINEATE strongly supports the lineage to be conspecific to *L. probata*. Maddison and Sproul (2020) assigned the lineage to *L. probata*, but they suggested that the lineage is molecularly distinctive enough to warrant further study.

Speciation in Australian Carlia Skinks

The dataset of Australian *Carlia* skinks gives equal support to the BDM model, where the rate of evolving intrinsic reproductive isolation is proportional to the mutation rate: $\lambda_c(r) = \beta e^r$, and the model with constant rate: $\lambda_c = \beta$ (log marginal likelihood: BDM model = −135.4424; constant rate =−135.5186), with Bayesian factor between the 2 models equal to 1.11. The marginal posterior distributions for all the parameters are very similar between the 2 models. Both suggest a faster shift from grass/littoral habitat to rock habitat than vice versa and a slightly faster rate of losing blue male color than gaining (Supplementary Fig. S5). The speciation completion rate for morphological species is on average 5 times faster than that for cryptic species (Supplementary Fig. S5).

ProSSE method using both models suggest very similar probabilistic delimitation of the 17 lineages with uncertain species identities. Table 2 lists the delimitation with the highest posterior probability for each species complex. For all species complex, except for *C. rhomboidalis*, the most supported delimitation assigns each lineage in the complex to a distinct new species. Considering all possible delimitations over all species complexes, ProSSE expects about 13 new species in *Carlia*.

For *C. gracilis* complex and *C. schmeltzii* complex, the probability for the mostly supported delimitation is relatively low, so we report here their second most supported delimitation. In *C. gracilis* complex, the second most supported delimitation has a posterior probability of 0.30. It assigns maningrida and ete lineages to 2 distinct new species and assigns the melville lineage to *C. gracilis*. In *C. schmeltzii* complex, the second mostly supported delimitation has a posterior probability of 0.40 and considers meq and seq lineages conspecific as a new species.

ProSSE delimitation is consistent with published preliminary evidence. Preliminary evidence supports the ete lineage in the *C. gracilis* complex as a new species and supports the ete lineage and gulf lineage in the *C. amax* complex as 2 different new species (Supplementary Table S1). ProSSE predicts the ete lineage of the *C. gracilis* complex to be a new species that is distinct from any other lineages in the complex with a probability of 0.71 (Table 2). Summing the probabilities of all delimitations for the *C. amax* complex that assign the ete lineage and the gulf lineage to be 2 different new species, ProSSE predicts the ete lineage and the gulf lineage to be 2 different new species with a probability of 0.95.

For island-only lineages, ProSSE infers high probabilities that lineages in English Company's Islands (eci) are new species that are different from any other lineages

TABLE 2. ProSSE delimitation results for *Carlia* skinks. In each species complex, one lineage is randomly chosen to represent the existing species, and ProSSE is applied to infer the species identities of the other lineages. For example, there are 2 lineages in the *C. vivax* complex, north lineage and south lineage. The south lineage is chosen to represent *C. vivax*, so we report delimitation results on the north lineage. For each species complex, the delimitation with the highest posterior probability is reported. Marginal posterior probability of being a distinct new species is calculated by summing the probabilities of all delimitations for the species complex that assign the lineage to a new species and this new species is distinct from any other lineages in the species complex.

Species complex	Lineage name	Delimitation with the highest posterior probability	Marginal posterior probability of being a distinct new species
C. vivax	North	new species (0.84)	0.84
C. dogare	*C.* cf. *dogare*	new species (0.95)	0.95
C. jarnoldae	Cy	new species (0.98)	0.98
C. gracilis	maningrida	all lineages distinct new species (0.36)	0.93
	Ete		0.71
	melville		0.48
C. amax	Eci	all lineages distinct new species (0.86)	0.98
	Gulf		0.95
	Ete		0.95
C. rufilatus	Eci	new species (0.98)	0.98
C. munda	melville	new species (0.69)	0.69
C. storri	cy_ng	all lineages distinct new species (0.79)	0.82
	tv_cy		0.90
C. schmeltzii	Meq	all lineages distinct new species (0.46)	0.50
	Seq		0.49
C. sexdentata	*C.* cf. *sexdentata*	*new species (0.97)*	0.97
C. rhomboidalis	North	*C. rhomboidalis* (0.54)	0.54

in the same complex. These include the eci lineage in *C. rufilatus* complex with a probability of 0.98 and the eci lineage in *C. amax* complex with a probability of 0.98 (Table 2). In contrast, ProSSE infers low probabilities that lineages on Melville island are new species. These include the melville lineage in *C. munda* complex with 0.69 and the melville lineage in *C. gracilis* complex with probability 0.48 (Table 2). This contrasting result for island lineages is consistent with the population genomic analysis that suggested ongoing migration from continental lineages to the Melvillie lineage in both *C. gracilis* complex and *C. munda* complex, but no migration from continental lineages to the eci lineages in both *C. amax* complex and *C. rufilatus* complex (Potter et al. 2018).

DISCUSSION

This study provides a new, efficient, and flexible phylogenetic approach for species delimitation and for modeling speciation processes. It extends the protracted speciation model ProSSE to separately model how fast lineages evolve to meet different delimitation criteria, including differences in qualitative traits, divergence in quantitative traits, and sufficient accumulation of genetic differences between lineages. By modeling these different delimitation criteria and by conditioning on the phylogeny and known species identities of some tip lineages, this extension of ProSSE not only allows us to infer the species identities of tip lineages that we are not certain about but also allows us to include species delimited by different criteria in the same phylogenetic comparative analysis. Applying extended ProSSE to simulated data, we show that the model has high accuracy in parameter estimation and the ProSSE method has high power in estimating species identities over a wide range of conditions. When applied to real data, the ProSSE method gives a prediction of species identities that is consistent with existing taxonomy and evidence for the completion of speciation. We also show how to apply the ProSSE method to test the BDM model that links mutation and speciation. This provides us with a phylogenetic framework to explicitly link microevolutionary and macroevolutionary models of speciation.

Speciation-Based Delimitation

Our ProSSE method is a type of speciation-based delimitation (first introduced by Sukumaran et al. 2021) that uses lineages with known species identities in a species group to infer how fast lineages become distinct species, in order to predict the species identities of the other lineages in the group. Unlike other delimitation methods, speciation-based delimitation does not force a fixed species concept or criteria for species delimitation but summarizes whatever criteria researchers have used to decide species boundaries into a probabilistic model. The advantages of doing so are that it does not require us to redelimit species boundaries using a single species concept before performing a comparative analysis and it keeps the species delimitation consistent before and after we add new species into a comparative analysis.

Despite these advantages, we emphasize that speciation-based delimitation should not be used to replace any existing delimitation criterion. Instead, speciation-based delimitation uses the results from the existing delimitation criteria. For example, Sukumaran et al. (2021) recommended speciation-based delimitation as a complementary method to coalescent-based species delimitation, because it distinguishes genetic structure within species (phylogenetic relationship

among intraspecific lineages) from that among species (phylogenetic relationship among interspecific lineages). However, we still need other delimitation criteria to tell us which lineages are intraspecific and interspecific before we use speciation-based delimitation either on lineages for which we cannot test any delimitation criterion or on lineages for which we have only tested some delimitation criteria and want to use test results from other lineages to inform its probable results for the remaining delimitation criteria. Essentially, the probability estimated from ProSSE for each possible species identity of a lineage can be thought of as the prior probability for delimitation before the delimitation criteria are tested on the lineage.

The fundamental assumption of speciation-based delimitation is that lineages in the same species group have similar rates of becoming distinct species. One extreme example of the assumption is used in DELINEATE, the method developed by Sukumaran et al. (2021), which assumes a single speciation completion rate across all lineages in a species group. This assumption can be unreasonable for some species groups. For example, genus *Cryptoblepharus* has both ecological species and cryptic species (Blom et al. 2016) and it often takes a much shorter time to become a new ecological species than a new cryptic species (Gavrilets 2003). In *Carlia* skinks, we estimated the speciation completion rate in morphological species on average 5 times faster than the speciation completion rate in cryptic species. If assuming morphological species and cryptic species share the same speciation completion rate in *Carlia*, we would overestimate the number of cryptic species.

To make the assumptions of the speciation-based delimitation model generally reasonable, extended ProSSE allows lineages to have different rates to evolve to meet different delimitation criteria. For example, if 2 lineages differ in some morphological characters, and 2 other nonallopatric lineages are morphologically similar, but have no hybrids in the contact zone, then researchers may have applied different species concepts and criteria for species delimitation to them. Extended ProSSE can estimate separate rates of a lineage evolving to meet the criterion of morphological difference and the criterion of no hybridization. Now, if we have another lineage with an unknown species identity, but we know how long it has been independently evolving from other lineages, then we can use this biological information of the lineage to calculate the probabilities that the lineage has evolved to meet one or both criteria. This promises a new approach to integrating different delimitation criteria under the same delimitation framework. Existing approaches of integrative delimitation test if a lineage passes some combinations of different delimitation criteria using different types of data from the lineage (e.g., Edwards and Knowles 2014; Solís-Lemus et al. 2015). Extended ProSSE differs from these approaches in that it predicts if a lineage passes any or all delimitation criteria using results of delimitation criteria from some other lineages in the same species group. More conveniently, each of these other lineages does not need to be tested by multiple criteria as along as different lineages have been tested by different criteria. This allows us to achieve integrative delimitation without the need to reassess existing species boundaries in a species group by multiple criteria.

We note that the current version of the ProSSE model uses the Poisson process to model when a lineage meets a particular delimitation criterion, which assumes all lineages have the same rate of evolving to meet this criterion. Consequently, the time duration between speciation initiation and speciation completion decided by this delimitation criterion follows an exponential distribution. The Poisson process is likely to be insufficient to fit the observed distribution of the time duration. For example, if delimitation uses the multispecies coalescent method, then with gene flow, a longer time for lineage divergence may be required for a lineage to pass the multispecies coalescent analysis. To account for this additional variation in the time duration, ProSSE algorithm allows us to easily generalize the Poisson process to any absorbing continuous-time Markov chain where the absorbing state is when speciation completes. A general absorbing continuous-time Markov chain has a continuous Phase-Type distribution that has much more flexible shapes than exponential distribution (Cox 1955), which allows us to fit almost any observed distribution of the time duration between speciation initiation and speciation completion.

In addition to this advantage, our ProSSE method outperforms DELINEATE in analyzing both simulated data and real species group even when the assumption of a single speciation completion rate holds. In particular, the amount of memory required by ProSSE method barely depends on the number of lineages with unknown species identities or on the number of species in the tree, while the amount of memory required by DELINEATE becomes impossible for high-performance computers for just about 50 species and 10 lineages with unknown species identities. The power of DELINEATE to correctly infer species identity also decreases faster than ProSSE method with an increasing proportion of lineages with unknown species identities.

Linking Microevolutionary and Macroevolutionary Models of Speciation

ProSSE is not only a species delimitation method but also a phylogenetic framework that allows us model speciation process more flexibly along phylogenetic trees and conditional on the known species identities of extant lineages. By explicitly modeling speciation initiation and speciation completion along lineage-level phylogenies, extended ProSSE models are able to separate the effect of speciation initiation rate and extinction rate (analogous to birth rate and death rate in birth–death model) on topologies and branch lengths from the effect of speciation completion rates on the species identities of tip lineages. An advantage of this is that we can

model the effects of different factors on speciation initiation rate and speciation completion rate separately. For example, mutation rate may affect speciation completion rate, but not speciation initiation rate.

Another advantage is that we always have closed-form solutions to model speciation process because tree likelihoods are calculated from linear coupled first-order systems: $D_R(t)$ and $D_I(t)$. This allows us to further extend ProSSE to model a more complicated speciation process. For example, we can explicitly model speciation completion events as a result of substitutions of BDM incompatible genes by introducing genotype-specific $D_R(t)$ and $D_I(t)$ (Hua, manuscript in preparation). For another example, we can explicitly model secondary contact by labeling speciation completion due to intrinsic reproductive isolation as the final stage of speciation completion, which is after the stage of being considered a distinct species by delimitation criteria other than intrinsic reproductive isolation. In other words, separating the *R* state into *R* state with, and *R* state without intrinsic reproductive isolation, only the latter of which can transit back to *I* state during secondary contact. The algorithm will be very similar to the algorithm for incompatible genes.

Using ProSSE, we demonstrate how to use phylogenetic information to test the BDM model of intrinsic reproductive isolation across all lineages in a species group, which are usually tested by comparing genomes only between 2 sister lineages. Under the BDM model, mathematical models predict that speciation completion rate is proportional to mutation rate. Using the *Carlia* skink dataset, we cannot reject the BDM model as the genetic basis of cryptic species, because the data gives equal evidence for the model where speciation completion rate is proportional to mutation rate and the model with constant speciation completion rate. This result is likely due to little variation in mutation rate among *Carlia* species (Supplementary Figure S1) compared to the entire Eugongylinae radiation (Ivan et al. 2021), because simulation results suggest high accuracy in estimating the coefficient between speciation completion rate and a continuous trait, such as mutation rate. This is further supported by the result that the posterior distributions of all the parameters (not just β) are very similar between the 2 models. As a result, the variation in speciation completion rate is relatively small compared to the inherent variance of time taken to complete speciation even under a constant speciation completion rate.

There are still 3 major confounding factors that could mask the true signal for the BDM model even if we apply the test to a species group with a large variation in mutation rate. One factor is that we usually don't have accurate information on the tip mutation rate in a species group. Even for *Carlia*, there is still a 50% variation in tip mutation rate that cannot be explained by temperature (Ivan et al. 2021). This problem can be solved by using ProSSE as the tree prior in molecular dating analysis and modeling speciation completion rate as a function of synonymous substitution rate along the tree. So instead of requiring prior information on tip mutation rate, we use sequence data to inform variation in synonymous substitution rates across the tree. This new molecular dating method is currently under development.

A related problem is that speciation-based delimitation requires a time tree and so its performance depends on how accurate the time tree is. For example, the time tree we use for *Carlia* skinks is inferred by a species tree method that uses a multispecies coalescent model, under which the estimated branch length may be confounded by mutation rate (Rannala and Yang 2003). This could mask the true signal for the BDM model because we use the branch lengths to estimate speciation completion rate. By co-estimating substitution rates, speciation completion rates, and branch lengths in the same analysis, we can avoid the circular reasoning of using substitutions to estimate branch lengths, using branch lengths to estimate speciation completion rates, and testing correlation between substitution rates and speciation completion rates.

The last confounding factor is that speciation completion rate is also proportional to the product of effective population size and selection coefficient if BDM genes are under direct selection (Gavrilets 2003), such that stronger and more efficient selection leads to faster speciation completion. There are 2 main types of direct selection on BDM genes: divergent selection under ecological speciation model and uniform selection under mutation-order speciation (Nosil and Flaxman 2011). Because lineages under divergent selection are likely diverged in morphology and in use of different habitats, to account for the effect of divergent selection in *Carlia*, we model speciation completion due to habitat transition and due to assortative mating separately from the other delimitation criteria. We also fit separate speciation completion rates for morphological species and cryptic species who do not show obvious morphological and ecological difference among each other. We show that, in *Carlia*, the speciation completion rate of morphological species is indeed faster than that of cryptic species. However, uniform selection is likely present among cryptic species in *Carlia*, especially because these cryptic species have similar level of genetic divergence to morphological species (Supplementary Fig. S1), which suggests that there is sufficient time for morphological divergence among cryptic species and yet they retain morphologically similar to each other possibly due to stabilizing selection (Struck et al. 2018). So the true signal for the BDM model can be masked by selection when variation in the product of effective population size and selection coefficient is larger than that of mutation rate among *Carlia* cryptic species. One possible solution to separate the effect of mutation and selection on speciation completion rate is to combine ProSSE, species tree method, and molecular dating in the same analysis, so that we can

test whether speciation completion rate is better predicted by mutation rate alone, by effective population size alone, or by both effective population size and mutation rate, assuming that the strength of uniform selection on BDM genes vary randomly across lineages of cryptic species.

CONCLUSION

This study develops ProSSE to generalize speciation-based delimitation, which mathematically models the use of different criteria to delimit species in a species group and uses the model to predict the species identities of lineages that have not been tested by these criteria. ProSSE not only provides us with a complementary method to existing species delimitation criteria, by estimating how fast a lineage evolves to meet different delimitation criteria but also gives us a flexible way to model the speciation process along a phylogenetic tree, which allows us to test speciation theories under a phylogenetic framework. Future work will combine ProSSE with substitution models to explicitly link molecular changes and speciation.

SUPPLEMENTARY MATERIAL

Supplemental information is available at https://github.com/huaxia1985/ProSSE/blob/main/SI_Hua_Moritz_2025.pdf.

ACKNOWLEDGEMENTS

We thank Lindell Bromham for valuable comments to the manuscript. We thank Stephen Zozaya for advice on *Carlia* taxonomy and Jason Bragg and Sally Potter for the updated *Carlia* species tree.

AUTHOR CONTRIBUTIONS

Both authors contributed critically to the drafts and gave final approval for publication. XH conceived the project, developed the method, and led the writing of the manuscript. CM curated the *Carlia* dataset and participated in the writing.

Conflict of Interest statement

None declared.

FUNDING

This work was supported by the Australian Research Council (DE190100491).

DATA AVAILABILITY

The source code of extended ProSSE is maintained at github.com/huaxia1985/ProSSE. The data and the R code for case studies are available at https://github.com/huaxia1985/ProSSE/tree/main/case%20study.

REFFERENCES

Afonso Silva A.C., Santos N., Ogilvie H.A., Moritz C. 2017. Validation and description of two new north-western Australian Rainbow skinks with multispecies coalescent methods and morphology. PeerJ. 5:e3724.

Anderson E., Thompson E. 2002. A model-based method for identifying species hybrids by using multilocus genetic data. Genetics. 160:1217–1229.

Barton N, Gale K. 1993 Genetic analysis of hybrid zones. In: Harrison R., editor. Hybrid zones and the evolutionary process. Oxford, UK: Oxford University Press. pp. 13–45.

Bateson W. 1909. Heredity and variation in modern lights. In: Seward A.C., editor. Darwin and modern science. Cambridge: Cambridge University Press. p. 85–101.

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

Bragg J.G., Potter S., Afonso Silva A.C., Blom M.P.K., Hoskin C.J., Ogilvie H.A., Mortiz C. 2024. Unrecognized lineages transform our understanding of diversification in a clade of lizards. EcoEvoRxiv. doi:10.32942/X2J043

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.

Cadena C.D., Zapata F., Jiménez I. 2018. Issues and perspectives in species delimitation using phenotypic data: Atlantean evolution in Darwin's finches. Syst. Biol. 67:181–194.

Cox D.R. 1955. A use of complex probabilities in the theory of stochastic processes. Math. Proc. Camb. Philos. Soc. 51:313–319.

De Queiroz K. 1998. The general lineage concept of species, species criteria, and the process of speciation: A conceptual unification and terminological recommendations. In: Howard D.J. and Berlocher S.H., editor. Endless Forms. Species and Speciation. Oxford, UK: Oxford University Press, pp. 57–75.

De Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.

Dobzhanksy, T. 1937. Genetics and the origin of species. NY: Columbia University Press.

Dolman G. 2008. Evidence for differential assortative female preference in association with refugial isolation of rainbow skins in Australia's tropical rainforest. PLoS One 3:e3499.

Edwards D.L., Knowles L.L. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy?. Proc. R. Soc. Lond. B Biol. Sci. 281:20132765.

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.

Fišer C., Robinson C.T., Malard F. 2018. Cryptic species as a window into the paradigm shift of the species concept. Mol. Ecol. 27:613–635.

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.

Gavrilets S. 2003. Models of speciation: what we have learned in 40 years. Evolution 57:2197–2215.

Hua X., Herdha T., Burden C.J. 2022. Protracted speciation under the state-dependent speciation and extinction approach. Syst. Biol. 71:1362–1377.

Ivan J., Mortiz C., Potter S., Bragg J., Turakulov R., Hua X. 2021. Temperature predicts the rate of molecular evolution in Australian Eugongylinae skinks. Evolution 76:252–261.

Jackson N., Carstens B., Morales A., O'Meara B.C. 2017. Species delimitation with gene flow. Syst. Biol. 66:799–812.
Jones G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74:447–467.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: 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.
Maddison D., Sproul J.S. 2020. Species delimitation, classical taxonomy and genome skimming: a review of the ground beetle genus *Lionepha* (Coleoptera: Carabidae). Zool. J. Linn. Soc. 189:1313–1358.
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.
Maliet O., Hartig F., Morlon H. 2019. A model with many small shifts for estimating species-specific diversification rates. Nat. Ecol. Evol. 3:1086–1092.
Muller H.J. 1942. Isolating mechanisms,evolution and temperature. Biol. Symp. 6:71–125.
Nabholz B., Lanfear R., Fuchs J. 2016. Body mass-corrected molecular rate for bird mitochondrial DNA. Mol. Ecol. 25:4438–4449.
Nosil P., Flaxman S.M. 2011. Conditions for mutation-order speciation. Proc. R. Soc. Lond. B, Biol. Sci. 278:399–407.
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.
Pagel M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. B Biol. Sci. 255:37–45.
Potter S., Bragg J.G., Peter B.M., Bi K., Moritz C. 2016. Phylogenomics at the tips: Inferring lineages and their demographic history in a tropical lizard, *Carlia amax*. Mol. Ecol. 25:1367–1380.
Potter S., Xue A.T., Bragg J.G., Rosauer D.F., Roycroft E.J., Moritz C. 2018. Pleistocene climatic changes drive diversification across a tropical savanna. Mol. Ecol. 27:520–532.
Pritchard J., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 145:945–959.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Reichl J. 2020. Estimating marginal likelihoods from the posterior draws through a geometric identity. Monte Carlo Methods Appl. 26:205–221.
Simpson A.G., Wagner P.J., Wing S.L., Fenster C.B. 2018. Binary-state speciation and extinction method is conditionally robust to realistic violations of its assumptions. BMC Evol. Biol. 18:69.
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.
Solís-Lemus C., Knowles L.L., Ané C. 2014. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507.
Struck T.H., Feder J.L., Bendiksby M., Birkeland S., Cerca J., Gusarov V.I., Kistenich S., Larsson K., Liow L.H., Nowak M.D., Stedje B., Bachmann L., Dimitrov D. 2018. Finding evolutionary processes hidden in cryptic species. Trends Ecol. Evol. 33:153–163.
Sukumaran J., Holder M.T., Knowles L.L. 2021. Incorporating the speciation process into species delimitation. PLoS Comput. Biol. 17:e1008924.
Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
Van Valen L. 1976. Ecological species, multispecies, and oaks. Taxon. 25:233–239.
Wiens J.J., Servedio M.R. 2010. Species delimitation in systematics: inferring diagnostic differences between species. Proc. R. Soc. Lond. B. 267:631–636.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA 107:9264–9269.
Zhang C., Zheng D., Zhu T., Yang Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. Syst. Biol. 60:747–761.