Supplementary Materials for:

A phylogenetic approach to delimitate species in a probabilistic way

Author: Xia Hua, Craig Moritz

Online Appendix A.

Online Appendix B.

Online Appendix C.

Supplementary Tables:

Supplementary Figures:

Appendix A.

We briefly describe the ProSSE algorithm, for the convenience of explaining our extension of the model in Appendix B and C. The basic idea of ProSSE is to calculate two probabilities: $D_R(t)$ and $D_I(t)$, along each edge of a tree from the present to the root, which are the probabilities that an edge in state R and in state I at time t leads to extant descendants as observed in the tree. State R means the edge represents a species at time t. For example, at the present when $t = 0$, any one of the tips that belong to the orange species in Figure 1b can represent the orange species. Hua et al. (2022) showed that which tip we choose to represent the species does not affect the likelihood. It is actually not necessary to assign a state to any edge that connects lineages of the same species, because $D_R(t)$ and $D_I(t)$ can be calculated by the same differential equation:

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

with the initial condition $D(0) = 1$, where $E(t) = \frac{\mu - \mu e^{(b-\mu)t}}{\mu - be^{(b-\mu)t}}$, λ is the speciation completion rate, b is the speciation initiation rate, and μ is extinction rate. At an internal node connecting two edges that belong to the same species, for example, at the descendant node of edge u at time t connecting two edges v and w (Figure 1b),

$$D_u(t) = bD_v(t)D_w(t), \quad (2)$$

which gives the boundary condition for edge u.

Along edges connecting different species, $D_R(t)$ and $D_I(t)$ are derived from

$$D_R'(t) = -(b + \mu)D_R(t) + 2bD_R(t)E(t) \quad (3)$$

$$D_I'(t) = -(\lambda + b + \mu)D_I(t) + 2bD_I(t)E(t) + \lambda D_R(t). \quad (4)$$

At an internal node connecting two edges that belong to different species, for example, at the descendant node of edge x at time t connecting edges y and z (Figure 1b),

$$D_{R,x}(t) = bD_{R,y}(t)D_{I,z}(t) + bD_{I,y}(t)D_{R,z}(t) \quad (5)$$

$$D_{I,x}(t) = bD_{I,y}(t)D_{I,z}(t) \quad (6)$$

These give the boundary conditions for edge x. In this particular example, $D_{R,y}(t) = 0$, because the edge y leads to the blue species, so it cannot represent the orange species at time t (Figure 1b). Then, the calculation continues along edge x until reaching the root the tree.

Appendix B

To account for co-occurrence between transition in trait states (or in habitat types) and speciation completion event, we introduce parameter q_{ij} for the rate of transition from state i to state j of the trait (or habitat) and parameter p_{ij} for the probability that the transition from state i to state j co-occurs with speciation completion event. To distinguish speciation due to gradual accumulation of genetic difference from that of morphological difference, we introduce speciation completion rate λ_c for cryptic species and λ_m for morphological species, and denote state *I* and state *R* with a subscript *c* for edges connecting cryptic species and with a subscript *m* for edges connecting morphological species.

For cryptic species, state transition is modelled only for habitat shifts that do not cause obvious morphological changes in cryptic species. For in total K states, $D_{R_c}(t)$ and $D_{I_c}(t)$ become row vectors $\vec{D}_{R_c}(t)$ and $\vec{D}_{I_c}(t)$, with each element corresponding to a state. Similar to equation 1, along edges connecting lineages of the same cryptic species, $\vec{D}_{R_c}(t)$ and $\vec{D}_{I_c}(t)$ can be calculated by the same differential equation for $\vec{D}_c(t)$, where for state i:

$$\begin{aligned} D_{c,i}'(t) = {} & -(\lambda_c + b + \mu)D_{c,i}(t) + 2bD_{c,i}(t)E(t) \\ & - \textstyle\sum_{j\neq i}^{K} q_{ij}D_{c,i}(t) && \text{No state transition} \\ & + \textstyle\sum_{j\neq i}^{K} q_{ij}(1-p_{ij})D_{c,j}(t) && \text{Transition does not co-occur with speciation completion. e.g., event 2 in Figure 1c} \end{aligned} \quad (7)$$

, because speciation completion event cannot happen along edges connecting lineages of the same species. The solution of equation 7 over h amount of time since time t is

$$\vec{D}_c(t+h) = A(h,t)e^{-\lambda_c h}e^{\boldsymbol{Q}h}\vec{D}_c(t) \tag{8}$$

, where $A(h,t) = \frac{(b-\mu)^2 e^{(b-\mu)h}}{\left[be^{(b-\mu)h}-\mu+bE(t)(1-e^{(b-\mu)h})\right]^2}$ and $\boldsymbol{Q}$ is a matrix with the i^{th} diagonal $-\sum_{j\neq i}^{K} q_{ij}$ and the (i,j) off-diagonal $q_{ij}(1-p_{ij})$.

Along edges connecting different cryptic species within the same species complex, $\vec{D}_{R_c}(t)$ is derived from the following equation for state i:

$$\begin{aligned} D'_{R_c,i}(t) = \; & -(b+\mu)D_{R_c,i}(t) + 2bD_{R_c,i}(t)E(t) \\ & -\sum_{j\neq i}^{K} q_{ij}D_{R_c,i}(t) && \text{No state transition} \\ & +\sum_{j\neq i}^{K} q_{ij}(1-p_{ij})D_{R_c,j}(t) && \text{Transition does not co-occur with speciation completion} \end{aligned} \tag{9}$$

, because if habitat shifts do not cause obvious morphological changes in cryptic species, then habitat shifts do not co-occur with speciation completion events. The solution of equation 9 for all trait states over h amount of time since t is

$$\vec{D}_{R_c}(t+h) = A(h,t)e^{\boldsymbol{Q}h}\vec{D}_{R_c}(t). \tag{10}$$

$\vec{D}_{I_c}(t)$ is derived from the following equation for state i:

$$\begin{aligned} D'_{I_c,i}(t) = \; & -(\lambda_c+b+\mu)D_{I_c,i}(t) + 2bD_{I_c,i}(t)E(t) \\ & +\lambda_c D_{R_c,i}(t) \\ & -\sum_{j\neq i}^{K} q_{ij}D_{I_c,i}(t) && \text{No state transition} \\ & +\sum_{j\neq i}^{K} q_{ij}\left(1-p_{ij}\right)D_{I_c,j}(t) && \text{Transition does not co-occur with speciation completion} \end{aligned} \tag{11}$$

The solution of equation 11 for all habitat types over h amount of time since t is

$$\vec{D}_{I_c}(t+h) = A(h,t)e^{-\lambda_c h}e^{\boldsymbol{Q}h}\vec{D}_{I_c}(t) - A(h,t)[e^{-\lambda_c h}-1]\vec{D}_{R_c}(t). \tag{12}$$

At the MRCA of a species complex, we set $\vec{D}_{R_m}(t) = \vec{D}_{R_c}(t)$ and $\vec{D}_{I_m}(t) = \vec{D}_{I_c}(t)$, because the MRCA of different species complexes must be morphologically distinct, and so the edge leading to the MRCA of a species complex is considered a morphological species.

For morphological species, states i and j are for both discrete morphological traits and habitat types. Along edges connecting lineages of the same morphological species, $\vec{D}_{R_m}(t)$ and $\vec{D}_{I_m}(t)$ can be calculated by the same differential equation for $\vec{D}_m(t)$. The equation is the same as equation 7, but with $\vec{D}_c(t)$ and λ_c replaced by $\vec{D}_m(t)$ and λ_m. Along edges connecting lineages of different morphological species, $\vec{D}_{R\,m}(t)$ is derived from, for state i:

$$\begin{aligned} D'_{R_m,i}(t) = {} & -(b+\mu)D_{R_m,i}(t) + 2bD_{R_m,i}(t)E(t) \\ & -\textstyle\sum_{j\neq i}^{K} q_{ij} D_{R_m,i}(t) && \text{No state transition} \\ & +\textstyle\sum_{j\neq i}^{K} q_{ij} D_{R_m,j}(t) && \text{State transition} \end{aligned} \quad (13)$$

, because edge already in R state stays R state, no matter if state transitions co-occur with speciation completion events. The solution of equation 13 for all trait states over h amount of time since t is:

$$\vec{D}_{R_m}(t+h) = A(h,t)e^{\boldsymbol{Q}_R h}\vec{D}_{R_m}(t) \quad (14)$$

, where $\boldsymbol{Q}_R$ is a matrix with the i^{th} diagonal $-\sum_{j\neq i}^{K} q_{ij}$ and the (i,j) off-diagonal q_{ij}.

$\vec{D}_{I_m}(t)$ is derived from, for state i:

$$\begin{aligned} D'_{I_m,i}(t) = {} & -(\lambda_m + b + \mu)D_{I_m,i}(t) + 2bD_{I_m,i}(t)E(t) \\ & +\lambda_m D_{R_m,i}(t) \\ & -\textstyle\sum_{j\neq i}^{K} q_{ij} D_{I_m,i}(t) && \text{No state transition} \\ & +\textstyle\sum_{j\neq i}^{K} q_{ij}\left(1-p_{ij}\right) D_{I_m,j}(t) && \text{Transition does not co-occur with speciation completion} \end{aligned} \quad (15)$$

$+\sum_{j\neq i}^{K} q_{ij}p_{ij}D_{R_m,j}(t)$ Transition co-occurs with speciation completion

, because edge in I_m state becomes R_m state after speciation completes, so we need to consider state transitions that co-occur with speciation completion events separately from those that do not co-occur with speciation completion events. The solution of equation 15 for all trait states over h amount of time since t is

$$\vec{D}_{I_m}(t+h) = \vec{D}_{R_m}(t+h) + A(h,t)e^{-\lambda_m h}e^{\boldsymbol{Q}h}\left[\vec{D}_{I_m}(t) - \vec{D}_{R_m}(t)\right]. \tag{16}$$

Appendix C

To allow speciation completion rate in cryptic species vary with mutation rate, we model mutation rate on log-scale r to evolve under a diffusion process along the tree with drift term $\phi(r,t)$ and diffusion term $\sigma^2(r,t)$, and make λ_c a function of r, specifically, $\lambda_c(r) = \beta e^r$, $\beta > 0$. As a result, equation 7 becomes

$$\begin{aligned}\frac{\partial D_{c,i}(r,t)}{\partial t} = &-\left(\lambda_c(r) + b + \mu + \sum_{j\neq i}^{K} q_{ij}\right)D_{c,i}(r,t) + 2bD_{c,i}(r,t)E(t) \\ &+ \sum_{j\neq i}^{K} q_{ij}\left(1-p_{ij}\right)D_{c,j}(r,t) + \phi(r,t)\frac{\partial D_{c,i}(r,t)}{\partial r} \\ &+ \frac{\sigma^2(r,t)}{2}\frac{\partial^2 D_{c,i}(r,t)}{\partial r^2}.\end{aligned} \tag{15}$$

Equation 9 becomes

$$\frac{\partial D_{R_c,i}(r,t)}{\partial t} = -\left(b+\mu+\sum_{j\neq i}^{K} q_{ij}\right) D_{R_c,i}(r,t) + 2bD_{R_c,i}(r,t)E(t)$$
$$+\sum_{j\neq i}^{K} q_{ij}(1-p_{ij}) D_{R_c,j}(r,t) + \phi(r,t)\frac{\partial D_{R_c,i}(r,t)}{\partial r} \quad (16)$$
$$+\frac{\sigma^2(r,t)}{2}\frac{\partial^2 D_{R_c,i}(r,t)}{\partial r^2}.$$

Equation 11 becomes

$$\frac{\partial D_{I_c,i}(r,t)}{\partial t} = -\left(\lambda_c(r)+b+\mu+\sum_{j\neq i}^{K} q_{ij}\right) D_{I_c,i}(r,t) + 2bD_{I_c,i}(r,t)E(t)$$
$$+\sum_{j\neq i}^{K} q_{ij}\left(1-p_{ij}\right) D_{I_c,j}(r,t) + \lambda(r) D_{R_c,i}(t) + \phi(r,t)\frac{\partial D_{I_c,i}(r,t)}{\partial r} \quad (17)$$
$$+\frac{\sigma^2(r,t)}{2}\frac{\partial^2 D_{I_c,i}(r,t)}{\partial r^2}.$$

Following FitzJohn (2009), these equations are solved efficiently under Brownian motion by discretising both the quantitative state axis r and the time axis t. In each time step of size h and for each qualitative trait state i, we use equations 8, 10, 12 to update $D_{c,i}(r,t)$, $D_{R_c,i}(r,t)$, and $D_{I_c,i}(r,t)$ over t and then use fast Fourier transformation (FFT) to perform convolutions over r. We solve equation 15 for $D_{c,i}(r,t)$ along edges connecting lineages of the same cryptic species and solve equation 16 for $D_{R_c,i}(r,t)$ and equation 17 for $D_{I_c,i}(r,t)$ along edges that connect different cryptic species of the same species complex. At the MRCA of a species complex, we set $D_{R_m,i}(r,t) = D_{R_c,i}(r,t)$ and $D_{I_m,i}(r,t) = D_{I_c,i}(r,t)$. Along an edge connecting lineages of morphological species, $D_{m,i}(r,t)$, $D_{R_m,i}(r,t)$, and $D_{I_m,i}(r,t)$ at the ancestral node of the edge are solved by multiplying the solutions for $D_{m,i}(t)$, $D_{R_m,i}(t)$, and $D_{I_m,i}(t)$ by the solution for Brownian motion starting at r, over time t, and integrated over the value at the descendant node of the edge using FFT. When a morphological species

is nested within a species complex, at the ancestral node of the edge that leads to the morphological species, we set $D_{R_c,i}(r,t) = D_{R_m,i}(r,t)$ and $D_{I_c,i}(r,t) = D_{I_m,i}(r,t)$. The $D_{R_m,i}(r,t)$ at the root gives the probability of the tree that connects all the extant lineages of certain species identities given that the root is in state i and have value r. Following the argument in FitzJohn et al. (2009), the final probability of the tree is calculated as $\sum_{i=1}^{K} D_{R,i}(t) \frac{D_{R,i}(t)}{\sum_{i=1}^{K} D_{R,i}(t)}$, where $D_{R,i}(t)$ is the integral of $D_{R,i}(r,t)$ over axis r.

Reference:

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

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.

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

Table S1. Published evidence on gene flow among non-allopatric lineages within species complex of Australian *Carlia* skinks. In the type of evidence, proportion of hybrids is inferred from population admixture analysis; the demographic history is inferred from the coalescent simulation that fit best to the observed site frequency spectrum out of various gene flow histories; Multispecies coalescent analysis (MSC) uses at least two methods, including BPP (Yang and Rannala 2014), BFD (Leaché et al. 2014), and STACEY (JONES 2017).

Species complex	Lineage pairs	Gene flow	Type of evidence	Reference
C. gracilis	wte vs. kim ete vs. kim ete vs. wte	Present Absent Absent	Proportion of hybrids; Demographic history	Potter et al. 2018
C. amax	ete vs. wte ete vs. gulf wte vs. gulf	Absent Absent Absent	D-statistic; MSC; Proportion of hybrids; Demographic history;	Potter et al. 2016; 2018 Fenker et al. 2021
C. triacantha	*C. triacantha* vs. *C. isostriacantha*	Absent	MSC; Morphological analysis	Afonso Silva et al. 2017
C. johnstonei	*C. johnstonei* vs. *C. insularis*	Absent	MSC; Morphological analysis	Afonso Silva et al. 2017
C. rufilatus	te vs. kim	Present	Proportion of hybrids; Demographic history	Potter et al. 2018
C. munda	ete vs. broad	Present	Proportion of hybrids; Demographic history	Potter et al. 2018
C. rubrigularis	*C. rubrigularis* vs. *C. crypta*	Absent	D-statistic; MSC; cline widths; LD; Morphological analysis; Proportion of hybrids	Singhal et al .2018

Reference:

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. Peer J. 5:e3724.

Fenker J., Tedeschi L.G., Melville J., Moritz C. 2021. Predictors of phylogeographic structure among codistributed taxa across the complex Australian monsoonal tropics. Mol. Ecol. 30:4276–4291.

Jones G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74:447–467.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR. 2014. Species delimitation using genome-wide SNP data. Systematic Biology 63:534–542.

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.

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.

Yang Z., Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. Mol. Biol. Evol. 31:3125–3135.

Figure S1. The time tree and the tip information of Australian *Carlia* skinks. The time tree is the maximum clade credibility tree of Australian *Carlia* skinks at lineage level. Coloured boxes at each tip show its male throat color, habitat, and the average temperature (on natural log scale) over its geographic distribution. Lineages of unknown species identities are marked with asterisk. A lineage in a box could be conspecific to any other lineage in the same box. Lineages in the same box usually belong to the same species complex. The only exception is *C. cf sexdentata* te, because the two lineages of *C. sexdentata* are not monophyletic, so their closest relatives are also included in the same box, indicating that *C. cf sexdentata* te could be conspecific with *C. sexdentata, or C. longipes, or C. quinquecarinata.*

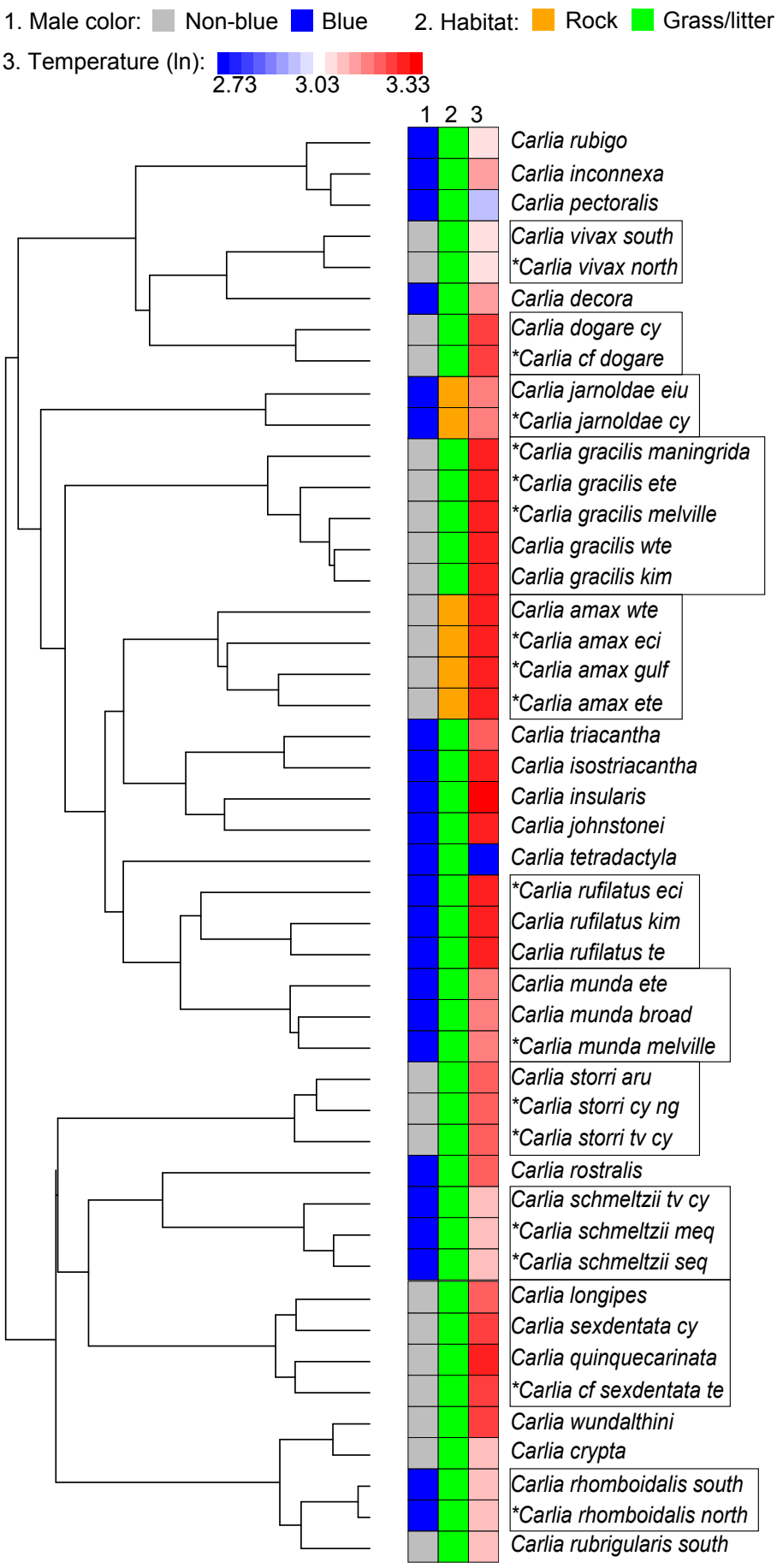
1. Male color: Non-blue Blue
2. Habitat: Rock Grass/litter
3. Temperature (ln):
2.73
3.03
3.33
1 2 3
Carlia rubigo
Carlia inconnexa
Carlia pectoralis
Carlia vivax south
*Carlia vivax north
Carlia decora
Carlia dogare cy
*Carlia cf dogare
Carlia jarnoldae eiu
*Carlia jarnoldae cy
*Carlia gracilis maningrida
*Carlia gracilis ete
*Carlia gracilis melville
Carlia gracilis wte
Carlia gracilis kim
Carlia amax wte
*Carlia amax eci
*Carlia amax gulf
*Carlia amax ete
Carlia triacantha
Carlia isostriacantha
Carlia insularis
Carlia johnstonei
Carlia tetradactyla
*Carlia rufilatus eci
Carlia rufilatus kim
Carlia rufilatus te
Carlia munda ete
Carlia munda broad
*Carlia munda melville
Carlia storri aru
*Carlia storri cy ng
*Carlia storri tv cy
Carlia rostralis
Carlia schmeltzii tv cy
*Carlia schmeltzii meq
*Carlia schmeltzii seq
Carlia longipes
Carlia sexdentata cy
Carlia quinquecarinata
*Carlia cf sexdentata te
Carlia wundalthini
Carlia crypta
Carlia rhomboidalis south
*Carlia rhomboidalis north
Carlia rubrigularis south

Figure S2. Error in ML estimates of the parameters related to the rate of trait evolution for each simulation scenario. These parameters are a) σ^2, the diffusion coefficient in the Brownian Motion of the continuous trait; b) q_{12} and c) q_{21}, the rate of state transition in the discrete trait, 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 five coloured boxplots represent the five 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.

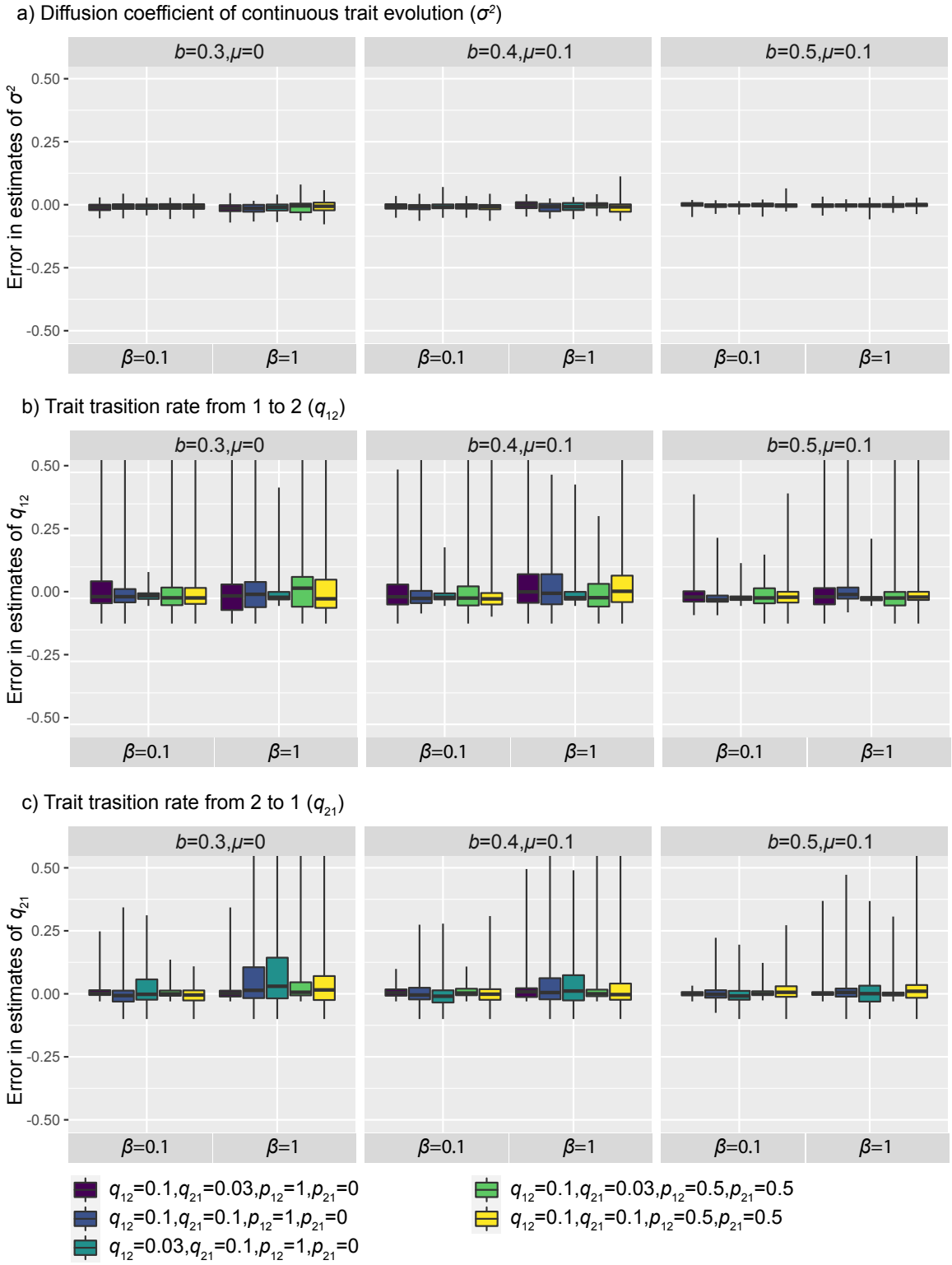
a) Diffusion coefficient of continuous trait evolution (σ^2)
b) Trait trasition rate from 1 to 2 (q_{12})
c) Trait trasition rate from 2 to 1 (q_{21})
b=0.3,μ=0
b=0.4,μ=0.1
b=0.5,μ=0.1
β=0.1
β=1
Error in estimates of σ²
Error in estimates of q12
Error in estimates of q21
0.50
0.25
0.00
-0.25
-0.50
q12=0.1,q21=0.03,p12=1,p21=0
q12=0.1,q21=0.1,p12=1,p21=0
q12=0.03,q21=0.1,p12=1,p21=0
q12=0.1,q21=0.03,p12=0.5,p21=0.5
q12=0.1,q21=0.1,p12=0.5,p21=0.5

Figure S3. The trend in the mean and the range of errors with tree size when speciation completion rate is low. Estimators are median-unbiased and consistent if both the median and the range of errors diminish to close to zero with tree size. Each plot corresponds to each parameter. Each dot is an estimate of the parameter from each of the 100 trees simulated under a simulation scenario labelled by the color.

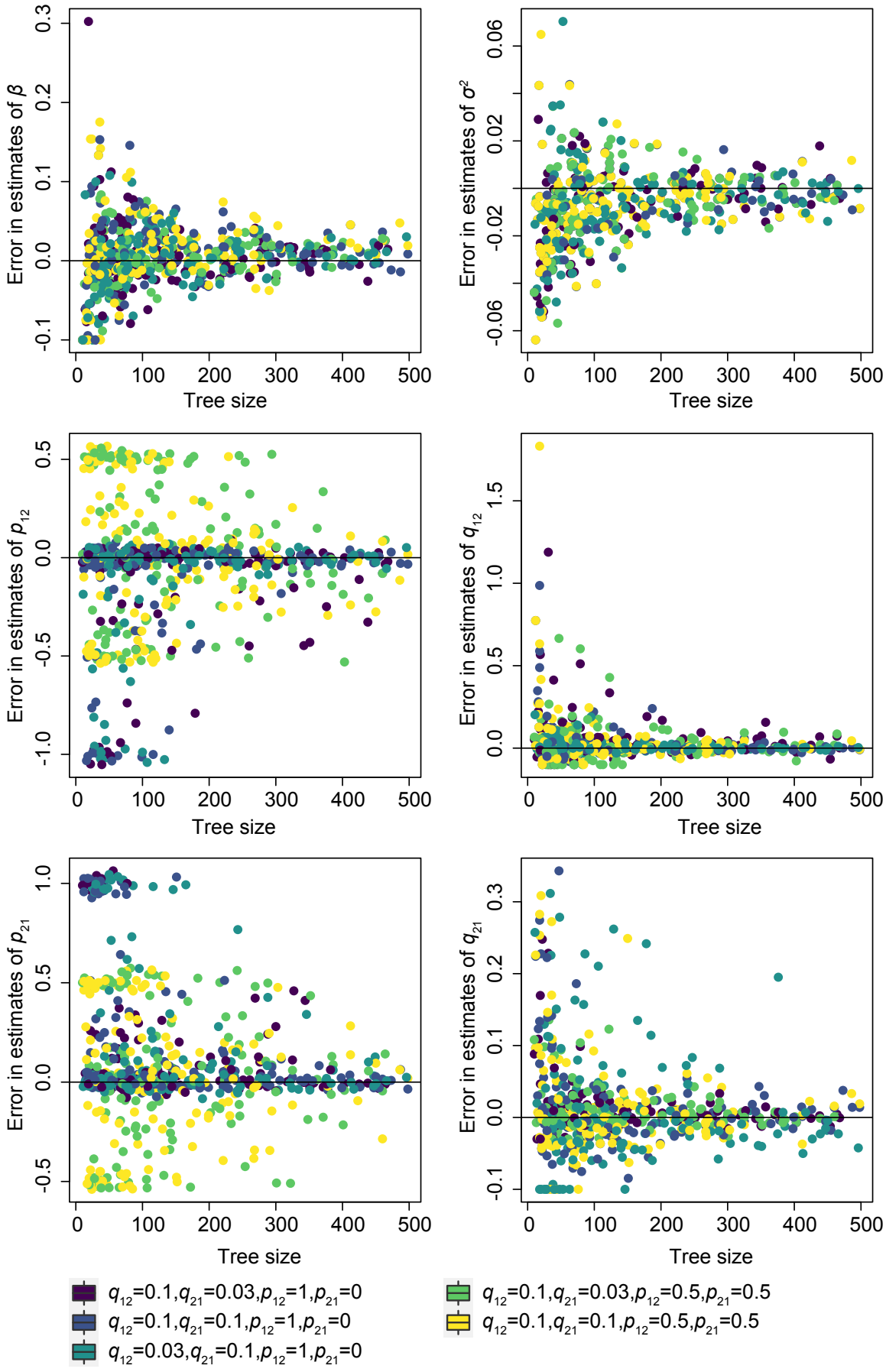
Error in estimates of β
Tree size
Error in estimates of σ²
Tree size
Error in estimates of p12
Tree size
Error in estimates of q12
Tree size
Error in estimates of p21
Tree size
Error in estimates of q21
Tree size
q12=0.1,q21=0.03,p12=1,p21=0
q12=0.1,q21=0.1,p12=1,p21=0
q12=0.03,q21=0.1,p12=1,p21=0
q12=0.1,q21=0.03,p12=0.5,p21=0.5
q12=0.1,q21=0.1,p12=0.5,p21=0.5

Figure S4. The trend in the mean and the range of errors with tree size when speciation completion rate is high. Estimators are median-unbiased and consistent if both the median and the range of errors diminish to close to zero with tree size. Each plot corresponds to each parameter. Each dot is an estimate of the parameter from each of the 100 trees simulated under a simulation scenario labelled by the color.

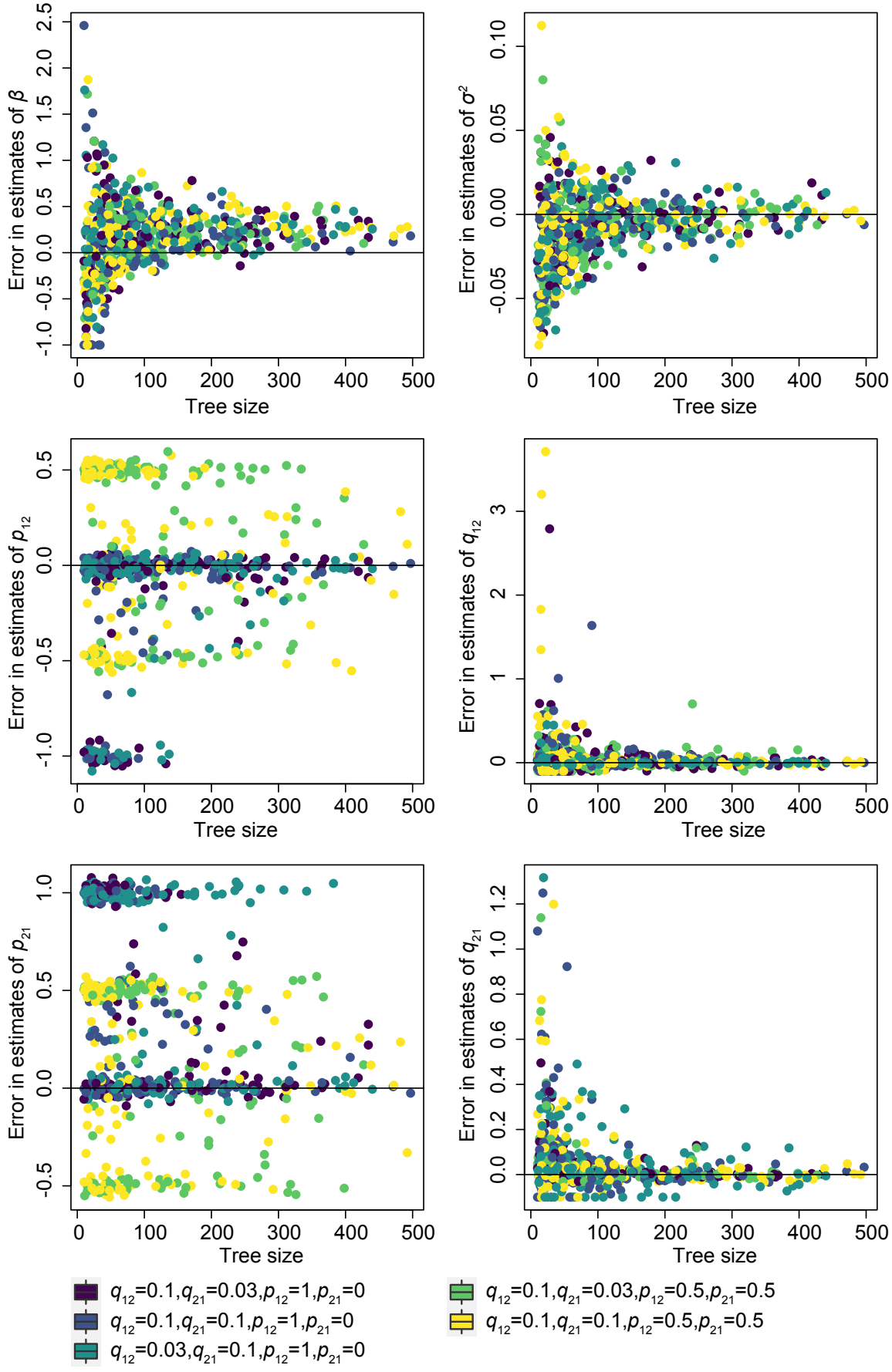
Error in estimates of β
Error in estimates of σ²
Error in estimates of p_{12}
Error in estimates of q_{12}
Error in estimates of p_{21}
Error in estimates of q_{21}
Tree size
q_{12}=0.1,q_{21}=0.03,p_{12}=1,p_{21}=0
q_{12}=0.1,q_{21}=0.1,p_{12}=1,p_{21}=0
q_{12}=0.03,q_{21}=0.1,p_{12}=1,p_{21}=0
q_{12}=0.1,q_{21}=0.03,p_{12}=0.5,p_{21}=0.5
q_{12}=0.1,q_{21}=0.1,p_{12}=0.5,p_{21}=0.5

Figure S5. Fitted density curves to the posterior samples of parameters under the BDM model and under the model with constant speciation completion rate for cryptic species. A) plots the difference in transition rates between blue male throat color and non-blue male throat color, where shaded area is the parameter space where transition from blue to non-blue is faster than that from non-blue to blue, with the numerical value showing its marginal posterior probability. B) plots the difference in transition rates between rock habitat and grass/littoral habitat, where shaded area is the parameter space where transition from grass/littoral habitat to rock habitat is faster than transition from rock habitat to grass/littoral habitat, with the numerical value showing its marginal posterior probability. C) plots the ratio between speciation completion rate in morphological species (λ_m) and in cryptic species (β), where shaded area and the value show the marginal posterior probability that speciation completion rate is higher in morphological species than in cryptic species.