

NEW PHYLOGENETIC COMPARATIVE APPROACHES FOR STUDYING VARIATION IN  
RATES OF CONTINUOUS TRAIT EVOLUTION

By

Bruce Stagg Martin

A DISSERTATION

Submitted to  
Michigan State University  
in partial fulfillment of the requirements  
for the degree of

Plant Biology—Doctor of Philosophy  
Ecology, Evolutionary Biology and Behavior—Dual Major

2024

## ABSTRACT

Rates of phenotypic evolution vary tremendously across the tree of life, generating vast disparities in phenotypic diversity across space, time, and taxa. Unfortunately, elucidating the factors driving such “rate heterogeneity” remains challenging due to various methodological limitations. In particular, most available methods for inferring variation in rates of continuous trait evolution assume rates are either influenced by only a few factors (i.e., variables hypothesized to affect rates) or change infrequently over the course of a clade’s history. However, rates of phenotypic evolution are likely affected by a dynamic, tangled web of countless environmental, life history, and genetic factors. By ignoring “residual” rate variation stemming from unobserved factors and assuming relatively simple rate variation patterns, available methods for modeling continuous trait evolution tend to underfit empirical data and mislead hypothesis testing by inflating support for complex models assuming spurious factor-rate associations. Here, to address these challenges, I develop, test, and apply new phylogenetic comparative methods capable of accurately inferring variation in rates of continuous trait evolution and robustly testing for factor-rate associations.

In chapter 1, I develop a novel continuous trait evolution model whereby rates constantly and incrementally change over time and across lineages, resulting in continuous, stochastic rate variation across a clade with closely-related lineages more likely to exhibit similar rates. I implement a Bayesian approach for fitting this model to empirical data in an R package, *evorates* (<https://github.com/bstaggmartin/evorates/>), along with comprehensive tools for analyzing and visualizing model results. Through simulation, I demonstrate that this method yields accurate inferences and can more reliably detect general decreases/increases in rates over time (i.e., “early/late bursts” of trait evolution) than previous methods by accounting for residual rate variation around overall time-dependent trends. Additionally, I use *evorates* to show that rates of body size evolution among whales and dolphins have generally declined over time yet exhibit substantial residual variation, with oceanic dolphins and beaked whales exhibiting anomalously fast and slow rates, respectively.

In chapter 2, I generalize stochastic character mapping or “simmapping”-based pipelines for

inferring relationships between rates of continuous trait evolution and discrete factors (e.g., habitat, diet) to also accommodate continuous factors (e.g., temperature, generation time). Simmapping is a popular method for imputing the uncertain evolutionary history of a trait (or factor) by sampling probable histories along a phylogeny under a given trait evolution model. However, available simmapping implementations only work with discrete variables. Accordingly, I develop a new R package, *contsimmap* (<https://github.com/bstaggmartin/contsimmap/>), which implements both a scalable algorithm for simmapping continuous variables and methods for inferring relationships between simmapped continuous factors and continuous trait evolution dynamics. I go on to verify the accuracy and robustness of this new pipeline in estimating factor-rate relationships via an extensive simulation study, even devising a pragmatic new approach to account for residual rate variation, which was ultimately crucial for controlling the pipeline's error rates. Lastly, I use the pipeline to show that rates of leaf and flower evolution are heterogeneous yet unrelated to overall size in a clade of eucalyptus trees ranging from  $\sim 1$  to nearly 100 meters in maximum height.

In chapter 3, I devise a new approach for inferring associations between discrete factors and continuous trait evolution dynamics by jointly modeling the evolution of both discrete factors and continuous traits under a unified process. A key advantage of this method is that it allows the continuous trait data to directly influence the likelihood of different factor histories, enabling inference unobserved discrete factors or “hidden states” potentially driving residual rate variation in continuous trait evolution. I implement this method in an R package, *sce* (<https://github.com/bstaggmartin/sce/>), and show that the method can effectively detect and quantify heterogeneity in rates of continuous trait evolution driven by both observed and unobserved factors under a wide variety of simulated evolutionary scenarios. Further, I demonstrate the empirical utility of the new method by using it to rigorously show that tropical sage lineages exhibit elevated rates of flower size evolution compared to temperate lineages.

This thesis is dedicated to my wife, Eleanore Jeanne Ritter, who always believed in me even when I didn't believe in myself, and my dog, Mulberry Ritter-Martin, who reminded me to stop and sniff the flowers from time to time. You two are my best friends in the entire world, and I would not have gotten to this point without your generous love and support throughout this crazy journey.

## ACKNOWLEDGEMENTS

Science is, at its core, a cumulative and community-driven endeavor that thrives on open, supportive communication and collaboration. As scientists, we all stand on the shoulders of countless dreamers, thinkers, and innovators that came before us. Further, the legacy of our research only reaches its full potential when embedded in the larger ecosystem of contemporary and historical work that motivates, contextualizes, and challenges our findings. Every individual has their own way of interpreting and describing the world around them—it is only through the sharing and reconciliation of all our unique perspectives that humanity achieves greater understanding of the universe’s greatest mysteries. I digress—but, in any case, this dissertation would not have been possible without the greater scientific community, not to mention the direct support—both professional and personal—of numerous mentors, colleagues, friends, and family.

I would first like to thank my advisor, Dr. Marjorie Weber, who gave me the freedom and independence to chase my scientific passions along with the support to both turn my craziest ideas into practical research projects and overcome the countless obstacles one faces in carrying out any long-term research project. Marjorie, your unwavering emphasis on the importance of inclusivity, creativity, and awe in scientific research continues to inspire me, and I will never be able to thank you enough for all you’ve taught me over the course of my dissertation.

Next, I would like to thank the rest of my guidance committee—Dr. Gideon Bradburd, Dr. Jeffrey Conner, Dr. Luke Harmon, and Dr. William Wetzel—who have all also provided invaluable inspiration and support over the years. Gideon, I would not be where I am today if it weren’t for your incredible mentorship in statistics, not to mention your wisdom in navigating the boundary between empirical research and method development. Jeff, our conversations have repeatedly challenged me to reconsider the “conventional macroevolutionary wisdom” I sometimes take for granted, and I am undoubtedly a much more learned and critical evolutionary biologist for it. Luke, your book was my initial ticket into the world of phylogenetic comparative methods development—I greatly appreciate all the enthusiasm you’ve shown for my research over the years, and I look forward to meeting up at Evolution and collaborating on research projects for many years to come.

Will, your excitement for science is positively infectious, and I've greatly appreciated our correspondence over the years—teaching a statistics course with you was an absolute blast!

I am of course indebted to countless faculty and staff at Michigan State University beyond my committee. In particular, I would like to thank Dr. Chad Niederhuth and Dr. Lars Brudvig for providing me with both lab space and community after Marjorie took up her new position at the University of Michigan. Additionally, the Michigan State University Herbarium was an invaluable resource to me throughout graduate school, thanks in no small part to the incredible efforts of both Matt Chansler and Dr. Alan Prather. I also want to thank Sara Kraeuter, whose devotion to helping graduate students is nothing short of incredible, along with the rest of the Plant Biology front office staff—Kelley Rose, Heather Stallone, Krystal Witt, and Trevor Simmons—who, along with the tireless efforts of Dr. Andrea Case, have breathed new life into the Plant Biology community following its pandemic slump.

Next, I want to thank Weber lab members both past and present, official and honorary. Thank you to Vincent Pan, Caroline Edwards, Erika LaPlante, Michael Foisy, Carolyn Graham, Rosy Glos, Sylvie Martin-Eberhardt, Abbey Soule, Dr. Ash Zemenick, Dr. Eric LoPresti, Dr. Andrew Myers, among many others. There are too many of you and unfortunately not enough time and space to write out all the things I could thank you for, but suffice to say I couldn't have asked for a more supportive and engaging lab community. Together, we form a collective force of scientific passion, creativity, and brilliance to be reckoned with!

I am also grateful to the broader community of my graduate student and postdoctoral peers at Michigan State University—special shout outs to Dr. Nate Catlin, Dr. Leslie Kollar, Riley Pizza, Meaghan Clark, Olivia Fitch, Brooke Jeffery, Maya Wilson-Brown, Miles Roberts, Sophie Buysse, Madison Plunkert, Andrew Bleich, Julia Brose—the list goes on and on. You all rock, and I'll really miss the found family dinners, Friday coffee hours, and DnD sessions. Thank you all for building such an open, welcoming, and supportive community in which all could find a niche and thrive.

Outside of Michigan State University, I would not be where I am today without the support of my undergraduate mentors based at Skidmore College. In particular, I would like to thank

Dr. Elaine Larsen, Denise McQuade, and Dr. Patti Steinberger for initializing my training in laboratory technician and teaching skills; Dr. Abbey Drake, Dr. Monica Raveret Richter, and Dr. Corey Freeman-Gallant for providing me with fascinating and invaluable undergraduate research opportunities; and Dr. Lucy Oremland and Dr. Julie Douglas for sparking my interest in the more mathematical aspects of biology.

There are of course many others located at a wide variety of other institutions that played key roles in the development of my academic career. Thanks to Dr. Michael Brett-Surman, who took me on as a collections management intern at the Smithsonian Natural History Museum when I was but a mere high school student, thus sparking a passion for museums and research collections that persists to this day. I am also grateful for the formative research experiences afforded to me under through the Mountain Lake Biological Station Research Experience for Undergraduates program under Dr. Chloe Lash and Charles Kwit, as well as through the School for Field Studies study abroad program in Queensland, Australia under Dr. Catherine Pohlmann. By crystallizing my interests in botany and entomology, these research experiences played vital roles in shaping my future academic career.

I also really want to thank the various phylogenetic comparative developers and systematists who have taken the time to talk to me throughout my graduate school career, offering valuable insights and wisdom that helped me (and my research) get to where it is today. So thank you to Dr. James Leebens-Mack, Dr. James Boyko, Dr. Josef Uyeda, Dr. Rosana Zenil-Ferguson, Dr. Michael Landis, Dr. Jeremy Beaulieu, Dr. Daniel Caetano, Dr. Liam Revell, and Dr. Michael Donoghue.

Additionally, I must thank my family for all their love and support. Thank you to my father, Andy, for exposing me to the wonders of the natural world from a young age and always admitting when he didn't know the answers to my incessant questions. Dad, you taught me how to find answers on my own and think for myself, nurturing my curiosity and contributing in no small part to where I am today. Thank you to my mother, Karen, for always encouraging me to follow my dreams while nonetheless emphasizing the importance of academics and hard work in achieving

those dreams. Mom, balancing idealism with pragmatism is often a difficult task, and I'm grateful to have your teachings and example to follow in this regard. I also want to thank my siblings, Bobby and Clare, as well as my extended family and in-laws, who always showed excitement for my career choice and projects even as my research interests grew more esoteric.

Lastly, but of course not least, I extend my deepest, sincerest gratitude to my wife, Eleanore, and dog, Mulberry. Both of you provided me with more companionship and encouragement I could have ever asked for throughout this process, both grounding me whenever my head "was in the clouds" and building me back up in the moments I felt defeated. I certainly would not have gotten anywhere close to this point without both of you in my life. Eleanore, you are the most brilliant and generous person I know, and I will never be able to fully express how thankful I am for all the support you gave me even as you arduously worked on your dissertation. As this chapter of our lives draws to a close, I look forward to the future, excited to see what life has in store for us.

## TABLE OF CONTENTS

CHAPTER 1	MODELING THE EVOLUTION OF RATES OF CONTINUOUS TRAIT EVOLUTION . . . . .	1
1.1	Introduction . . . . .	1
1.2	Materials and Methods . . . . .	4
1.3	Results . . . . .	17
1.4	Discussion . . . . .	26
	BIBLIOGRAPHY . . . . .	32
APPENDIX 1A	SUPPLEMENTAL TABLES AND FIGURES . . . . .	40
APPENDIX 1B	APPROXIMATING GEOMETRIC BROWNIAN MOTION TIME-AVERAGES . . . . .	44
APPENDIX 1C	AVERAGE CHANGES IN TRAIT EVOLUTION RATES . . . . .	62
APPENDIX 1D	PRIOR SENSITIVITY STUDY . . . . .	71
CHAPTER 2	STOCHASTIC CHARACTER MAPPING OF CONTINUOUS TRAITS ON PHYLOGENIES . . . . .	89
2.1	Introduction . . . . .	89
2.2	Materials and Methods . . . . .	91
2.3	Results . . . . .	109
2.4	Discussion . . . . .	117
	BIBLIOGRAPHY . . . . .	125
APPENDIX 2A	SUPPLEMENTAL TABLES AND FIGURES . . . . .	134
APPENDIX 2B	STOCHASTIC APPROXIMATION OF LIKELIHOOD FUNCTION GRADIENTS . . . . .	142
APPENDIX 2C	GENERATING CONTSIMMAPS UNDER EVORATES MODELS . . . . .	147
CHAPTER 3	A NEW APPROACH FOR INFERRING STATE-DEPENDENT VARIATION IN CONTINUOUS TRAIT EVOLUTION DYNAMICS . . . . .	158
3.1	Introduction . . . . .	158
3.2	Materials and Methods . . . . .	162
3.3	Results . . . . .	177
3.4	Discussion . . . . .	186
	BIBLIOGRAPHY . . . . .	192
APPENDIX 3A	SUPPLEMENTAL TABLES AND FIGURES . . . . .	200
APPENDIX 3B	PRUNING ALGORITHM DETAILS . . . . .	211

## CHAPTER 1

### MODELING THE EVOLUTION OF RATES OF CONTINUOUS TRAIT EVOLUTION

**This chapter has been published in *Systematic Biology*:**

Martin B.S.<sup>1</sup>, Bradburd G.S.<sup>2</sup>, Harmon L.J.<sup>3</sup>, and Weber M.G.<sup>2</sup> 2023. Modeling the evolution of rates of continuous trait evolution. *Syst Biol* 72:590–605.

<sup>1</sup>Department of Plant Biology, Michigan State University, East Lansing, MI, USA

<sup>2</sup>Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

<sup>3</sup>Department of Biological Sciences, University of Idaho, Moscow, ID, USA

#### 1.1 Introduction

The rates at which traits evolve is markedly heterogeneous across the tree of life, as evidenced by the uneven distribution of phenotypic diversity across space, time, and taxa (e.g., Simpson, 1944; Reaney et al., 2020; Brusatte et al., 2012; Chartier et al., 2021). While understanding the drivers of such patterns can provide critical insights into macroevolutionary processes, general consensus on what factors are most important in accelerating and decelerating trait evolution remain elusive (Chira et al., 2018). There is a vast, interconnected web of factors hypothesized to affect trait evolution rates, typically divided into extrinsic and intrinsic components. Extrinsic factors relate to the environment of an evolving lineage, commonly including aspects of biogeography like climate or habitat (e.g., Clavel and Morlon, 2017; Mihalitsis and Bellwood, 2019), as well as interactions with other species (e.g., Slater, 2015; Borstein et al., 2019; Drury et al., 2021). Intrinsic factors instead involve properties of the evolving lineage itself, including life history attributes such as behavior or developmental traits (e.g., Muñoz and Bodensteiner, 2019; Fabre et al., 2020) and genetic features like trait heritability and effective population size (e.g., Arnold et al., 2008; Villar et al., 2014). The effects of all these variables are interrelated and depend on the particular traits being studied, further complicating matters (Cooper and Purvis, 2009; Muñoz et al., 2018; see also Donoghue and Sanderson, 2015).

Unfortunately, the evolutionary histories of many factors hypothesized to affect trait evolution rates are largely unobserved. Thus, methods testing for associations between rates and variables

of interest must first estimate the history of the explanatory variables themselves (but see Hansen et al., 2022). This limits researchers to considering only a few, relatively simple hypotheses (Revell, 2013; Caetano and Harmon, 2019), causing trait evolution models to often underfit observed data (Pennell et al., 2015; Chira and Thomas, 2016; Chira et al., 2018). This underfitting generally oversimplifies inferred rate variation patterns and artificially increases statistical support for complex models which may imply spurious links between trait evolution rates and explanatory variables (May and Moore, 2020; see also Rabosky and Goldberg, 2015; Beaulieu and O’Meara, 2016). Thus, these “hypothesis-driven” approaches to modeling trait evolution should be integrated with “data-driven” approaches that agnostically model variation in trait evolution rates based on observed trait data alone. Such approaches can account for rate variation unrelated to some focal hypothesis, or even be used to generate novel hypotheses regarding what factors may have driven inferred rate variation patterns (Uyeda et al., 2018; May and Moore, 2020; see also Beaulieu and O’Meara, 2016).

Several data-driven methods for inferring trait evolution rates are already available and widely used (Eastman et al., 2011; Thomas and Freckleton, 2012; Rabosky et al., 2014; Pagel et al., 2022), but such methods generally work by splitting phylogenetic trees into subtrees and assigning a unique rate to each subtree (sometimes termed “macroevolutionary regimes”). These models implicitly assume trait evolution rates stay constant over long periods of time with sudden shifts in particular lineages. This mode of rate variation would be expected if rates are primarily influenced by only a few, discretely varying factors of large effect. However, this assumption could be problematic given the sheer number of factors hypothesized to affect trait evolution rates, as well as the fact that many of these factors vary continuously (Cooper and Purvis, 2009). If rates are instead affected by many factors, mostly with subtle effects, we would expect trait evolution rates to constantly shift in small increments over time within a given lineage, resulting in gradually changing rates over time and phylogenies. In other words, rates themselves would “evolve” and be similar, but not identical, among closely-related lineages (i.e., phylogenetic autocorrelation; see Sakamoto and Venditti, 2018). By assuming that rates change infrequently, current data-driven methods

likely oversimplify rate variation patterns, collapsing heterogeneous evolutionary processes into homogeneous regimes (but see May and Moore, 2020; Fisher et al., 2021). To this end, Revell (2021) recently developed a data-driven method that models trait evolution as gradually changing, but this method is limited in requiring *a priori* specification of how much trait evolution rates vary across the phylogeny. Further, the method offers no way to rigorously test whether lineages exhibit different rates (Revell, 2021).

Notably, some *hypothesis-driven* methods model trait evolution rates as gradually changing over time. However, such models most commonly assume that rates only follow a simple trend of exponential decrease or increase over time (Blomberg et al., 2003; but see Clavel and Morlon, 2017; Slater et al., 2017). In this context, declining trait evolution rates, or “early bursts” (EB), are often invoked as signatures of adaptive radiation (Harmon et al., 2010), while increasing trait evolution rates, or “late bursts” (LB), are sometimes linked to processes like character displacement (Weber et al., 2016; Skeels and Cardillo, 2019). Unfortunately, current methods lack statistical power to detect decreasing trends in rates when just a few lineages deviate from an overall EB pattern (Slater and Pennell, 2014). Essentially, by assuming a perfect correspondence between time and rates across all lineages, inference under these methods is misled by subclades exhibiting anomalously low or high trait evolution rates. New methods that explicitly model such “residual” rate variation may more accurately detect general trends in trait evolution rates by accounting for these anomalous lineages/subclades.

Here we develop a new, data-driven method that models trait evolution rates as gradually changing over time, ultimately resulting in stochastic, continuously distributed rates that are more similar among closely-related lineages. We take advantage of recent developments in Bayesian inference and develop new strategies for efficiently estimating autocorrelated rates on phylogenetic trees while dealing with uncertain trait values, resulting in relatively fast, reliable inference. We call this method (and its corresponding software implementation) “evolving rates” or *evorates* for short. *Evorates* is both flexible and intuitive, allowing researchers to infer both how and where rates vary on a phylogeny. Through simulation, we demonstrate that *evorates* recovers accurate parameter

estimates on ultrametric phylogenies spanning a range of sizes and that it is more sensitive and robust in detecting trends in trait evolution rates than conventional EB/LB models. We also use *evorates* to model body size evolution among extant whales and dolphins (order cetacea) and find evidence for declining rates of body size evolution and moderate rate heterogeneity in this clade, unifying and expanding on previous results (Slater et al., 2010; Slater and Pennell, 2014; Sander et al., 2021).

## 1.2 Materials and Methods

*Evorates* uses comparative data on a univariate continuous trait to infer how trait evolution rates change over time as well as which lineages in a phylogeny exhibit anomalous rates. Here, comparative data refers to a fixed, rooted phylogeny with branch lengths proportional to time and trait values associated with its tips. We generally caution against using *evorates* with univariate ordinations of multivariate trait data such as principal component scores because ordination can bias rate inference from comparative data (Uyeda et al., 2015). *Evorates* is designed to work with raw trait measurements; both missing data and multiple trait values per tip are allowed (i.e., tips with 0 and  $> 1$  observations, respectively). In the case of averaged trait measurements, estimated mean trait values and standard error can be used to specify normal priors on trait values at particular tips. The current implementation also allows for assigning raw trait measurements and priors to internal nodes as well, perhaps reflecting fossil data and/or strong prior beliefs, though we do not test this feature here. Conditional on these trait data, *evorates* uses Bayesian inference to estimate two key parameters governing the process of rate change: rate variance, controlling how quickly rates diverge among independently evolving lineages, and a trend, determining whether rates tend to decrease or increase over time. When rate variance is 0, rates do not accumulate random variation over time and are constant across contemporaneous lineages. In this case, trait evolution follows the same exact process as expected under a conventional EB/LB model, with negative trends corresponding to EBs, no trend to Brownian motion (BM), and positive trends to LBs. The method also infers branchwise rates, which are estimates of average trait evolution rates along each branch in the phylogeny, indicating which lineages exhibit unusually low or high rates.

### 1.2.1 The Model

At its core, *evorates* works by extending a typical Brownian motion (BM) model of univariate trait evolution to include stochastic, incremental changes in trait evolution rates,  $\sigma^2$ . Specifically,  $\sigma^2$  follows a process approximating geometric BM (GBM) with a constant rate, meaning that  $\ln(\sigma^2)$  follows a homogeneous BM-like process. GBM is a natural process to describe “rate evolution” because it ensures rates stay positive and implies rates vary on a multiplicative, as opposed to additive, scale (Limpert et al., 2001; Gingerich, 2009). To render inference under this model tractable, we treat it as a hierarchical model with a trait evolution process dependent on the unknown—but estimable—branchwise rates, which are themselves dependent on a rate evolution process controlled by the estimated rate variance and trend parameters. The overall posterior probability of the model can be summarized as:

$$P(\overline{\sigma^2}, \theta | x, \psi) \propto P(x | \psi, \overline{\sigma^2}) P(\overline{\sigma^2} | \psi, \theta) P(\theta) \quad (1)$$

Where  $\psi$  is a phylogeny with  $e$  branches and  $n$  tips,  $\overline{\sigma^2}$  is an  $e$ -length vector of branchwise rates,  $x$  is an  $n$ -length vector of trait values for each tip, and  $\theta$  is a vector of parameters governing the rate evolution process. Cases with missing data and multiple trait values per tips are covered in a later section. In our notation, time is 0 at the root of the phylogeny and increases towards the tips.  $P(x | \psi, \overline{\sigma^2})$  is the likelihood of  $x$  given the trait evolution process,  $P(\overline{\sigma^2} | \psi, \theta)$  is the probability of branchwise rates given the rate evolution process, and  $P(\theta)$  is the prior probability of the rate evolution process parameters. We explicitly estimate and condition likelihood calculations on branchwise rates (a type of “data augmentation”; see May and Moore, 2020) because the likelihood of the trait data while marginalizing over branchwise rates (i.e.,  $P(x | \psi, \theta)$ ) does not follow a known probability distribution and would require complex, numerical approximations to compute. On the other hand,  $P(x | \psi, \overline{\sigma^2})$  follows a straight-forward multivariate normal density:

$$x \sim MVN(\alpha, C) \quad (2)$$

where  $\alpha$  is a vector of the trait value at the root of the phylogeny repeated  $n$  times and  $C$  is an  $n \times n$  matrix. The entries of  $C$  are given by:

$$C_{i,j} = \sum_{k \in \text{anc}(i,j)} \overline{\sigma^2}_k t_k \quad (3)$$

where  $t$  is an  $e$ -length vector of branch lengths,  $i$  and  $j$  are indices denoting specific tips,  $k$  is an index denoting a particular branch, and  $\text{anc}(i, j)$  is a function that returns all ancestral branches shared by  $i$  and  $j$ . Note that when branchwise rates are constant across the tree,  $C_{i,j}$  is proportional to the elapsed time between the root of the phylogeny and the most recent common ancestor of  $i$  and  $j$ . Branchwise rates can be thought of as “squashing” and “stretching” the branch lengths of a phylogeny, such that certain lineages have evolved for effectively shorter or longer amounts time, respectively.

Unfortunately, there is no general solution for calculating  $P(\overline{\sigma^2} | \psi, \theta)$  under a true GBM process (Lepage et al., 2007), so we instead use a multivariate log-normal approximation (e.g., Dufresne, 2004; Welch and Waxman, 2008) of the distribution of branchwise rates and calculate probabilities under this approximation. Briefly, this approximation decomposes branchwise rates into their expected values,  $\beta$ , determined solely by the trend parameter, and a “noise” component,  $\gamma$ , sampled from a multivariate normal distribution controlled by the rate variance parameter:

$$\ln(\overline{\sigma^2}) \approx \beta + \gamma \quad (4)$$

Here, the noise component is approximate because it follows the distribution of geometric, rather than arithmetic, averages of trait evolution rates along each branch assuming there is no trend (i.e.,  $\overline{\ln(\sigma^2)}$  rather than  $\ln(\overline{\sigma^2})$ ; see APPENDIX 1B for further details). The entries of  $\beta$  are given by:

$$\beta = \ln(\sigma_0^2) + \begin{cases} 0 & \text{if } \mu_{\sigma^2} = 0 \\ \ln(|\exp[\mu_{\sigma^2} \tau_2] - \exp[\mu_{\sigma^2} \tau_1]|) - \ln(|\mu_{\sigma^2}|) - \ln(t) & \text{if } \mu_{\sigma^2} \neq 0 \end{cases} \quad (5)$$

where  $\ln(\sigma_0^2)$  is the estimated rate at the root of the phylogeny,  $\mu_{\sigma^2}$  is the trend parameter,  $t$  is an  $e$ -length vector of branch lengths, and  $\tau_1$  and  $\tau_2$  are  $e$ -length vectors of the start and end times of each branch in the phylogeny (Blomberg et al., 2003). The entries of  $\gamma$  are given by:

$$\gamma \sim MVN(0, \sigma_{\sigma^2}^2 D) \quad (6)$$

where 0 is a vector of 0s repeated  $e$  times,  $\sigma_{\sigma^2}^2$  is the rate variance parameter, and  $D$  is an  $e \times e$  matrix. The entries of  $D$  are given by:

$$D_{i,j} = \sum_{k \in \text{anc}(i,j)} t_k - \begin{cases} 2t_i/3 & \text{if } i = j \\ t_i/2 & \text{if } i \in \text{anc}(j, j) \\ t_j/2 & \text{if } j \in \text{anc}(i, i) \\ 0 & \text{if } i \neq j, i \notin \text{anc}(j, j), j \notin \text{anc}(i, i) \end{cases} \quad (7)$$

where  $i$ ,  $j$ , and  $k$  are all indices denoting branches and  $\text{anc}(i, j)$  is a function that returns all ancestral branches shared by  $i$  and  $j$  (Devreese et al., 2010; see APPENDIX 1B for further details). Overall, this approximation closely matches the distribution of branchwise rates obtained via fine-grained simulations of GBM on phylogenies under plausible parameter values and is negligibly different from other computationally efficient approximations (e.g., Thorne et al., 1998; Lartillot and Poujol, 2011; Revell, 2021; Figs. 1B.1–1B.14; Tables 1B.1–1B.3). We prefer this approximation because it is convenient to work with and directly focuses on estimating branchwise rates rather than rates at the nodes of the phylogeny, which is what other strategies focus on.

Under this approximation, the final expression for the posterior probability is:

$$P(\ln(\overline{\sigma^2}), \alpha, \ln(\sigma_0^2), \sigma_{\sigma^2}^2, \mu_{\sigma^2} | x, \Psi) \propto \frac{\exp[-\frac{1}{2}(x - \alpha)'C^{-1}(x - \alpha)]}{\sqrt{(2\pi)^n |C|}} \frac{\exp[-\frac{1}{2}(\ln(\overline{\sigma^2}) - \beta)'(\sigma_{\sigma^2}^2 D)^{-1}(\ln(\overline{\sigma^2}) - \beta)]}{\sigma_{\sigma^2} \sqrt{(2\pi)^e |D|}} \quad (8)$$

$$P(\alpha, \ln(\sigma_0^2), \sigma_{\sigma^2}^2, \mu_{\sigma^2})$$

### 1.2.2 Model Implementation

*Evorates* estimates the posterior distribution of parameters given a phylogeny and associated trait data via Hamiltonian Monte Carlo (HMC) using the probabilistic programming language Stan, interfaced through R (Carpenter et al., 2017; Stan Development Team, 2019, 2020). Unlike conventional Markov Chain Monte Carlo algorithms like Metropolis-Hastings samplers, HMC uses derivatives and physics simulations to efficiently explore posterior distributions, which is particularly helpful for complex, high-dimensional posteriors (see Neal, 2011 and Hoffman and Gelman, 2014 for further information). To optimize sampling efficiency and avoid numerical issues, *evorates* estimates branchwise rates with an uncentered parameterization (Betancourt and Girolami, 2019) and marginalizes over unobserved trait values at the root and tips of the tree (Freckleton, 2012; Hassler et al., 2022). Under an uncentered parameterization, the HMC algorithm does not directly estimate branchwise rates, but instead estimates the distribution of  $e$  independent standard normal random variables,  $z$ , which are transformed to follow the distribution of branchwise rates:

$$\ln(\overline{\sigma^2}) = \sigma_{\sigma^2} Lz + \beta \quad (9)$$

where  $L$  is lower triangular Cholesky factorization of  $D$  (i.e.,  $D = LL'$ ; see Eq. (7)). This parameterization is particularly efficient because it avoids having to repeatedly manipulate  $D$  to calculate  $P(\ln(\overline{\sigma^2}) | \psi, \ln(\sigma_0^2), \sigma_{\sigma^2}^2, \mu_{\sigma^2})$ .

*Evorates* also uses Felsenstein’s pruning algorithm for quantitative traits to marginalize over the trait value at the root of the phylogeny and avoid repeatedly inverting  $C$  when calculating  $P(x | \ln(\overline{\sigma^2}))$  (Felsenstein, 1973; Freckleton, 2012; Caetano and Harmon, 2019). To simplify the pruning algorithm implementation, any multifurcations in the phylogeny are converted to series of bifurcations by adding additional “pseudo-branches” of length 0. This procedure does not alter the resulting likelihood calculations (Felsenstein, 2008), and our implementation does not estimate branchwise rates along pseudo-branches because these rates do not affect the likelihood of the observed trait data.

### 1.2.3 Accommodating Missing Data and Multiple Observations

Incorporating uncertainty in observed trait values in comparative studies is especially important for methods that model trait evolution rate variation because measurement error can inflate estimates of evolutionary rates, particularly in young clades (Felsenstein, 2008). To prevent such biases, *evorates* generally treats the mean trait values at the tips,  $x$ , as unknown parameters. We marginalize over  $x$  given raw trait measurements,  $y$  (potentially including 0 or  $> 1$  observations for some tips), and “tip error” variances for each tip,  $\sigma_y^2$ . While we use the term “raw” trait measurement for clarity, the data provided for certain tips could instead be the mean of a normal prior on the trait value. Entries of  $\sigma_y^2$  for such tips may be fixed to an associated variance for the prior. All other entries of  $\sigma_y^2$  are treated as unfixed, free parameters. To render the model more tractable, we assume tip error variance is constant across all tips with unfixed variance.

To marginalize over the mean trait values at the tips, we modify the initialization of Felsenstein’s pruning algorithm (Felsenstein, 1973). Prior to pruning, we assign each tip the expectation and variance of its mean trait value given its raw trait measurements. We then calculate each tip’s partial likelihood from contrasts between its associated raw trait measurements given its error variance,  $\sigma_{y,i}^2$ . Assuming the raw trait measurements are independently sampled from a normal distribution with variance  $\sigma_{y,i}^2$ , the mean trait value’s expectation is simply the mean of the raw trait measurements,  $\bar{y}_i$ , and its variance is given by  $\sigma_{y,i}^2/m_i$ , where  $m_i$  is the number of raw trait measurements (Felsenstein, 2008). Note that if there are no trait measurements for a particular tip (i.e.,  $m_i = 0$ ), the expectation of that tip’s true trait value is undefined with infinite variance (Hassler et al., 2022).

Because there are no contrasts for tips with one or fewer raw trait measurements, the partial likelihood associated with these tips is 1. Otherwise, we can derive a general formula for the partial likelihood by considering each tip as a small subtree and applying Felsenstein’s pruning algorithm. Specifically, each tip is treated as a star phylogeny consisting of  $m_i$  “sub-tips” of length  $\sigma_{y,i}^2$ , with trait values  $y_i$  (Felsenstein, 1973, 2008):

$$P(y_i | \sigma_{y,i}^2) = \prod_{k=1}^{m_i-1} \frac{\sqrt{k}}{\sigma_{y,i} \sqrt{2\pi(k+1)}} \exp \left[ -\frac{k}{2(k+1)} \left( \frac{y_{i,k+1} - \bar{y}_{i,1:k}}{\sigma_{y,i}} \right)^2 \right] \quad (10)$$

where  $i$  denotes a particular tip,  $y_i$  is a vector of  $m_i$  raw trait measurements for tip  $i$ ,  $\sigma_{y,i}^2$  is the tip error variance for tip  $i$ , and  $\bar{y}_{i,1:k}$  is the mean of measurements 1 through  $k$  in the vector  $y_i$ .

After initializing all tips in the phylogeny, Felsenstein's pruning algorithm can be applied normally, iterating over the internal nodes from the tips towards the root (e.g., Felsenstein, 1973, Freckleton, 2012, Caetano and Harmon, 2019). The presence of missing data, however, will cause some calculations to involve nodes with undefined expected trait values and infinite variance. Note that these "data-deficient" nodes do not contribute information to the expectation and variance of the trait value at their ancestral nodes. Thus, if both nodes descending from some focal node are data deficient, the focal node will also be data deficient, with undefined expectation and infinite variance. Otherwise, if only one descendant node is data deficient, the expectation and variance of the trait value at the focal node is solely determined by the descendant node that is *not* data deficient. Let the descendant, non-data deficient node have expected trait value and variance  $\hat{x}_i$  and  $\sigma_{\hat{x}_i}^2$ , respectively, and be connected to the focal node by a branch of length  $t_i$  with branchwise rate  $\bar{\sigma}_i^2$ . The focal node's expected trait value and variance will be  $\hat{x}_i$  and  $\sigma_{\hat{x}_i}^2 + \bar{\sigma}_i^2 t_i$ , respectively. Whether one or both descendant nodes are data deficient, there is no contrast associated with the focal node and the corresponding partial likelihood is 1.

In the case of univariate traits, tips with missing data have no effect on the likelihood of trait data or parameter inference. However, by including missing data one can estimate posterior distributions of the unobserved trait values at these tips (Goolsby, 2017; Hassler et al., 2022). *EvoRates* already includes functionality for sampling from the posterior distribution of trait values at all nodes and tips in a phylogeny given a fitted model. The inclusion of additional branches could theoretically affect the inferred rate evolution process because our GBM approximation improves along shorter branches. However, inference using *evorates* is robust to whether rate evolution is simulated under our GBM approximation or a true GBM process (Figs. 1B.10 and 1B.14; Tables 1B.1–1B.3), suggesting such effects are too minor to have practical consequences.

### 1.2.4 Priors

Despite their popularity, flat and uninformative priors tend to result in fat-tailed posteriors that explore unrealistic regions of parameter space, and Bayesian statisticians have increasingly advocated for the use of at least weakly informative priors in recent years (Lemoine, 2019). We follow this advice, choosing default priors for *evorates* that modestly regularize parameter estimates, promoting conservative inferences (i.e., little rate heterogeneity) while still allowing for a wide range of evolutionary dynamics. We also conducted a prior sensitivity study to document the impact of priors on inference using *evorates* (Figs. 1D.1–1D.7; Tables 1D.1–1D.12). Overall, *evorates* is fairly robust to alternate prior specifications, provided that priors are not overly informative, and the the default priors appear adequate under a variety of conditions.

By default, a normal prior with mean 0 and standard deviation  $10/T$  is placed on the trend parameter ( $\mu_{\sigma^2}$ ), while a Half-Cauchy prior with scale  $5/T$  is placed on rate variance ( $\sigma_{\sigma^2}^2$ ), where  $T$  is the height of the phylogeny. These priors are quite liberal: a trend of  $10/T$  corresponds to a  $e^{10} \sim 20,000$ -fold change in trait evolution rates over the timespan of a phylogeny, and data simulated with a rate variance of  $5/T$  on random trees with 50 tips or more (generated using the R package *ape* version 5.6-2; Paradis and Schliep, 2019) typically yield branchwise rates spanning two to four orders of magnitude. Of course, researchers may increase or decrease the standard deviation/scale of these priors if a phylogeny spans an especially long or short timescale, respectively. To penalize tip error variance ( $\sigma_y^2$ ) estimates that are large relative to the scale of the observed trait data, a half-Cauchy prior with scale  $\sigma_{raw}^2/2$  is placed on tip error variance, where  $\sigma_{raw}^2$  is the variance of the trait data.

It is somewhat more challenging to pick a default prior for the rate at the root ( $\sigma_0^2$ ) because this parameter depends on both the timescale of the phylogeny and scale of the observed trait data. By default, a log-normal prior with location  $\ln(\sigma_{raw}^2/T)$  and scale 10 is placed on the root rate. This prior is designed to regularize root rate estimation by roughly centering on trait evolution rates that could give rise to the observed trait data with little rate heterogeneity. Notably, decreasing and increasing trends will generally shift the location of this default prior downward and upwards,

respectively, relative to the true root rate. While more complex schemes for choosing a root rate prior (perhaps based on phylogenetic independent contrasts) could help mitigate this issue, we wanted to keep default prior settings as simple and transparent as possible. As a rule of thumb, the scale of the root rate prior should be roughly equal to the maximum plausible change in trait evolution rates over the timespan of a phylogeny. The default scale of 10, corresponding to a  $e^{10} \sim 20,000$ -fold change in rates, is quite liberal and should suffice for most purposes. In any case, we encourage researchers to alter the root rate prior to reflect biologically plausible trait evolution rates when such information is available.

### 1.2.5 Hypothesis Testing

We agree with other macroevolutionary biologists advocating for greater focus on interpreting parameter estimates and effect sizes inferred by comparative models (e.g., Beaulieu and O’Meara, 2016). Nonetheless, assessing statistical support for particular hypotheses remains important for biologically interpreting fitted models – particularly complex models with many parameters. In the context of *evorates*, we focus on two main hypotheses: 1) that significant rate heterogeneity, independent of any trend, occurred over the history of a clade ( $\sigma_{\sigma^2}^2 > 0$ ), and 2) rates generally declined or increased over time (i.e.,  $\mu_{\sigma^2} \neq 0$ ). Both hypotheses could be tested by fitting additional models with constrained rate variance and/or trend parameters and comparing among unconstrained and constrained models using Bayes factors. However, Bayes factor estimation requires additional, time-consuming computation. Thus, we developed alternative approaches that only require the posterior samples of a fitted, unconstrained model.

We use the posterior probability that  $\mu_{\sigma^2} > 0$  to test for overall trends in rates. If the posterior probability is 0.025 or less, we can conclude that there is substantial evidence that rates declined over time, and vice versa if the posterior probability is 0.975 or above. This corresponds to a two-tailed test with a critical value of 0.05. For rate variance, we instead use Savage-Dickey (SD) ratios because rate variance is bounded at 0 and the posterior probability that  $\sigma_{\sigma^2}^2 > 0$  will always be 1. SD ratios are ratios of the posterior to prior probability density at a particular parameter value corresponding to a null hypothesis. If this ratio is sufficiently less than 1, the data have “pulled”

prior probability mass away from the null hypothesis, suggesting that the null hypothesis is likely incorrect. In general, a ratio of 1/3 or less is considered substantial evidence against the null hypothesis (Kass and Raftery, 1995). We use log spline density estimation implemented in the R package *logspline* (version 2.1.16) to estimate the posterior probability density at  $\sigma_{\sigma^2}^2 = 0$  (Stone et al., 1997; Wagenmakers et al., 2010).

Researchers may also wish to identify lineages evolving at anomalous rates. The most straightforward method to do so is to calculate the posterior probability that branchwise rates are greater than some “background rate”, analogous to the approach for trends. In this paper, we define the background trait evolution rate as the geometric mean of branchwise rates, weighted by their relative branch lengths. Rates are generally distributed with long right tails (Gingerich, 2009), particularly under our model whereby rate evolution follows a GBM-like process. Geometric means are less sensitive than arithmetic means to extremely high, outlier rates associated with these long tails, and are thus better-suited for rate comparisons. In the presence of a strong trend, only the oldest and youngest lineages will generally exhibit anomalous rates, rendering anomalous rate detection redundant with trend estimation. Thus, we define a helpful branchwise rate transformation, called “detrending”, which further facilitates interpretation of *evorates* results. Specifically, branchwise rates are detrended prior to calculating background rates and posterior probabilities by subtracting  $\beta$  from branchwise rates on the natural log scale (see Eq. (5)). These detrended rates yield a new set of transformed parameters, branchwise rate deviations,  $\ln(\overline{\sigma_{dev}^2})$ , defined as the difference between detrended branchwise rates and the background detrended rate on the natural log scale. When the posterior probability  $\ln(\overline{\sigma_{dev}^2}) > 0$  for a given branch is less than 0.025 or greater than 0.975, we can conclude that trait evolution is anomalously slow or fast along that branch, respectively, given the overall trend in rates through time. While we focus on comparing detrended branchwise and background rates based on geometric means in the current paper, we note that *evorates* can also compare untransformed branchwise and background rates based on either geometric or arithmetic means per user specifications.

Additionally, users may also calculate background trait evolution rates for subsets of branches

in a phylogeny, such that rates for specific lineages and/or subclades can be estimated and compared. Some caution, however, is warranted in first identifying lineages exhibiting anomalous rates and then testing for significant differences among them, as this could increase the risk of spuriously detecting rate differences. This potential issue is not unique to *evorates* and applies to any data-driven phylogenetic comparative method designed to identify shifts in evolutionary processes. In practice, we recommend users mainly focus on interpreting comparisons between branchwise rates and the overall background rate, calculating background rates for branch subsets only to effectively summarize and communicate model results. Of course, it is also perfectly reasonable to compare rates among specific lineages and/or subclades when these comparisons are planned prior to model fitting and/or have biological justification (e.g., comparing background rates among lineages that vary in some factor hypothesized to affect trait evolution rates).

Notably, relationships among Bayes factors, posterior probabilities, and frequentist p-values are not necessarily straight-forward and depend on sample size, priors, and posterior distribution shape, among other factors (Held and Ott, 2018; Wagenmakers et al., 2022). The hypothesis testing procedures we propose and test here are essentially useful heuristics developed to guide researchers in interpreting models fit through *evorates*, and these heuristics are not formally equivalent to conventional significance testing under a frequentist framework. Nonetheless, we use terms like “hypothesis testing”, “null hypothesis”, and “significance” in describing and analyzing the performance of these heuristics for ease of communication.

### 1.2.6 Simulation Study

To test the performance and accuracy of *evorates*, we applied it to continuous trait data simulated under the model of inference. We simulated data under all combinations of no, low, and high rate variance ( $\sigma_{\sigma^2}^2 = 0, 3, 6$ ) and decreasing, constant, and increasing trends ( $\mu_{\sigma^2} = -4, 0, 4$ ), for a total of 9 trait evolution scenarios. We picked these values to simulate data that appeared empirically plausible and represented a range of different trait evolution dynamics. Note that when rate variance is 0, the resulting simulations evolve under EB, BM, or LB models of trait evolution depending on the trend parameter. We simulated traits evolving along ultrametric, pure-birth phy-

logenies with 50, 100, and 200 tips generated using the R package *phytools* (version 1.0-1; Revell, 2012) to assess the effect of increasing sample size on model performance. While *evorates* can be applied to non-ultrametric trees, we focus on ultrametric trees here to render the simulation study more manageable. We simulated 10 phylogenies and associated trait data for each trait evolution scenario and phylogeny size for a total of 270 simulations. In all cases, phylogenies were rescaled to a total height of 1, ensuring the effect of parameters remained consistent across replicates. All simulations were simulated with a trait and log rate value of 0 at the root. Because we focused on the estimation of branchwise rate, rate variance, and trend parameters, we simulated trait data with only 1 observation per tip and no tip error.

To quantitatively assess the simulation study results, we calculated the median absolute error (MAE), breadth, and coverage of marginal posterior distributions for rate variance and trend parameters. Here, MAE is the median absolute difference between posterior samples and their corresponding true, simulated value, such that larger MAEs are associated with less accurate posteriors. We prefer *median* to *mean* absolute error because the former metric is less influenced by posterior precision and more directly reflects variation in posterior accuracy. Breadth refers to the width of the 95% equal-tailed interval (i.e., a type of credible interval that spans from the 2.5% to 97.5% posterior quantiles, hereafter simply termed credible intervals) and measures posterior precision, with smaller breadths corresponding to more precise (though not necessarily accurate) posteriors. Lastly, coverage is a binary metric equal to one when the true value falls within the 95% credible interval and zero otherwise. For branchwise rate parameters, we averaged the MAEs, breadths, and coverage of all branchwise rate marginal posterior distributions (on the natural log scale) for each model fit. Additionally, we calculated the statistical power and false positive error rate (i.e., type I error rate, hereafter error rate) of *evorates* for detecting significant rate variance and decreasing/increasing trends. Due to the continuous nature of branchwise rates, we assessed power and error rates for detecting anomalous branchwise rates by calculating the proportion of times a branch is detected as exhibiting anomalously slow or rapid trait evolution rates across different values of true branchwise rate deviations.

### 1.2.7 Empirical Example

We applied *evorates* to model body size evolution in extant cetaceans using a recently estimated timetree of both fossil and extant cetaceans (Lloyd and Slater, 2021), pruned to consist of 88 extant species (we excluded one extant species, *Balaenoptera brydei*, due to its uncertain taxonomic status; see Constantine et al., 2018), and associated trait data on log-transformed maximum female body lengths for each species. Most body length data was compiled in a previous comparative study, but we supplemented these data with published measurements for an additional 15 species (Table 1A.1). We chose this example because previous research detected notable signatures of declining body size evolution rates over time in this clade, despite conventional model selection failing to yield support for an EB model of trait evolution. This puzzling result seems primarily due to a few recently-evolved lineages exhibiting unusually rapid shifts in body size (Slater et al., 2010; Slater and Pennell, 2014; see also Sander et al., 2021). While previous work used a mix of simulation and outlier detection techniques to arrive at this conclusion, we predicted that our method would identify these patterns in a more cohesive modeling framework.

### 1.2.8 HMC configuration and diagnostics

When fitting models to simulated and the empirical data, we ran 4 HMC chains consisting of 3,000 iterations. After discarding the first 1,500 iterations as warmup and checking for convergence, chains were combined for a total of 6,000 HMC samples for each simulation. We repeated this procedure while constraining the rate variance parameter to 0 to see if our method could detect trends in trait evolution rates with more power than conventional EB/LB models. We set tip error for the simulation study to 0 *a priori* because we do not focus on inference of this parameter here, though we did allow the method to estimate tip error in the empirical example. For each model fit, chains mixed well (greatest  $\hat{R} \approx 1.013$ ) and achieved effective sample sizes of at least 3,000 for every parameter. Divergent transitions, a feature of HMC which can be indicative of sampling problems, were relatively rare, with only six simulation model fits exhibiting 1-3 divergent transitions. Overall, diagnostic tests suggested all HMC chains converged and sampled posterior distributions thoroughly.

## 1.3 Results

### 1.3.1 Performance of Method

Overall, the method exhibited accurate inference and appropriate coverage for all parameters, though posterior breadth was often quite large, especially for trees with 50 tips (1.1–1.3, Fig. 1.1). Posterior accuracy and precision were highly dependent on trait evolution scenario and tree size. In general, higher values of trends and rate variance were associated with larger posterior MAEs and breadth for their respective parameters, such that increasing trends and high rate variance are estimated with the least accuracy and precision. In some cases, higher trends seemed to increase the MAEs and breadth of rate variance posteriors and vice versa, but this pattern was weak overall. On the other hand, larger tree sizes resulted in smaller posterior MES and breadth, such that trees with 200 tips yielded the most accurate, precise posteriors. Coverage for trend and rate variance parameters across all trait evolution scenarios and tree sizes remained consistent at around the theoretical expectation of 95%.

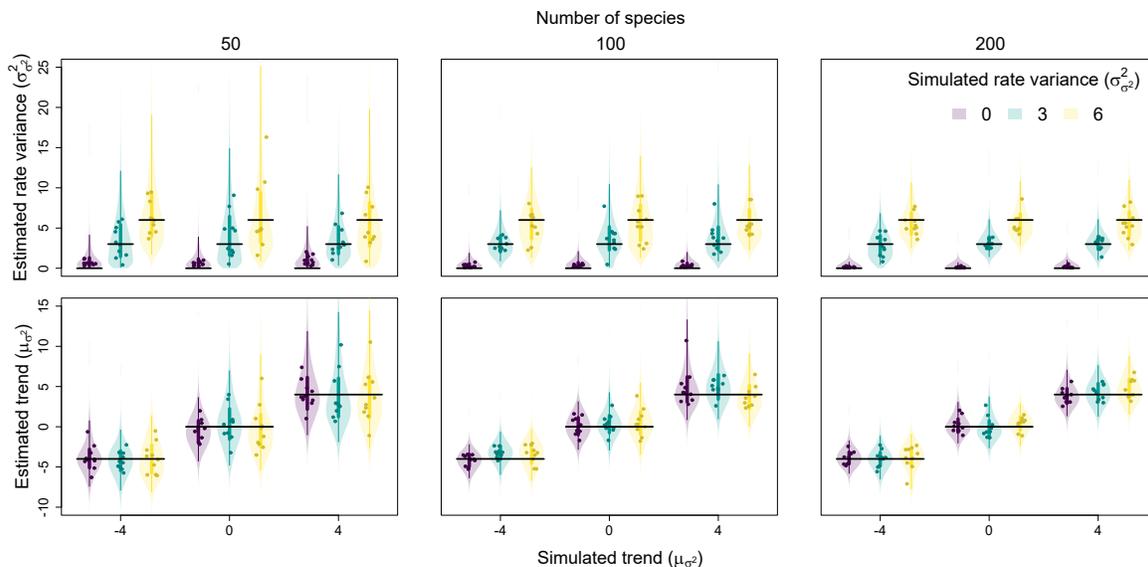


Figure 1.1 Relationship between simulated and estimated rate variance ( $\sigma_{\sigma^2}^2$ ) and trend ( $\mu_{\sigma^2}$ ) parameters. Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values.

Both the statistical power and error rates of our method were appropriate for detecting trends

Table 1.1 Median absolute errors of rate variance, trend, and branchwise rate posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates			
	0	3	6	0	3	6	0	3	6	
	50 species									
$\mu_{\sigma^2} =$	-4	0.66	1.96	2.55	1.36	1.29	1.83	0.47	0.81	1.00
	0	0.57	2.48	3.69	1.49	2.09	2.45	0.48	0.86	1.06
	4	0.99	1.75	3.00	2.06	2.79	2.91	0.60	0.87	1.01
	100 species									
$\mu_{\sigma^2} =$	-4	0.30	1.01	2.03	0.77	1.08	1.31	0.31	0.73	0.90
	0	0.37	1.62	2.37	1.12	1.20	1.59	0.37	0.76	0.89
	4	0.34	1.56	1.87	1.89	1.63	1.54	0.44	0.83	0.90
	200 species									
$\mu_{\sigma^2} =$	-4	0.13	1.27	1.50	0.77	0.95	1.25	0.24	0.66	0.80
	0	0.11	0.75	1.44	0.92	1.13	0.95	0.23	0.71	0.85
	4	0.18	0.82	1.69	1.00	1.13	1.35	0.27	0.72	0.84

and significant rate variance. In general, power increased with larger trees, while error rates remained consistent. The ability of SD ratios to identify significant rate variance was particularly impressive, erroneously detecting rate variance only once while exhibiting high power (Fig. 1.2). Decreasing trends were notably easier to detect than increasing trends, particularly on small trees (Fig. 1.3). Trend error rates consistently remained below  $\sim 5\%$ , and decreasing trends were never mistaken for increasing trends and vice versa. Higher rate variance seemed to only slightly decrease the power to detect trends. Constraining rate variance to 0 resulted in either worse power or higher error rates for detecting trends, depending on whether trends were decreasing or increasing. As rate variance increased, the power of constrained models to detect decreasing trends dramatically diminished. On the other hand, constrained models detected increasing trends with greater power, at the cost of greatly inflated error rates. Overall, estimating rate variance allows for more sensitive detection of declining trait evolution rates while better safe-guarding against false

Table 1.2 Breadths of rate variance, trend, and branchwise rate posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates			
	0	3	6	0	3	6	0	3	6	
50 species										
$\mu_{\sigma^2} =$	-4	3.85	9.07	15.05	5.03	6.08	6.71	2.33	3.17	3.76
	0	3.65	10.07	14.82	5.92	8.26	8.28	2.29	3.41	3.90
	4	4.52	8.66	14.05	10.73	10.75	10.75	3.01	3.49	3.85
100 species										
$\mu_{\sigma^2} =$	-4	1.56	5.60	8.53	3.27	4.65	4.84	1.66	2.92	3.35
	0	1.91	6.45	9.01	4.31	5.27	6.01	1.87	3.10	3.42
	4	1.69	6.47	8.39	7.61	8.42	7.39	2.06	3.32	3.60
200 species										
$\mu_{\sigma^2} =$	-4	0.69	4.13	6.43	2.80	3.59	4.01	1.23	2.51	3.06
	0	0.62	4.23	6.21	3.39	3.99	4.06	1.18	2.72	3.23
	4	0.79	3.89	6.14	4.50	5.21	5.65	1.39	2.83	3.22

detection of increasing rates.

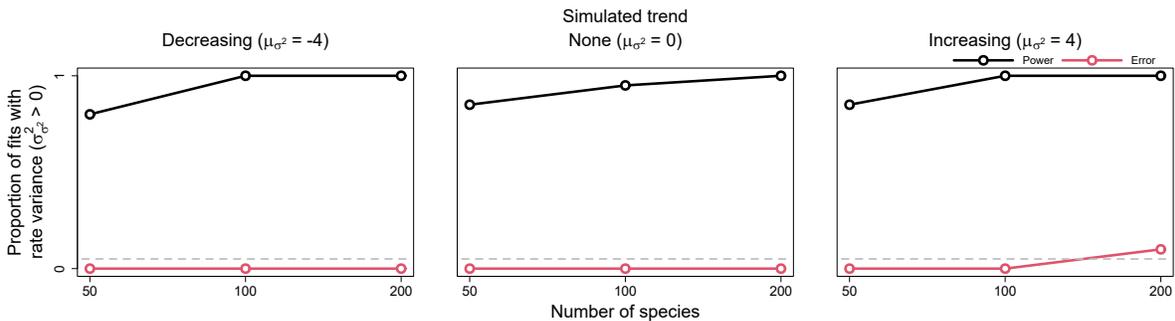


Figure 1.2 Power and error rates for the rate variance parameter ( $\sigma_{\sigma^2}^2$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for rate variance significantly greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of tree size.

Branchwise rate estimation also generally displayed appropriate coverage, accuracy, and statistical testing properties (Tables 1.1–1.3, Fig. 1.4). However, branchwise rate estimates were noticeably biased towards their overall mean (i.e., shrinkage). Linear regressions of median branchwise

Table 1.3 Coverage of rate variance, trend, and branchwise rate posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates			
	0	3	6	0	3	6	0	3	6	
50 species										
$\mu_{\sigma^2} =$	-4	—	0.90	1.00	0.80	1.00	1.00	0.98	0.95	0.92
	0	—	0.90	0.90	1.00	0.90	0.80	0.99	0.96	0.92
	4	—	1.00	0.90	1.00	0.90	0.90	0.99	0.96	0.92
100 species										
$\mu_{\sigma^2} =$	-4	—	1.00	0.90	1.00	1.00	1.00	1.00	0.97	0.92
	0	—	0.80	1.00	1.00	1.00	1.00	0.99	0.96	0.95
	4	—	0.90	1.00	0.90	1.00	1.00	0.97	0.95	0.96
200 species										
$\mu_{\sigma^2} =$	-4	—	0.90	1.00	1.00	1.00	0.90	1.00	0.94	0.94
	0	—	1.00	1.00	0.90	0.90	1.00	0.99	0.95	0.94
	4	—	1.00	0.90	1.00	1.00	0.90	1.00	0.96	0.95

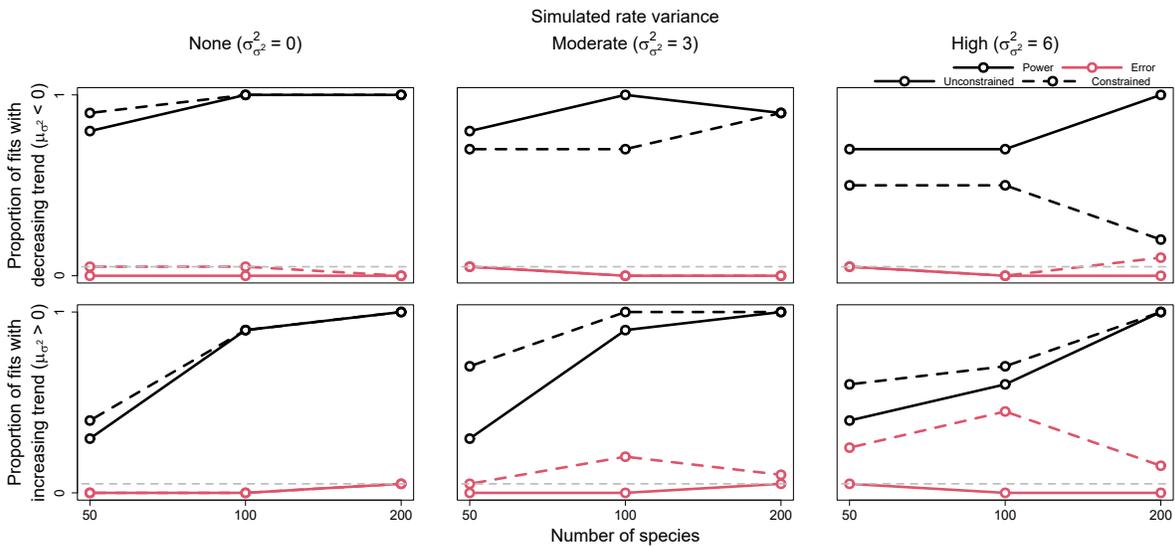


Figure 1.3 Power and error rates for the trend parameter ( $\mu_{\sigma^2}$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for trends significantly less and greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of tree size. Results are shown for both models allowed to freely estimate rate variance ( $\sigma_{\sigma^2}^2$ ) (i.e., unconstrained models, solid lines) and models with rate variance constrained to 0 (i.e., constrained models, dashed lines). The latter models are identical to conventional early/late burst models.

rate estimates on simulated branchwise rates yield an average slope of about 0.8 (Fig. 1.5). A similar pattern holds for linear regression of branchwise rate deviations (Fig. 1A.1). Branchwise rate posteriors for simulations with no rate variance exhibited especially high accuracy, precision, and coverage (notably above the theoretical expectation of 95%), perhaps due to the increased precision of rate variance posteriors under such trait evolution scenarios. In contrast to other parameters, increasing tree size only slightly decreased posterior MAEs and breadth for branchwise rates. After accounting for variation in simulated branchwise rate deviations, trait evolution scenario and tree size had little effect on statistical power and error rates for detecting anomalous branchwise rates. Averaging across all fits to simulations with significant rate variance detected, error rates for detecting anomalous rates remained negligible, peaking at around 0.5% for branchwise rate deviations around 0. In fact, this peak only increased to about 5% when we set the significant posterior probability thresholds to 10% and 90% (Fig. 1A.2). The method was somewhat more sensitive to positive than negative deviations, correctly and consistently detecting anomalous rates with deviations more extreme than  $-4$  (1/50th of background rate) or  $3$  (20 times background rate).

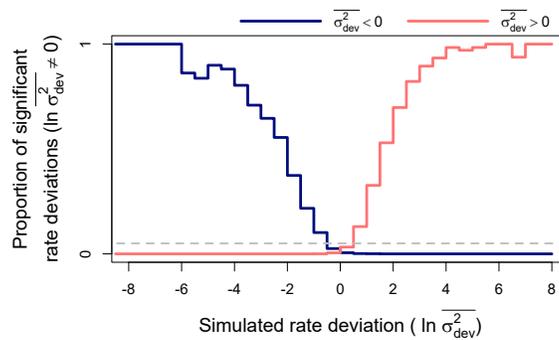


Figure 1.4 Power and error rates for branchwise rate parameters ( $\ln \overline{\sigma^2}$ ). Lines depict changes in proportions of branchwise rates considered anomalously slow (in dark blue) or fast (in light red) as a function of simulated rate deviations ( $\ln \overline{\sigma_{dev}^2}$ ). These results combine all fits to simulated data that detected rate variance ( $\sigma_{\sigma^2}$ ) significantly greater than 0. The proportions are equivalent to power when the detected rate deviation is of the same sign as the true, simulated deviation (left of 0 for anomalously slow rates in dark blue and right for anomalously fast rates in light red), and to error rate when the detected and true rate deviations are of opposite signs. Here, significant rate deviations for simulated rate deviations that are exactly 0 are considered errors regardless of sign.

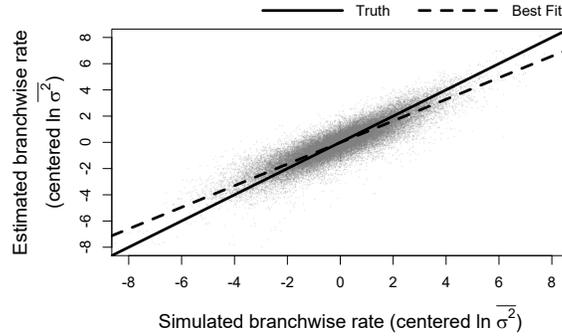


Figure 1.5 Relationship between simulated and estimated branchwise rate parameters ( $\ln \overline{\sigma^2}$ ). For each simulation and posterior sample, branchwise rates were first centered by subtracting their mean. We estimated centered branchwise rates by taking the median of the centered posterior samples. The solid line represents the position of the true centered branchwise rates, while the shallower, dashed line represents the observed line of best fit for these data.

### 1.3.2 Empirical Example

Overall, our model suggests that rates of body size evolution among extant cetaceans have generally slowed down over time, with considerable divergence in rates of body size evolution among key subclades (Fig. 1.6). We found marginally significant support for a decreasing trend in rates over time, with rates declining by about 7% every million years (95% credible interval (CI): 0 - 15% decrease, posterior probability (PP) of increasing trend: 2.5%). We also infer a moderate rate variance of about 0.06 per million years (CI: 0.01 - 0.22, SD ratio: 0.14). Combining these two results, changes in body size evolution rates over a million year time interval are expected to range from a 50% decrease to 60% increase for any particular lineage (Fig. 1.7).

We also identify a few regions of the cetacean phylogeny where rates of body size evolution seem to be especially low or high. After detrending, rates of body size evolution in the beaked whale genus *Mesoplodon* are about 34% slower than the background rate (CI: 13 - 77%, PP of positive rate deviation: < 1%). On the other end of the spectrum, some oceanic dolphin lineages exhibit unusually rapid body size evolution rates. In particular, pilot whales and allies (subfamily globicephalinae) and the orca (*Orcinus orca*) lineage exhibit body size evolution rates about 160% (CI: 10 - 900%, PP: 99%) and 200% (CI: 20 - 1300, PP: 99%) higher than the background rate, respectively. In fact, oceanic dolphins as a whole exhibit a marginally significant increase in body size evolution rates, even after excluding the pilot whale subfamily and orca lineage (CI: 90 - 300%

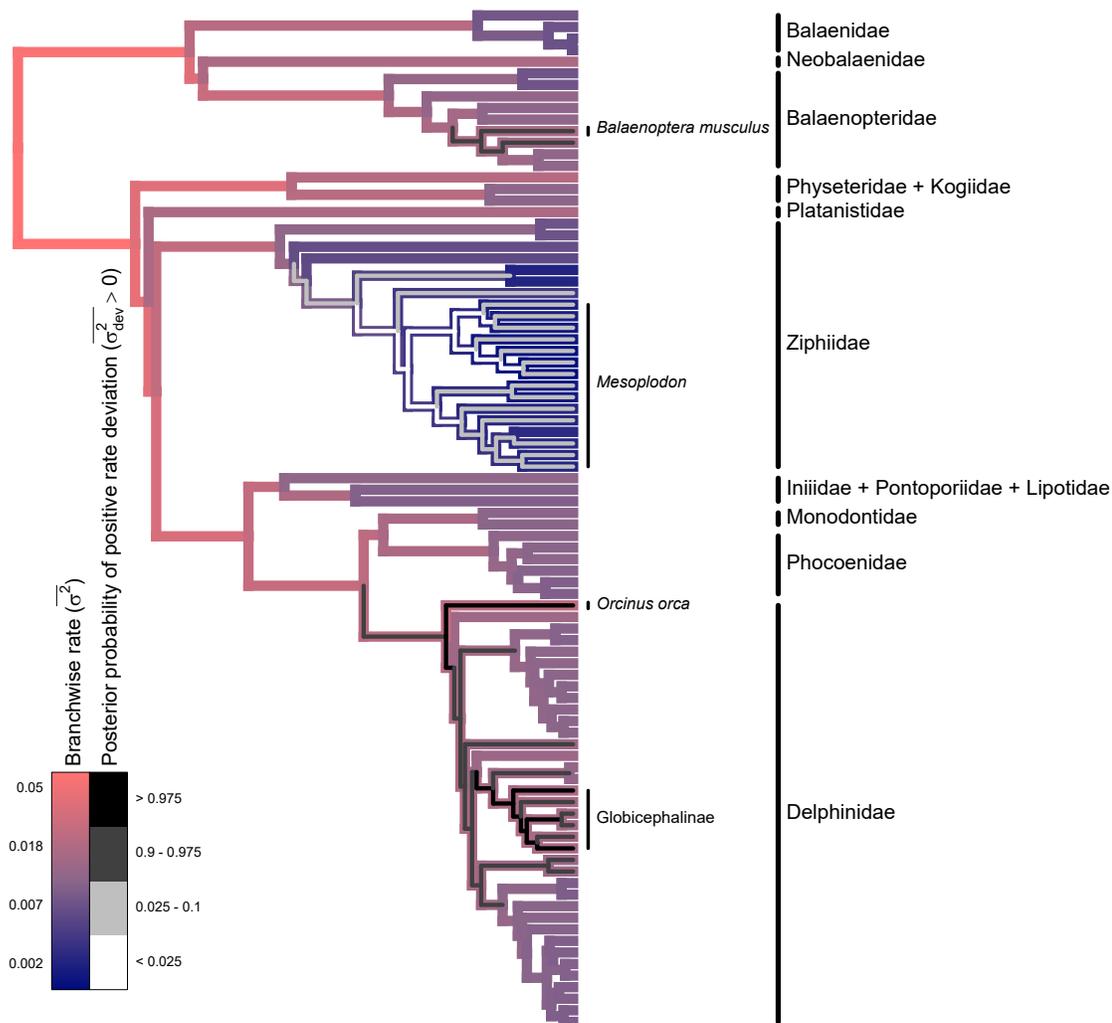


Figure 1.6 Phylogram of model results for cetacean body size data. Branch colors represent median posterior estimates of branchwise rates ( $\ln \sigma^2$ ) of body size evolution, with slower and faster rates in dark blue and light red, respectively. The thinner, inset colors represent the posterior probability that a branchwise rate is anomalously fast according to its rate deviation ( $\ln \sigma_{dev}^2$ ), with lower and higher posterior probabilities in light and dark gray, respectively.

background rate, PP: 95%). Similarly, the blue whale (*Balanoptera musculus*) lineage also exhibits a marginally significant increase in body size evolution rate, about 140% (CI: -10 - 1000%, PP: 95%) higher than the background rate.

Under the model with rate variance constrained to 0, rates of body size evolution decrease by only about 4% every million years (95% CI: -1 - 10% decrease, PP of increasing trend: 7.3%). While only a slight difference, the trend parameter estimated under the full model yields a marginally significant, two-tailed “p-value” of ~5%, while the constrained model yields a decidedly insignificant “p-value” of ~15%. This is reflected in a conventional sample-size corrected Akaike Information Criterion (AICc) comparison between simple BM and EB models of trait evolution fitted via maximum likelihood using the R package *geiger* (version 2.0.7; Pennell et al., 2014). In this case, a simple BM model receives nearly twice the AICc weight of an EB model (65% vs. 35%).

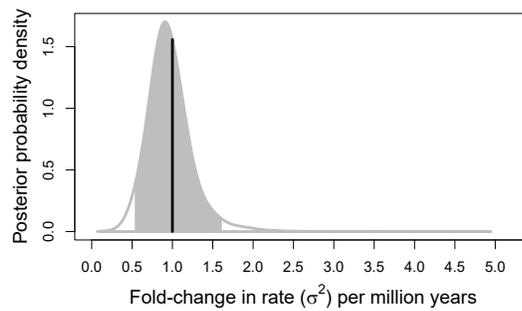


Figure 1.7 The posterior probability distribution of fold-changes in cetacean body size evolution rates ( $\sigma^2$ ) per 1 million years. This distribution is given by  $\exp[\mu_{\sigma^2} + \sigma_{\sigma^2}X]$ , where  $X$  is a random variable drawn from a standard normal distribution. The gray filled-in portion represent the 95% equal-tailed interval, while the vertical line represents the starting rate of 1.

## 1.4 Discussion

Here we implemented a novel data-driven method, *evorates*, for modeling stochastic, incremental variation in trait evolution rates. Part of the power of *evorates* is its ability to infer trait evolution rate variation independent of an *a priori* hypothesis on what factors influence rates. This allows for detailed, hypothesis-free exploration of trait evolution rate variation across time and taxa. Researchers may use such results to generate and refine hypotheses regarding what factors have influenced trait evolution rates across the tree of life (e.g., Uyeda et al., 2018). Overall, *evorates* performs well on simulated data, recovering accurate parameter estimates and exhibiting appropriate statistical power and error rates for hypothesis testing. Further, the method shows great promise for empirical macroevolutionary research, offering novel insights into the dynamics of cetacean body size evolution – a notably well-studied system (e.g., Slater et al., 2010, Pyenson and Sponberg, 2011, Montgomery et al., 2013, Slater and Pennell, 2014, Slater et al., 2017, Sander et al., 2021). The results of our study also build on previous work in demonstrating that estimating time-independent rate heterogeneity is critical for accurately quantifying temporal dynamics in trait evolution rates (Slater and Pennell, 2014). This finding has consequences for how EBs/LBs of trait evolution are practically identified and conceptually defined.

The simulation study results showcases *evorate's* ability to recover accurate parameter estimates across a range of tree sizes. Despite the high uncertainty of rate variance estimates under some trait evolution scenarios, rate heterogeneity could still be correctly detected about 90% of the time with an error rate substantially lower than 5%. Indeed, our hypothesis testing procedures seem conservative in general, exhibiting relatively low error rates. While it could be beneficial to relax significance thresholds for SD ratios and/or posterior probabilities for increased statistical power, our hypothesis testing procedures seem sufficiently powered and we thus do not explore alternative thresholds in great detail here (but see Fig. 1A.2). In any case, compared to conventional EB/LB models, *evorates* can detect decreasing trends in trait evolution rates with greater sensitivity and detect increasing trends with greater robustness. Notably, traits evolving with exponentially increasing rates on an ultrametric phylogeny (i.e., an LB model) exhibit the same probability distri-

bution expected under a single-peak Ornstein-Uhlenbeck (OU) model, where traits evolve towards some optimum at a constant rate (Blomberg et al., 2003). Therefore, the frequently observed support for single-peak OU models from ultrametric comparative data (e.g., Harmon et al., 2010; see also Cooper et al., 2016; Landis and Schraiber, 2017) may partially result from autocorrelated rate heterogeneity, which inflates support for LB/OU models based on our simulation study. Despite their mathematical similarities, LB, OU, and our new models have distinct biological interpretations regarding the importance of rate heterogeneity and selective forces in shaping the patterns of trait diversity within clades.

Interestingly, closer inspection of our simulation study results suggest that, in the presence of rate heterogeneity, models with rate variance constrained to 0 (i.e., conventional EB/LB models) estimate trend parameters corresponding to changes in average trait evolution rates over time. On the other hand, unconstrained *evorates* models estimate trend parameters corresponding to changes in median trait evolution rates over time, essentially determining whether most lineages in a clade exhibit rate decreases or increases (Figs. 1C.3–1C.5; Tables 1C.1–1C.3). Counterintuitively, when the trend parameter is only weakly negative relative to rate variance ( $-\sigma_{\sigma^2}^2/2 < \mu_{\sigma^2} < 0$ ), it is possible for a majority of lineages within a clade to exhibit declining trait evolution rates (i.e., an EB according to *evorates*) while rates averaged across the entire clade increase over time (i.e., an LB according to conventional methods). This occurs because rates evolve in a right-skewed manner under our model—in other words, a few anomalous lineages/subclades tend to evolve extremely high trait evolution rates in spite of declining rates among most other lineages, driving up a clade’s overall average rate (Fig. 1C.1 and 1C.2). We note that *evorates* still returns estimates of average changes in trait evolution rates per unit time via a simple parameter transformation ( $\mu_{\sigma^2} + \sigma_{\sigma^2}^2/2$ ). We choose to focus on the majority-based definition of EBs/LBs since, by accounting for anomalous lineages/subclades exhibiting unusual rates, this definition better matches many macroevolutionary biologists’ intuitive definition of EBs (Lloyd et al., 2012; Slater and Pennell, 2014; Benson et al., 2014; Hopkins and Smith, 2015; Wright, 2017; Puttick, 2018).

Our empirical example with cetacean body size directly demonstrates the practical importance

of these nuances in defining EB/LB dynamics. We find substantial evidence that body size evolution has slowed down in most cetacean lineages, despite the presence of “outlier” lineages exhibiting relatively rapid rates. Indeed, we find little evidence for a decline in body size evolution rates averaged across the clade (95% credible interval: 12% decrease - 5% increase in average rate per million years, posterior probability of increasing average rate: 16%). This broadly agrees with previous research, but *evorates* is able to offer novel insights and contextualize prior results by explicitly estimating branchwise rates in addition to overall trends (Slater and Pennell, 2014; Sander et al., 2021). For example, Slater and Pennell (2014) identified the orca and pilot whale lineages as outlier lineages exhibiting especially rapid rates of body size evolution. Our method recapitulates these findings while suggesting oceanic dolphins as a whole represent a relatively recent burst of body size evolution that has largely masked signals of an earlier burst towards the base of the clade. Such findings more generally agree with recent suggestions that bursts of trait evolution may be common but not limited to the base of “major” clades. This is likely due, in part, to major clades being arbitrarily designated based on taxonomic rank (Puttick, 2018). Alternatively, some propose that EBs may be hierarchical, with major clades exhibiting repeated bouts of rapid trait diversification as competing, closely-related lineages partition niche space more finely over time (Slater and Friscia, 2019). Ultimately, we are optimistic that *evorates* may be better able to resolve how frequently bursts of trait evolution—early or not—occur across the tree of life compared to more conventional methods.

The shrinkage of branchwise rates, whereby rate estimates are biased towards their overall mean, is presumably due to the assumption that rates are autocorrelated under our model. Because of this, rate estimates are partially informed by the rates in closely-related lineages, particularly when closely-related lineages are better sampled (i.e., more related to taxa with sampled trait values and/or consisting of many short branch lengths). This “diffusion” of rates across the phylogeny appears to cause under- and overestimation of unusually high and low rates, respectively. Fortunately, this renders *evorates* conservative in terms of identifying anomalous trait evolution rates, safeguarding against erroneous conclusions. In general, we view this behavior as a good com-

promise between model flexibility and robustness, allowing *evorates* to infer rate variation while avoiding ascribing significance to noise in data. We note that rate variance estimates under our model are largely unbiased, such that branchwise rates in a typical posterior sample should be as variable as the true rates. Thus, taking the joint distribution of branchwise rates into account by analyzing distributions of *differences* between rates, rather than just assessing marginal distributions of rates, appears important in accurately interpreting results under our model. In any case, despite this shrinkage phenomenon, the statistical power to identify overall rate heterogeneity and anomalous rates with *evorates* appears comparable to that of previous data-driven methods (Eastman et al., 2011).

*Evorates* is one of several recently developed methods that also estimate unique trait evolution rates for each branch in a phylogeny but assume an alternative mode of rate change (May and Moore, 2020; Fisher et al., 2021). These other methods assume that branchwise rates are independently distributed according to a log-normal distribution. The method we develop here differs from these “independent rate” (IR) models in assuming that rates evolve gradually and are thus phylogenetically autocorrelated (see also Revell, 2021). Theoretically, trait evolution rates should exhibit some degree of phylogenetic autocorrelation given that many factors hypothesized to affect trait evolution rates themselves exhibit phylogenetic autocorrelation. Indeed, a recent study found evidence for autocorrelation of trait evolution rates in a few vertebrate clades (Sakamoto and Venditti, 2018), and autocorrelation has also been found in lineage diversification (Savolaine et al., 2002; Caron and Pie, 2020) and molecular substitution rates (Lepage et al., 2006; Tao et al., 2019). Notably, there is also no known rate evolution process that would produce independent, log-normally distributed branchwise rates (Lepage et al., 2006, 2007). However, IR models could outperform “autocorrelated rate” (AR) models in some instances due to their tremendous flexibility in modeling how rates vary over time and phylogenies. In general, we expect that IR models will perform best in cases with many traits and/or non-ultrametric trees, where the flexibility of the model can be tempered by rich information content in the data. More work testing for rate autocorrelation or lack thereof in continuous trait data is needed as methods for inferring trait evolution

rate variation become more complex.

Revell (2021) independently developed a method, *multirateBM*, based on a model similar to the one we introduce here, though *evorates* offers several key advantages. In particular, the maximum likelihood (ML) implementation of *multirateBM* renders it impossible to estimate rate variance. To do so, one would need to analytically marginalize over uncertainty in branchwise rates. Here, we circumvent this issue by using Bayesian inference to numerically integrate over uncertainty in branchwise rates. This is analogous to how ML implementations of mixed effect models analytically marginalize over uncertainty in random effects, while Bayesian implementations of the same models sample random effects (Browne and Draper, 2006). Indeed, ML implementations of mixed effect models that treat random effects as parameters would be unable to estimate random effect variances due to the very same reasons *multirateBM* cannot estimate rate variance. Additionally, our model has the added advantage of accommodating both trends in rates and uncertainty in tip trait values. Lastly, we implement procedures to test the significance of rate heterogeneity, trends, and anomalous trait evolution rates. While *multirateBM* offers a quick and convenient means for comparative data exploration, our new method allows for more rigorous quantification and analysis of rate evolutionary processes and patterns from comparative data.

There are a number of ways the *evorates* might be improved or expanded. Assuming that trait evolution rates for different traits are correlated with one another, using data on multiple traits could improve inference of both the rate evolution process and branchwise rate parameters (May and Moore, 2020). Another promising future direction is integration of *evorates* with hypothesis-driven methods. This could be done *post hoc* by applying phylogenetic linear regression to “tip rates” estimated under the model (e.g., Rabosky and Huang, 2016) or analyzing distributions of branchwise rates associated with ancestral states estimated via stochastic character maps (Revell, 2013; but see May and Moore, 2020). Alternatively, one could explicitly model rates as the product of both a stochastic rate evolution process and a deterministic function of some factor of interest. We have already taken steps towards this model extension in our current implementation by allowing rates to change as a deterministic function of time. Lastly, despite our focus on gradually

changing rates, trait evolution rates might also exhibit sudden shifts of large magnitude (“jumps”) or short-lived fluctuations (“pulses”) in response to factors with particularly strong influence on rates. It would be ideal – but difficult – to model rates as evolving gradually, while potentially undergoing sudden jumps or pulses (e.g., Lartillot et al., 2016). An alternative strategy is developing methods to compare the fit of a model like ours against more conventional data-driven models whereby rates jump or even Lévy models whereby rates pulse (Landis et al., 2013). Assessing when and whether comparative data can distinguish between different modes of rate change will be important for future research on the dynamics of trait evolution.

#### **1.4.1 Conclusion**

Here, we introduced *evorates*, a method that models gradual change, rather than abrupt shifts, in continuous trait evolution rates from comparative data. Unlike nearly all other comparative methods for inferring rate variation, *evorates* goes beyond identifying lineages exhibiting anomalous rates by also estimating the process by which rates themselves evolve. Although there are many potential modes of rate variation over time and phylogenies, our model estimates rate evolution processes as the product of two parameters: one controlling how quickly rates accumulate random variation, and another determining whether rates tend to decrease or increase over time. The resulting method returns accurate estimates of evolutionary processes and provides a flexible and intuitive means of detecting and analyzing trait evolution rate variation. Looking forward, *evorates* has tremendous potential for improvement and elaboration, and we are optimistic that the future of macroevolutionary biology will benefit from increased focus not only on how traits evolve, but how the rates of trait evolution themselves evolve over time and taxa.

## BIBLIOGRAPHY

- Arnold P.W. and Heinsohn G.E. 1996. Phylogenetic status of the Irrawaddy dolphin *Orcaella brevirostris* (Owen in Gray): A cladistic analysis. *Mem Queensl Mus* 39:143–204.
- Arnold S.J., Bürger R., Hohenlohe P.A., Ajie B.C., and Jones A.G. 2008. Understanding the evolution and stability of the G-matrix. *Evolution* 62:2451–2461.
- Baker A.N. 1981. The southern right whale dolphin *Lissodelphis peronii* (Lacépède) in Australasian waters. *Natl Mus NZ Rec* 2:17–34.
- Barros N.B. 1991. Recent cetacean records for southeastern Brazil. *Mar Mamm Sci* 7:296–306.
- Beaulieu J.M. and O’Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst Biol* 65:583–601.
- Benson R.B.J., Campione N.E., Carrano M.T., Mannion P.D., Sullivan C., Upchurch P., and Evans D.C. 2014. Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLoS Biol* 12:e1001853.
- Betancourt M. and Girolami M. 2019. Hamiltonian Monte Carlo for hierarchical models. Pages 79–97 in *Current Trends in Bayesian Methodology with Applications* (S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan, eds.). Chapman and Hall/CRC Press, Boca Raton, FL.
- Beygelzimer A., Kakadet S., Langford J., Arya S., Mount D., and Li S. 2022. FNN: Fast nearest neighbor search algorithms and applications. R package version 1.1.3.1.
- Blomberg S.P., Garland T. Jr, and Ives A.R. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57:717–745.
- Borstein S.R., Fordyce J.A., O’Meara B.C., Wainwright P.C., and McGee M.D. 2019. Reef fish functional traits evolve fastest at trophic extremes. *Nat Ecol Evol* 3:191–199.
- Browne W.J. and Draper D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* 1:473–514.
- Brusatte S.L., Butler R.J., Prieto-Márquez A., and Norell M.A. 2012. Dinosaur morphological diversity and the end-Cretaceous extinction. *Nat Commun* 3:804.
- Caetano D.S. and Harmon L.J. 2019. Estimating correlated rates of trait evolution with uncertainty. *Syst Biol* 68:412–429.
- Caron F.S. and Pie M.R. 2020. The phylogenetic signal of diversification rates. *J Zool Syst Evol Res* 58:1432–1436.

- Carpenter B., Gelman A., Hoffman M.D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., and Riddell A. 2017. Stan: A probabilistic programming language. *J Stat Softw* 76:1–32.
- Charlton-Robb K., Gershwin L.A., Thompson R., Austin J., Owen K., and McKechnie S. 2011. A new dolphin species, the Burrunan dolphin *Tursiops australis* sp. nov., endemic to southern Australian coastal waters. *PLoS One* 6:e24047.
- Chartier M., von Balthazar M., Sontag S., Löfstrand S., Palme T., Jabbour F., Sauquet H., and Schönenberger J. 2021. Global patterns and a latitudinal gradient of flower disparity: Perspectives from the angiosperm order Ericales. *New Phytol* 230:821–831.
- Chira A.M., Cooney C.R., Bright J.A., Capp E.J.R., Hughes E.C., Moody C.J.A., Nouri L.O., Varley Z.K., and Thomas G.H. 2018. Correlates of rate heterogeneity in avian ecomorphological traits. *Ecol Lett* 21:1505–1514.
- Chira A.M. and Thomas G.H. 2016. The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. *J Evol Biol* 29:2502–2518.
- Clavel J. and Morlon H. 2017. Accelerated body size evolution during cold climatic periods in the Cenozoic. *Proc Natl Acad Sci USA* 114:4183–4188.
- Constantine R., Iwata T., Nieukirk S.L., and Penry G.S. 2018. Future directions in research on Bryde's whales. *Front Mar Sci* 5:1–7.
- Cooper N. and Purvis A. 2009. What factors shape rates of phenotypic evolution? A comparative study of cranial morphology of four mammalian clades. *J Evol Biol* 22:1024–1035.
- Cooper N., Thomas G.H., Venditti C., Meade A., and Freckleton R.P. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol J Linn Soc* 118:64–77.
- Dalebout M.L., Mead J.G., Baker C.S., Baker A.N., and Helden A.L. 2002. A new species of beaked whale *Mesoplodon perrini* sp. n. (Cetacea: Ziphiidae) discovered through phylogenetic analyses of mitochondrial DNA sequences. *Mar Mamm Sci* 18:577–608.
- Dalebout M.L., Scott Baker C., Steel D., Thompson K., Robertson K.M., Chivers S.J., Perrin W.F., Goonatilake M., Charles Anderson R., Mead J.G., Potter C.W., Thompson L., Jupiter D., and Yamada T.K. 2014. Resurrection of *Mesoplodon hotaula* Deraniyagala 1963: A new species of beaked whale in the tropical Indo-Pacific. *Mar Mamm Sci* 30:1081–1108.
- Devreese J.P.A., Lemmens D., and Tempere J. 2010. Path integral approach to Asian options in the Black-Scholes model. *Physica A: Statistical Mechanics and its Applications* 389:780–788.
- Donoghue M.J. and Sanderson M.J. 2015. Confluence, synnovation, and depauperons in plant diversification. *New Phytol* 207:260–274.

- Drury J.P., Clavel J., Tobias J.A., Rolland J., Sheard C., and Morlon H. 2021. Tempo and mode of morphological evolution are decoupled from latitude in birds. *PLoS Biol* 19:e3001270.
- Dufresne D. 2004. The log-normal approximation in financial and other computations. *Adv Appl Probab* 36:747–773.
- Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., and Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65:3578–3589.
- Fabre A.C., Bardua C., Bon M., Clavel J., Felice R.N., Streicher J.W., Bonnel J., Stanley E.L., Blackburn D.C., and Goswami A. 2020. Metamorphosis shapes cranial diversity and rate of evolution in salamanders. *Nat Ecol Evol* 4:1129–1140.
- Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet* 25:471–492.
- Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *Am Nat* 171:713–725.
- Fisher A.A., Ji X., Zhang Z., Lemey P., and Suchard M.A. 2021. Relaxed random walks at scale. *Syst Biol* 70:258–267.
- Fortune S.M.E., Moore M.J., Perryman W.L., and Trites A.W. 2021. Body growth of North Atlantic right whales (*Eubalaena glacialis*) revisited. *Mar Mamm Sci* 37:433–447.
- Freckleton R.P. 2012. Fast likelihood calculations for comparative analyses. *Methods Ecol Evol* 3:940–947.
- Gill M.S., Tung Ho L.S., Baele G., Lemey P., and Suchard M.A. 2017. A relaxed directional random walk model for phylogenetic trait evolution. *Syst Biol* 66:299–319.
- Gingerich P.D. 2009. Rates of evolution. *Annu Rev Ecol Evol Syst* 40:657–675.
- Goolsby E.W. 2017. Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. *Ecol Evol* 7:2791–2797.
- Hansen T.F., Bolstad G.H., and Tsuboi M. 2022. Analyzing disparity and rates of morphological evolution with model-based phylogenetic comparative methods. *Syst Biol* 71:1054–1072.
- Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeck M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte J.A. II, Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., and Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64:2385–2396.

- Hassler G., Tolkoﬀ M.R., Allen W.L., Ho L.S.T., Lemey P., and Suchard M.A. 2022. Inferring phenotypic trait evolution on large trees with many incomplete measurements. *J Am Stat Assoc* 117:678–692.
- Held L. and Ott M. 2018. On p-values and Bayes factors. *Annu Rev Stat Appl* 5:393–419.
- Hoffman M.D. and Gelman A. 2014. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 15:1593–1623.
- Hopkins M.J. and Smith A.B. 2015. Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. *Proc Natl Acad Sci USA* 112:3758–3763.
- Hunt G. 2006. Fitting and comparing models of phyletic evolution: Random walks and beyond. *Paleobiology* 32:578–601.
- Jefferson T.A. and Rosenbaum H.C. 2014. Taxonomic revision of the humpback dolphins (*Sousa spp.*), and description of a new species from Australia. *Mar Mamm Sci* 30:1494–1541.
- Kass R.E. and Raftery A.E. 1995. Bayes factors. *J Am Stat Assoc* 90:773–795.
- Konishi K., Tamura T., Zenitani R., Bando T., Kato H., and Walløe L. 2008. Decline in energy storage in the Antarctic minke whale (*Balaenoptera bonaerensis*) in the Southern Ocean. *Polar Biol* 31:1509–1520.
- Landis M.J. and Schraiber J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes. *Proc Natl Acad Sci USA* 114:13224–13229.
- Landis M.J., Schraiber J.G., and Liang M. 2013. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Syst Biol* 62:193–204.
- Lartillot N., Phillips M.J., and Ronquist F. 2016. A mixed relaxed clock model. *Philos Trans R Soc B* 371:20150132.
- Lartillot N. and Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol* 28:729–744.
- Lemoine N.P. 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos* 128:912–928.
- Lepage T., Bryant D., Philippe H., and Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol* 24:2669–2680.
- Lepage T., Lawi S., Tupper P., and Bryant D. 2006. Continuous and tractable models for the variation of evolutionary rates. *Math Biosci* 199:216–233.

- Limpert E., Stahel W.A., and Abbt M. 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51:341–352.
- Lloyd G.T. and Slater G.J. 2021. A total-group phylogenetic metatree for Cetacea and the importance of fossil data in diversification analyses. *Syst Biol* 70:922–939.
- Lloyd G.T., Wang S.C., and Brusatte S.L. 2012. Identifying heterogeneity in rates of morphological evolution: discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). *Evolution* 66:330–348.
- Lodi L., Sicilian S., and Capistran L. 1990. Mass stranding of *Peponocephala electra* (Cetacea, Globicephalinae) on Piracange Beach, Baria, northeastern Brazil. *Sci Rep Cetacean Res* 1:79–84.
- May M.R. and Moore B.R. 2020. A Bayesian approach for inferring the impact of a discrete character on rates of continuous-character evolution in the presence of background-rate variation. *Syst Biol* 69:530–544.
- Mead J.G., Walker W.A., and Houck W.J. 1982. Biological observations on *Mesoplodon carlhubbsi* (Cetacea, Ziphiidae). *Smithson Contrib Zool* Pages 1–25.
- Mihalitsis M. and Bellwood D.R. 2019. Morphological and functional diversity of piscivorous fishes on coral reefs. *Coral Reefs* 38:945–954.
- Molina D.M. and Oporto J.A. 1993. Comparative study of dentine staining techniques to estimate age in the Chilean dolphin, *Cephalorhynchus eutropia* (Gray, 1846). *Aquat Mamm* 19:45–48.
- Montgomery S.H., Geisler J.H., McGowen M.R., Fox C., Marino L., and Gatesy J. 2013. The evolutionary history of cetacean brain and body size. *Evolution* 67:3339–3353.
- Muñoz M.M. and Bodensteiner B.L. 2019. Janzen’s hypothesis meets the Bogert effect: Connecting climate variation, thermoregulatory behavior, and rates of physiological evolution. *Integr Org Biol* 1:oby002.
- Muñoz M.M., Hu Y., Anderson P.S.L., and Patek S.N. 2018. Strong biomechanical relationships bias the tempo and mode of morphological evolution. *Elife* 7:e37621.
- Neal R. 2011. MCMC using Hamiltonian dynamics. Pages 113–162 *in* Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. L. Jones, and M. Xiao-Li, eds.). Chapman and Hall/CRC, Boca Raton, FL.
- Pagel M., O’Donovan C., and Meade A. 2022. General statistical model shows that macroevolutionary patterns and processes are consistent with Darwinian gradualism. *Nat Commun* 13:1113.
- Paradis E. and Schliep K. 2019. ape 5.0: An environment for modern phylogenetics and evolution-

- ary analyses in R. *Bioinformatics* 35:526–528.
- Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., and Harmon L.J. 2014. *geiger* v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30:2216–2218.
- Pennell M.W., FitzJohn R.G., Cornwell W.K., and Harmon L.J. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *Am Nat* 186:E33–50.
- Plön S., Albrecht K.H., Cliff G., and Froneman P.W. 2012. Organ weights of three dolphin species (*Sousa chinensis*, *Tursiops aduncus* and *Delphinus capensis*) from South Africa: Implications for ecological adaptation? *J Cetacean Res Manag* 12:265–276.
- Puttick M.N. 2018. Mixed evidence for early bursts of morphological evolution in extant clades. *J Evol Biol* 31:502–515.
- Pyenson N.D. and Sponberg S.N. 2011. Reconstructing body size in extinct crown Cetacea (Neoceti) using allometry, phylogenetic methods and tests from the fossil record. *J Mamm Evol* 18:269.
- Rabosky D.L., Donnellan S.C., Grundler M., and Lovette I.J. 2014. Analysis and visualization of complex macroevolutionary dynamics: An example from Australian scincid lizards. *Syst Biol* 63:610–627.
- Rabosky D.L. and Goldberg E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst Biol* 64:340–355.
- Rabosky D.L. and Huang H. 2016. A robust semi-parametric test for detecting trait-dependent diversification. *Syst Biol* 65:181–193.
- Raj Pant S., Goswami A., and Finarelli J.A. 2014. Complex body size trends in the evolution of sloths (*Xenarthra*: *Pilosa*). *BMC Evol Biol* 14:184.
- Reaney A.M., Bouchenak-Khelladi Y., Tobias J.A., and Abzhanov A. 2020. Ecological and morphological determinants of evolutionary diversification in Darwin’s finches and their relatives. *Ecol Evol* 10:14020–14032.
- Revell L.J. 2012. *phytools*: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223.
- Revell L.J. 2013. A comment on the use of stochastic character maps to estimate evolutionary rate variation in a continuously valued trait. *Syst Biol* 62:339–345.
- Revell L.J. 2021. A variable-rate quantitative trait evolution model using penalized-likelihood. *PeerJ* 9:e11997.

- Reyes J.C., Mead J.G., and Van Waerebeek K. 1991. A new species of beaked whale *Mesoplodon peruvianus* sp. n. (Cetacea: Ziphiidae) from Peru. *Mar Mamm Sci* 7:1–24.
- Safak A. and Safak M. 2002. Moments of the sum of correlated log-normal random variables. Pages 140–144 in *Proceedings of IEEE Vehicular Technology Conference (VTC) vol. 1 IEEE*.
- Sakamoto M. and Venditti C. 2018. Phylogenetic non-independence in rates of trait evolution. *Biol Lett* 14:20180502.
- Sander P.M., Griebeler E.M., Klein N., Juarbe J.V., Wintrich T., Revell L.J., and Schmitz L. 2021. Early giant reveals faster evolution of large body size in ichthyosaurs than in cetaceans. *Science* 374:eabf5787.
- Savolaine V., Heard S.B., Powell M.P., Jonathan Davies T., and Mooers A.Ø. 2002. Is cladogenesis heritable? *Sys Biol* 51:835–843.
- Simpson G.G. 1944. *Tempo and Mode in Evolution*. Columbia University Press, New York, NY.
- Skeels A. and Cardillo M. 2019. Equilibrium and non-equilibrium phases in the radiation of *Hakea* and the drivers of diversity in Mediterranean-type ecosystems. *Evolution* 73:1392–1410.
- Slater G.J. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. *Proc Natl Acad Sci USA* 112:4897–4902.
- Slater G.J. and Friscia A.R. 2019. Hierarchy in adaptive radiation: A case study using the Carnivora (Mammalia). *Evolution* 73:524–539.
- Slater G.J., Goldbogen J.A., and Pyenson N.D. 2017. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proc R Soc B* 284:20170546.
- Slater G.J. and Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Syst Biol* 63:293–308.
- Slater G.J., Price S.A., Santini F., and Alfaro M.E. 2010. Diversity versus disparity and the radiation of modern cetaceans. *Proc R Soc B* 277:3097–3104.
- Sookias R.B., Butler R.J., and Benson R.B.J. 2012. Rise of dinosaurs reveals major body-size transitions are driven by passive processes of trait evolution. *Proc R Soc B* 279:2180–2187.
- Stan Development Team . 2019. *Stan Modeling Language Users Guide and Reference Manual. Version 2.21.0*.
- Stan Development Team . 2020. *RStan: The R interface to Stan. R package version 2.21.1*.
- Stone C.J., Hansen M.H., Kooperberg C., and Truong Y.K. 1997. Polynomial splines and their

- tensor products in extended linear modeling. *Ann Stat* 25:1371–1425.
- Tao Q., Tamura K., U Battistuzzi F., and Kumar S. 2019. A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. *Mol Biol Evol* 36:811–824.
- Thomas G.H. and Freckleton R.P. 2012. MOTMOT: Models of trait macroevolution on trees. *Methods Ecol Evol* 3:145–151.
- Thompson K., Baker C.S., van Helden A., Patel S., Millar C., and Constantine R. 2012. The world's rarest whale. *Curr Biol* 22:R905–R906.
- Thorne J.L., Kishino H., and Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15:1647–1657.
- Uyeda J.C., Caetano D.S., and Pennell M.W. 2015. Comparative analysis of principal components can be misleading. *Syst Biol* 64:677–689.
- Uyeda J.C., Zenil-Ferguson R., and Pennell M.W. 2018. Rethinking phylogenetic comparative methods. *Syst Biol* 67:1091–1109.
- Vehtari A., Gelman A., Simpson D., Carpenter B., and Bürkner P.C. 2021. Rank-normalization, folding, and localization: An improved  $\hat{R}$  for assessing convergence of MCMC (with discussion). *Bayesian Anal* 16:667–718.
- Villar D., Flicek P., and Odom D.T. 2014. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat Rev Genet* 15:221–233.
- Wagenmakers E.J., Gronau Q.F., Dablander F., and Etz A. 2022. The support interval. *Erkenntnis* 87:589–601.
- Wagenmakers E.J., Lodewyckx T., Kuriyal H., and Grasman R. 2010. Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn Psychol* 60:158–189.
- Weber M.G., Mitko L., Eltz T., and Ramírez S.R. 2016. Macroevolution of perfume signalling in orchid bees. *Ecol Lett* 19:1314–1323.
- Welch J.J. and Waxman D. 2008. Calculating independent contrasts for the comparative study of substitution rates. *J Theor Biol* 251:667–678.
- Wright D.F. 2017. Phenotypic innovation and adaptive constraints in the evolutionary radiation of Palaeozoic crinoids. *Sci Rep* 7:13745.
- Zhao P. and Lai L. 2021. On the convergence rates of KNN density estimation. Pages 2840–2845 in 2021 IEEE International Symposium on Information Theory (ISIT) IEEE.

## APPENDIX 1A

### SUPPLEMENTAL TABLES AND FIGURES

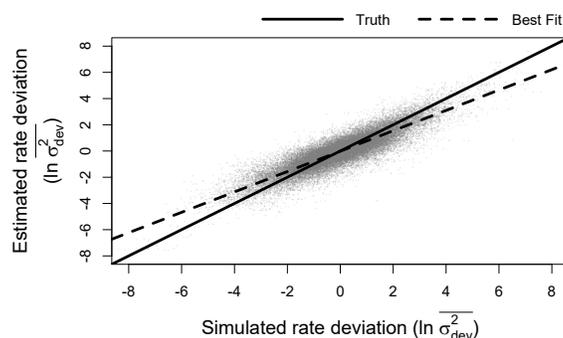


Figure 1A.1 Relationship between simulated and estimated branchwise rate deviation parameters ( $\ln \overline{\sigma_{dev}^2}$ ). The solid line represents the position of the true branchwise rate deviations, while the shallower, dashed line represents the observed line of best fit for these data.

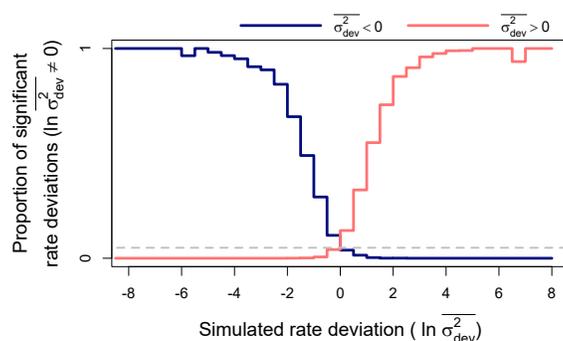


Figure 1A.2 Power and error rates for branchwise rate parameters ( $\ln \overline{\sigma^2}$ ) under relaxed significance thresholds (posterior probability  $< 0.1$  or  $> 0.9$ ). Lines depict changes in proportions of branchwise rates considered anomalously slow (in dark blue) or fast (in light red) as a function of simulated rate deviations ( $\ln \overline{\sigma_{dev}^2}$ ). These results combine all fits to simulated data that detected rate variance ( $\sigma_{\sigma^2}$ ) significantly greater than 0. The proportions are equivalent to power when the detected rate deviation is of the same sign as the true, simulated deviation (left of 0 for anomalously slow rates in dark blue and right for anomalously fast rates in light red), and to error rate when the detected and true rate deviations are of opposite signs. Here, significant rate deviations for simulated rate deviations that are exactly 0 are considered errors regardless of sign.

Table 1A.1 Cetacean body length data and associated references used for empirical example.

species	length (m)	reference
<i>Balaena mysticetus</i>	18.0	Slater et al., 2010
<i>Balaenoptera acutorostrata</i>	10.7	Slater et al., 2010
<i>Balaenoptera bonaerensis</i>	10.2	Konishi et al., 2008
<i>Balaenoptera borealis</i>	16.1	Slater et al., 2010
<i>Balaenoptera edeni</i>	15.4	Slater et al., 2010
<i>Balaenoptera musculus</i>	33.6	Slater et al., 2010
<i>Balaenoptera omurai</i>	10.7	Slater et al., 2010
<i>Balaenoptera physalus</i>	21.2	Slater et al., 2010
<i>Berardius arnuxii</i>	8.9	Slater et al., 2010
<i>Berardius bairdii</i>	12.0	Slater et al., 2010
<i>Caperea marginata</i>	6.2	Slater et al., 2010
<i>Cephalorhynchus commersoni</i>	1.5	Slater et al., 2010
<i>Cephalorhynchus eutropia</i>	1.5	Molina and Oporto, 1993
<i>Cephalorhynchus heavisidii</i>	1.7	Slater et al., 2010
<i>Cephalorhynchus hectori</i>	1.5	Slater et al., 2010
<i>Delphinapterus leucas</i>	3.8	Slater et al., 2010
<i>Delphinus capensis</i>	2.5	Plön et al., 2012
<i>Delphinus delphis</i>	2.3	Slater et al., 2010
<i>Eschrichtius robustus</i>	14.6	Slater et al., 2010
<i>Eubalaena australis</i>	13.9	Slater et al., 2010
<i>Eubalaena glacialis</i>	13.7	Slater et al., 2010
<i>Eubalaena japonica</i>	17.4	Fortune et al., 2021
<i>Feresa attenuata</i>	2.4	Slater et al., 2010
<i>Globicephala macrorhynchus</i>	4.8	Slater et al., 2010
<i>Globicephala melas</i>	5.1	Slater et al., 2010
<i>Grampus griseus</i>	3.7	Slater et al., 2010
<i>Hyperoodon ampullatus</i>	7.9	Slater et al., 2010
<i>Hyperoodon planifrons</i>	7.5	Slater et al., 2010
<i>Indopacetus pacificus</i>	7.2	Slater et al., 2010
<i>Inia geoffrensis</i>	2.0	Slater et al., 2010
<i>Kogia breviceps</i>	3.4	Slater et al., 2010
<i>Kogia sima</i>	2.4	Slater et al., 2010
<i>Lagenodelphis hosei</i>	2.6	Slater et al., 2010
<i>Lagenorhynchus albirostris</i>	3.0	Slater et al., 2010
<i>Leucopleurus acutus</i>	2.4	Slater et al., 2010
<i>Lipotes vexillifer</i>	2.0	Slater et al., 2010
<i>Lissodelphis borealis</i>	2.3	Slater et al., 2010
<i>Lissodelphis peronii</i>	2.3	Baker, 1981
<i>Megaptera novaeangliae</i>	18.0	Slater et al., 2010
<i>Mesoplodon bidens</i>	5.1	Slater et al., 2010

Table 1A.1 (cont'd)

species	length (m)	reference
<i>Mesoplodon bowdoini</i>	4.5	Slater et al., 2010
<i>Mesoplodon carlhubbsi</i>	5.3	Mead et al., 1982
<i>Mesoplodon densirostris</i>	4.7	Slater et al., 2010
<i>Mesoplodon europaeus</i>	5.2	Slater et al., 2010
<i>Mesoplodon ginkgodens</i>	4.9	Slater et al., 2010
<i>Mesoplodon grayi</i>	5.3	Slater et al., 2010
<i>Mesoplodon hectori</i>	4.4	Slater et al., 2010
<i>Mesoplodon hotaula</i>	4.8	Dalebout et al., 2014
<i>Mesoplodon layardii</i>	6.2	Slater et al., 2010
<i>Mesoplodon mirus</i>	5.1	Slater et al., 2010
<i>Mesoplodon perrini</i>	4.4	Dalebout et al., 2002
<i>Mesoplodon peruvianus</i>	3.7 <sup>a</sup>	Reyes et al., 1991
<i>Mesoplodon stejnegeri</i>	5.7	Slater et al., 2010
<i>Mesoplodon traversii</i>	5.3	Thompson et al., 2012
<i>Monodon monoceros</i>	4.3	Slater et al., 2010
<i>Neophocaena phocaenoides</i>	1.4	Slater et al., 2010
<i>Orcaella brevirostris</i>	2.2	Slater et al., 2010
<i>Orcaella heinsohni</i>	2.2	Arnold and Heinsohn, 1996
<i>Orcinus orca</i>	7.9	Slater et al., 2010
<i>Peponocephala electra</i>	2.8	Lodi et al., 1990
<i>Phocoena dioptrica</i>	1.9	Slater et al., 2010
<i>Phocoena phocoena</i>	1.9	Slater et al., 2010
<i>Phocoena sinus</i>	1.1	Slater et al., 2010
<i>Phocoena spinipinnis</i>	1.7	Slater et al., 2010
<i>Phocoenoides dalli</i>	1.9	Slater et al., 2010
<i>Physeter macrocephalus</i>	11.0	Slater et al., 2010
<i>Platanista gangetica</i>	2.5	Slater et al., 2010
<i>Pontoporia blainvillii</i>	1.5	Slater et al., 2010
<i>Pseudorca crassidens</i>	5.1	Slater et al., 2010
<i>Sagmatias australis</i>	2.1	Slater et al., 2010
<i>Sagmatias cruciger</i>	1.8	Slater et al., 2010
<i>Sagmatias obliquidens</i>	2.4	Slater et al., 2010
<i>Sagmatias obscurus</i>	1.9	Slater et al., 2010
<i>Sotalia fluviatilis</i>	1.5	Slater et al., 2010
<i>Sotalia guianensis</i>	2.1	Barros, 1991
<i>Sousa chinensis</i>	2.4	Slater et al., 2010
<i>Sousa teuszii</i>	2.5	Jefferson and Rosenbaum, 2014
<i>Stenella attenuata</i>	2.1	Slater et al., 2010
<i>Stenella clymene</i>	1.9	Slater et al., 2010
<i>Stenella coeruleoalba</i>	2.3	Slater et al., 2010

Table 1A.1 (cont'd)

species	length (m)	reference
<i>Stenella frontalis</i>	2.1	Slater et al., 2010
<i>Stenella longirostris</i>	2.0	Slater et al., 2010
<i>Steno bredanensis</i>	2.6	Slater et al., 2010
<i>Tasmacetus shepherdi</i>	6.5	Slater et al., 2010
<i>Tursiops aduncus</i>	2.1	Slater et al., 2010
<i>Tursiops australis</i>	2.8 <sup>b</sup>	Charlton-Robb et al., 2011
<i>Tursiops truncatus</i>	2.4	Slater et al., 2010
<i>Ziphius cavirostris</i>	6.4	Slater et al., 2010

<sup>a</sup>from male specimen because no mature females were measured

<sup>b</sup>sex not reported

## APPENDIX 1B

### APPROXIMATING GEOMETRIC BROWNIAN MOTION TIME-AVERAGES

Our model seeks to model rates ( $\sigma^2$ ) as “evolving” under a trended Geometric Brownian motion (GBM)-like process, whereby the natural log of rates evolve in a trended Brownian motion (BM)-like manner. Unfortunately, this requires an expression for the probability distribution of GBM time-averages along each branch in the phylogeny. Expressions for such distributions are infamously intractable, necessitating approximate solutions (Dufresne, 2004; Lepage et al., 2007). For our model, we use a multivariate log-normal approximation to model rate time-averages along each branch (branchwise averages,  $\bar{\sigma}^2$ ) based on two observations. First, as the rate variance parameter ( $\sigma_{\sigma^2}^2$ ) approaches 0, rates ( $\sigma^2$ ) will converge to following a simple exponential function with respect to time,  $\sigma^2 = \sigma_0^2 \exp[\mu_{\sigma^2} t]$ , where  $\sigma_0^2$  is the starting rate,  $\mu_{\sigma^2}$  is the trend, and  $t$  is time. In this case, the branchwise averages can be derived through integration and are equivalent to the time-averaged rates expected under a conventional “early/late burst” (EB/LB) model (Blomberg et al., 2003). Second, over short amounts of time and/or with low rate variance, the arithmetic and geometric time-averages of a GBM process approach one another. The geometric time-average of a GBM process is simply the exponentiated arithmetic time-average of the GBM process on the natural log scale, which has a straight-forward and tractable log-normal distribution (Devreese et al., 2010). Thus, assuming that branch lengths in a phylogeny are typically short and rate variance is relatively low, we can approximate the distribution of the natural log of branchwise averages by adding multivariate normal “noise”,  $\gamma$ , to the natural log of branchwise averages expected under a conventional EB/LB model,  $\beta$ . In other words:

$$\ln(\bar{\sigma}^2) \approx \beta + \gamma \tag{1}$$

$$\beta = \ln(\sigma_0^2) + \begin{cases} 0 & \text{if } \mu_{\sigma^2} = 0 \\ \ln(|\exp[\mu_{\sigma^2} \tau_2] - \exp[\mu_{\sigma^2} \tau_1]|) - \ln(|\mu_{\sigma^2}|) - \ln(t) & \text{if } \mu_{\sigma^2} \neq 0 \end{cases} \tag{2}$$

$$\gamma \sim MVN(0, \sigma_{\sigma^2}^2 D) \tag{3}$$

as in the main text. Here,  $t$  is a vector of branch lengths,  $\tau_1$  and  $\tau_2$  are vectors of the start and end times of each branch (i.e.,  $\tau_2 - \tau_1 = t$ ), and  $D$  is the variance-covariance matrix of branchwise averages for a value evolving under an untrended BM process on a phylogeny. Let  $\bar{x}$  and  $t$  be vectors of time-averaged trait values and edge lengths, respectively, for three edges: two sister edges,  $i$  and  $j$ , with ancestral edge,  $k$ . If traits evolve under an untrended BM process and the ancestral trait value of  $k$  is fixed, the variances of  $\bar{x}_i$  and  $\bar{x}_j$  are  $t_i/3 + t_k$  and  $t_j/3 + t_k$ , respectively. The covariance between  $\bar{x}_i$  and  $\bar{x}_j$  is simply  $t_k$ , and the covariances between either  $\bar{x}_i$  or  $\bar{x}_j$  and  $\bar{x}_k$  is  $t_k/2$  (Devreese et al., 2010). From this, we can derive an expression for the variance-covariance matrix of branchwise averages given an arbitrary phylogeny, as shown in the main text:

$$D_{i,j} = \sum_{k \in \text{anc}(i,j)} t_k \begin{cases} 2t_i/3 & \text{if } i = j \\ t_i/2 & \text{if } i \in \text{anc}(j,j) \\ t_j/2 & \text{if } j \in \text{anc}(i,i) \\ 0 & \text{if } i \neq j, i \notin \text{anc}(j,j), j \notin \text{anc}(i,i) \end{cases} \quad (4)$$

While this multivariate log-normal approximation is rough, we demonstrate here that it is largely sufficient for our purposes. Notably, we are not the first to approximate GBM time-averages using log-normal distributions in the context of comparative phylogenetics (Welch and Waxman, 2008). There are two other tractable strategies for approximating these distributions given in the comparative phylogenetics literature. Both of these strategies use the fact that values at the nodes of a phylogeny evolving under a GBM process follow an exact multivariate log-normal distribution, and instead focus on estimating nodewise values. Branchwise averages are then approximated by either averaging ancestral and descendant nodewise values for each edge (e.g., Thorne et al., 1998) or via the maximum likelihood estimate of branchwise averages given the ancestral and descendant nodewise values (e.g., Lartillot and Poujol, 2011; Revell, 2021). We term these strategies “end-point averaging” and “endpoint integration”, respectively. We prefer the log-normal approximation due to its convenient formulation and direct focus on estimating branchwise, rather than nodewise, quantities. In the spirit of thoroughness, however, we conducted three simulation experiments to

investigate the relative performance of these different approximation strategies.

We first conducted a simple experiment where we simulated 100,000 GBM time-averages on the natural log scale under each approximation strategy. We also estimated a “true” branchwise average distribution for comparison by simulating 100,000 fine-grained GBM sample paths (1,000 time points) and taking the natural log of each sample path’s average. We repeated these simulations for each combination of trend ( $\mu_{\sigma^2}$ ) and rate variance ( $\sigma_{\sigma^2}^2$ ) parameter values used in the main text’s simulation study (Fig. 1B.1). All simulations were standardized to occur over a time interval of 1, just as each phylogeny in our simulation study was rescaled to have a total height of 1. The results below thus represent how “off” each approximation would be for a single branch spanning the entire height of a phylogeny in our simulation study. The log-normal approximation notably lacks a right skew characteristic of the true distribution and other approximations. The log-normal approximation also appears to overestimate the variance of branchwise averages when trends are decreasing and underestimates variance when trends are increasing, particularly with high rate variance. On the other hand, the endpoint average approximation exhibits notable upward bias and consistently underestimates branchwise average variance. Additionally, this approximation fails to converge to the correct branchwise average when rate variance is 0. Lastly, the endpoint integration approximation exhibits no notable bias but underestimates branchwise average variance in the case of no or decreasing trends. The accuracy of branchwise average variance under the log-normal approximation might be improved by adapting the Fenton-Wilkinson approximation of log-normal sums for GBM processes (Safak and Safak, 2002), but we did not explore this here.

The above results help give a sense of where each approximation breaks down in parameter space, yet poorly represent the practical behavior of each approximation. In the context of our model, these approximations take place on individual branches of a phylogeny, which typically span relatively short intervals of time. For our next simulation experiment, we scaled up to simulating sets of branchwise averages on entire phylogenies. For each parameter combination (excluding combinations where rate variance is 0), we repeated the same simulations on 100 pure birth phylogenies with either 50, 100, or 200 species (generated using the R package *phytools*; Revell, 2012)

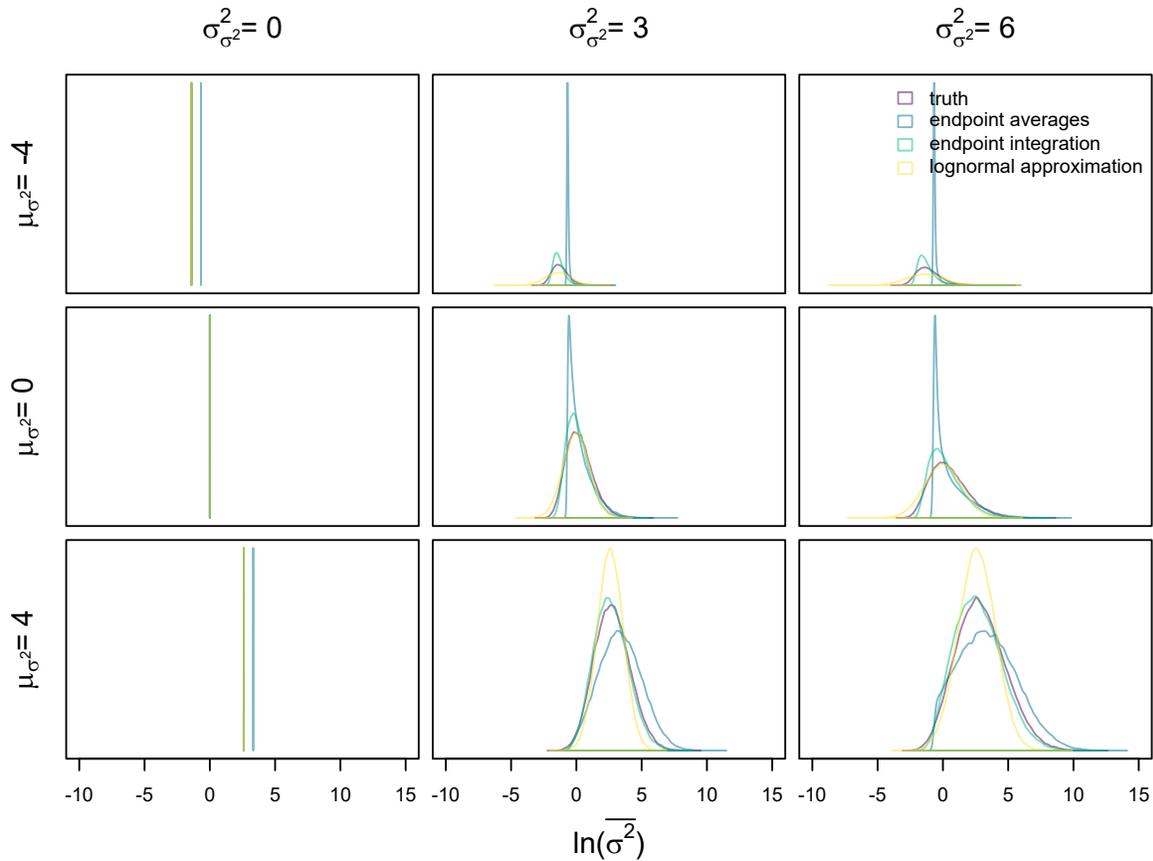


Figure 1B.1 Distributions of simulated branchwise averages under different approximation strategies and the true distribution given parameter combinations used in the main text’s simulation study. All simulations were run on single branches of length 1.

standardized to a height of 1. For each phylogeny, we simulated 1,000 sets of branchwise averages under each approximation strategy, as well as fine-grained GBM sample paths ( 1,000 time points across entire phylogeny’s height) representing the true distribution. Because these samples have a high number of dimensions (one for each branch in a phylogeny), we visualized how well these multivariate distributions match one another using summary statistics. Specifically, for each tree, we recorded the correlation coefficients between the means/(co)variances of branchwise averages simulated under each approximation strategy and the true distribution (Figs. 1B.2–1B.7). To have a null expectation for these correlation coefficients, we also simulated a second true distribution and estimated correlation coefficients for means/(co)variances between replicate true distributions.

Overall, the results indicate that all approximations do a fairly good job at recapitulating the means and (co)variances expected under the true distribution. The log-normal approximation no-

tably exhibits uncorrelated means in the case of no trend, in contrast to other approximations. This is due to the log-normal approximation lacking the right skew of the true distribution and other approximations (Fig. 1B.1), which naturally inflates the means of branchwise average distributions along long branches. In the case of any trend, the endpoint average approximation exhibits somewhat less strong correlations between branchwise average means compared to other approximations. When rate variance is high, the log-normal approximation exhibits performance intermediate between the endpoint average approximation and endpoint integration approximation/null distribution. However, even the worst performing simulations nearly always exhibit strong correlations in branchwise average means above 0.98. In contrast to means, correlations for branchwise average (co)variances consistently varied between about 0.98-0.99 regardless of simulation parameters or approximation strategy, closely matching the null distribution.

Because GBM time-averages are non-normally distributed, we also sought a non-parametric method of comparing samples from the approximations and true distributions. For this, we attempted to use the R package *FNN* (Beygelzimer et al., 2022) to estimate Kullback-Leibler (KL) divergence from each approximation to the true distribution. However, this estimator exhibited severe numerical issues, like negative KL divergence estimates. Thus, we instead implemented a crude K nearest neighbor probability density estimator (Zhao and Lai, 2021). For each tree in the simulation experiment above, we used this estimator to calculate local probability densities under each approximation and the true distribution around samples from a replicate true distribution. We then calculated log ratios of the true densities to densities under each approximation and averaged the distances between these log ratios and 0 (i.e., equal densities). These averaged distances give a rough sense of how well the probability density of each approximation matches that of the true distribution, with increased sampling in higher-density regions of the true distribution (Figs. 1B.8 and 1B.9). Overall, the average log density ratio distances under each approximation matches the null distribution well. The endpoint average and log-normal approximations exhibit marginally elevated distances in the case of non-zero trends and decreasing trends, respectively, likely due to these approximations' under/overestimation of branchwise average variance in certain regions of

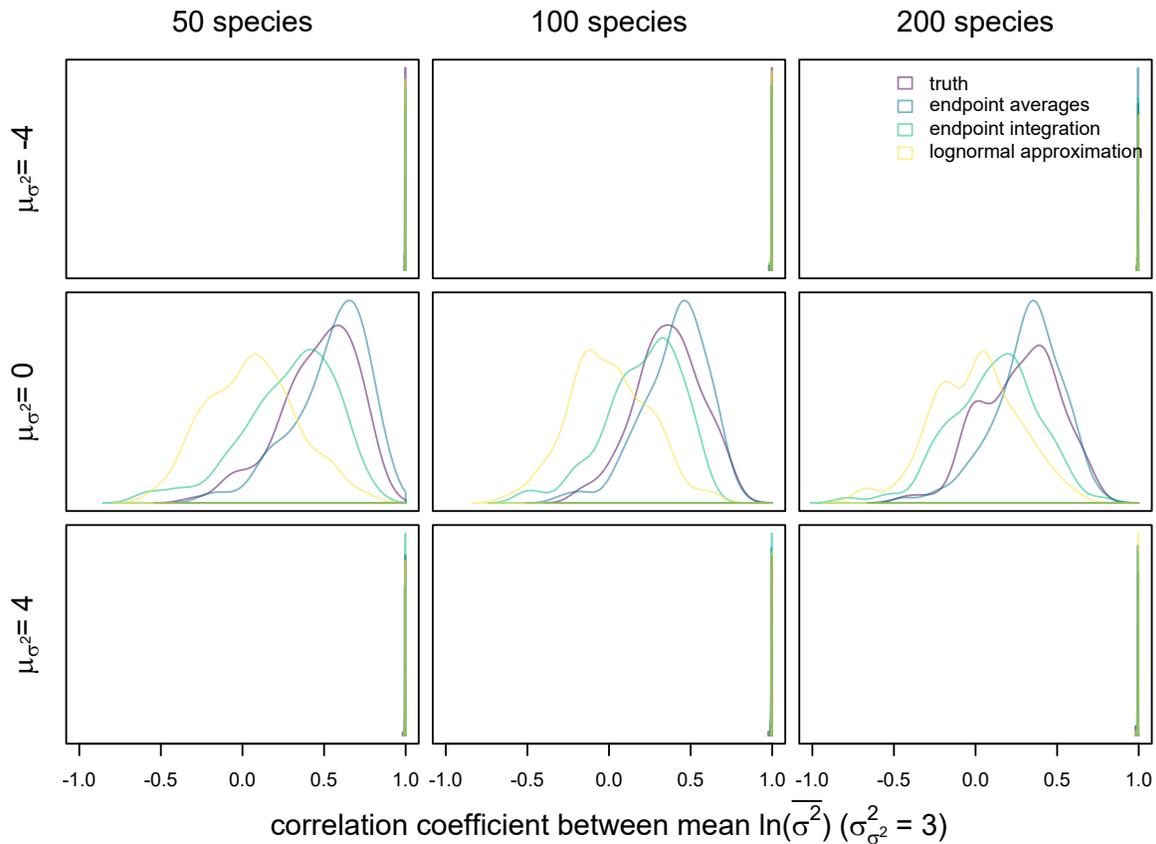


Figure 1B.2 Distributions of correlation coefficients between mean simulated branchwise averages under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 3. All simulations were run on pure-birth phylogenies of height 1.

parameter space (Fig. 1B.1).

Lastly, we redid our entire simulation study with trait evolution rates simulated as evolving under a fine-grained GBM process ( $\sim 500$  time points across entire phylogeny's height). We present all figures and tables for this simulation study below (Figs. 1B.10–1B.14; Tables 1.3, 1B.1 and 1B.2). In general, the results qualitatively match those of the simulation study presented in the main text, and we feel confident that the log-normal approximation of branchwise averages is sufficient for our model. While there is some discrepancy in the statistical power of trend detection compared to results in the main text, it is unlikely such discrepancies result from systematic bias. Notably, statistical power for trend detection even under conventional EB/LB models in this simulation study also differs from the main text results, suggesting that any discrepancies are attributable to variation in the simulated data.

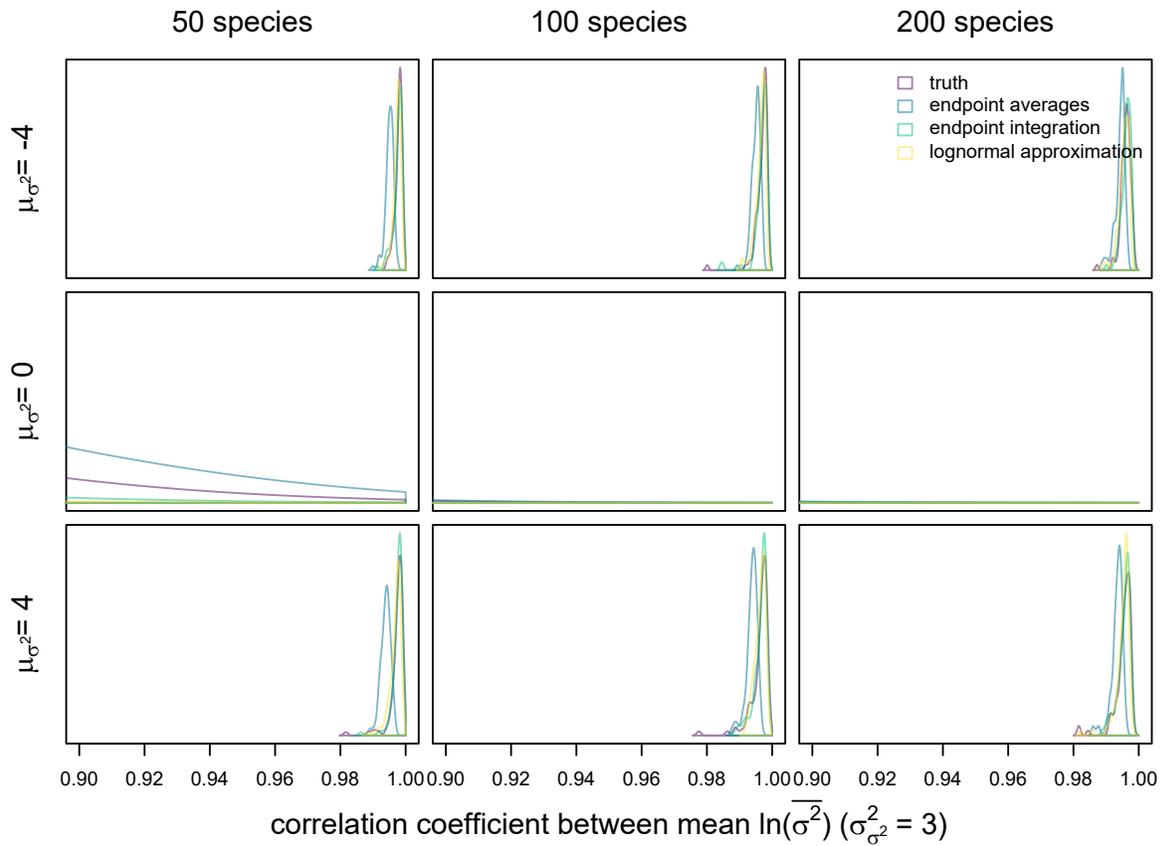


Figure 1B.3 Distributions of correlation coefficients between mean simulated branchwise averages under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 3. All simulations were run on pure-birth phylogenies of height 1. Plots are zoomed in on distributions close to 1.

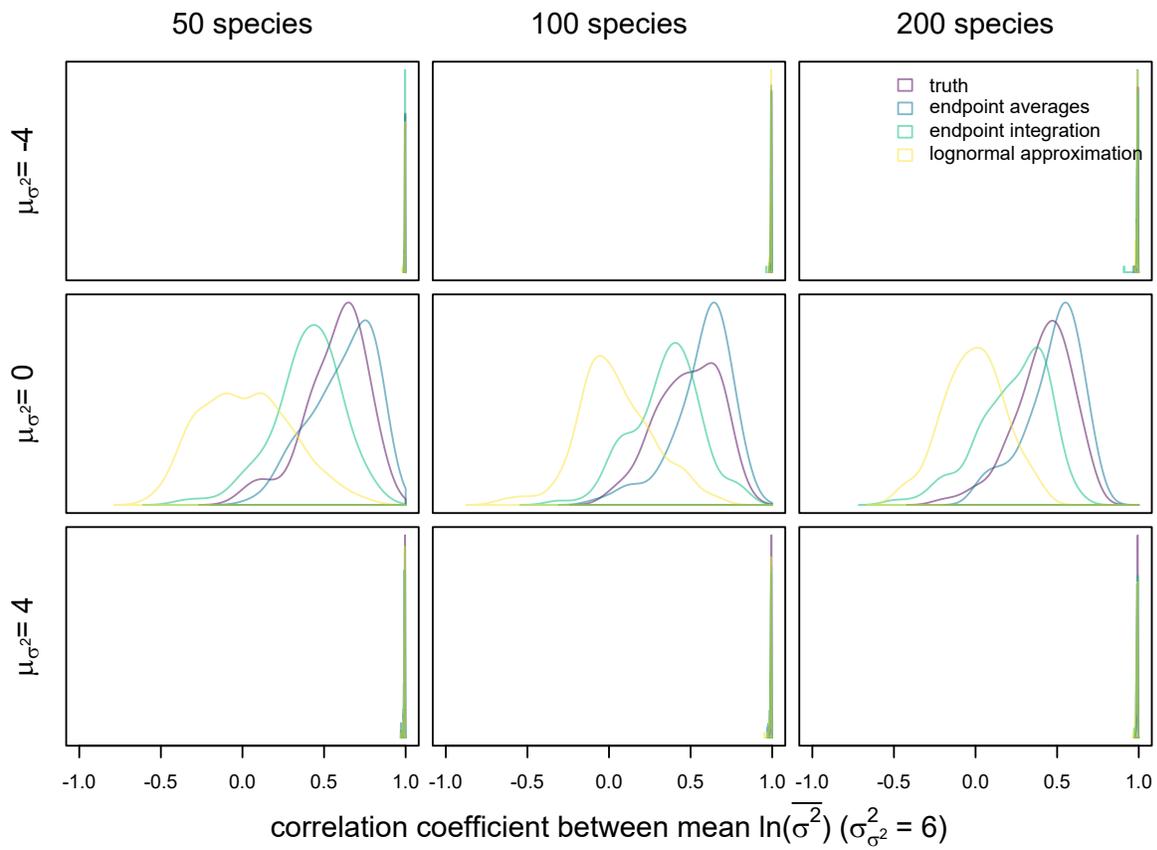


Figure 1B.4 Distributions of correlation coefficients between mean simulated branchwise averages under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 6. All simulations were run on pure-birth phylogenies of height 1.

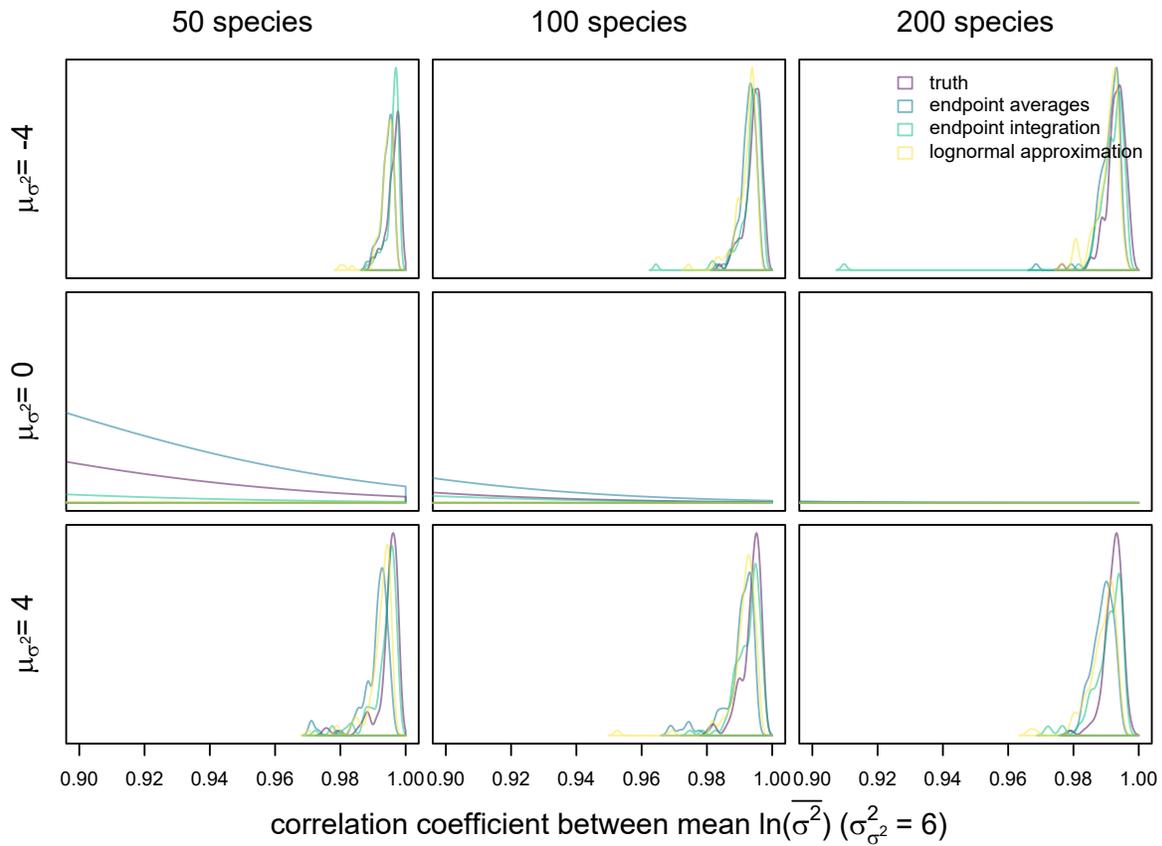


Figure 1B.5 Distributions of correlation coefficients between mean simulated branchwise averages under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 6. All simulations were run on pure-birth phylogenies of height 1. Plots are zoomed in on distributions close to 1.

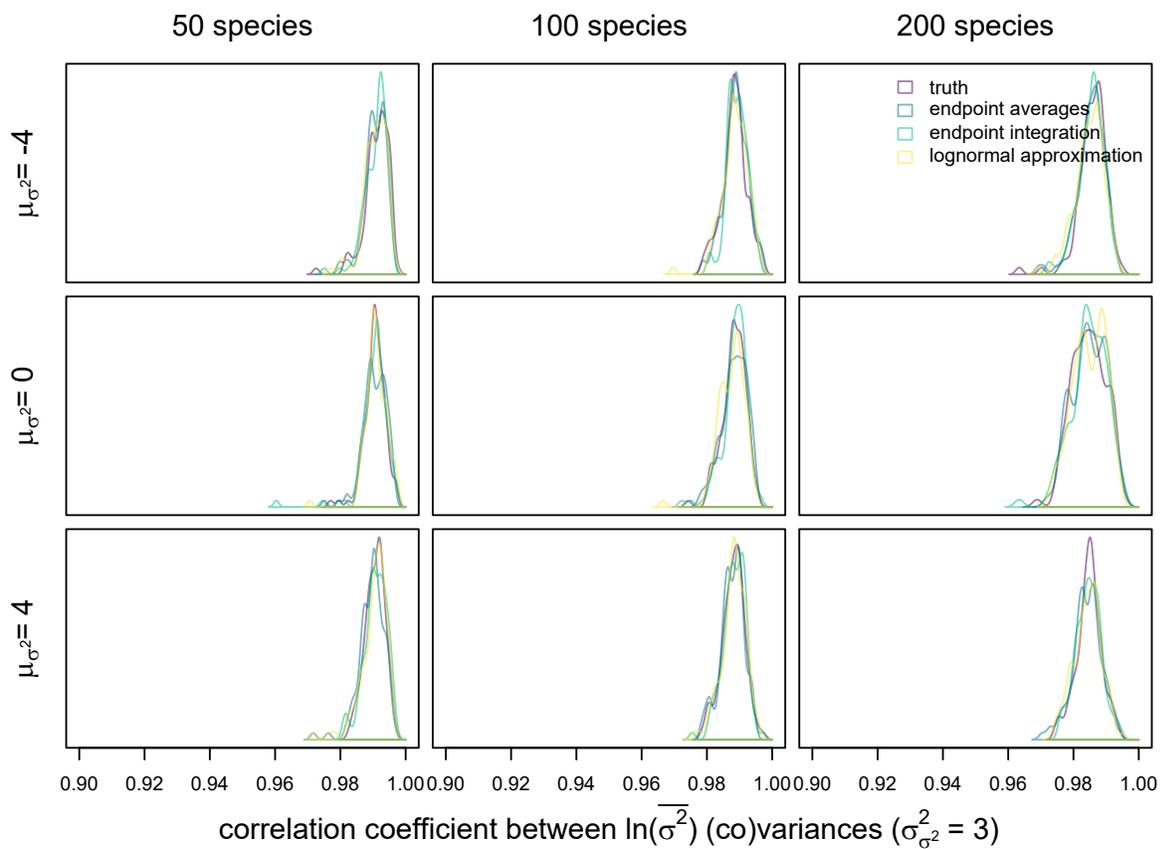


Figure 1B.6 Distributions of correlation coefficients between simulated branchwise average (co)variances under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 3. All simulations were run on pure-birth phylogenies of height 1.

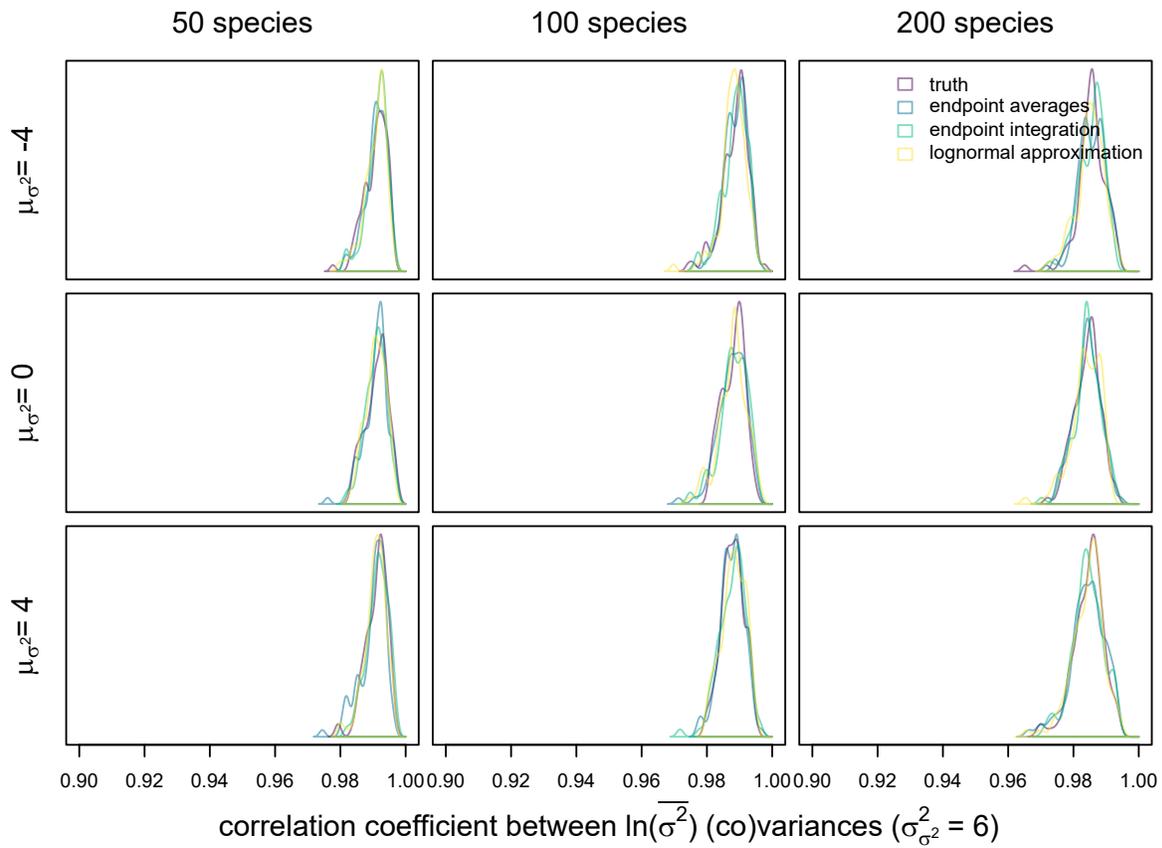


Figure 1B.7 Distributions of correlation coefficients between simulated branchwise average (co)variances under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 6. All simulations were run on pure-birth phylogenies of height 1.

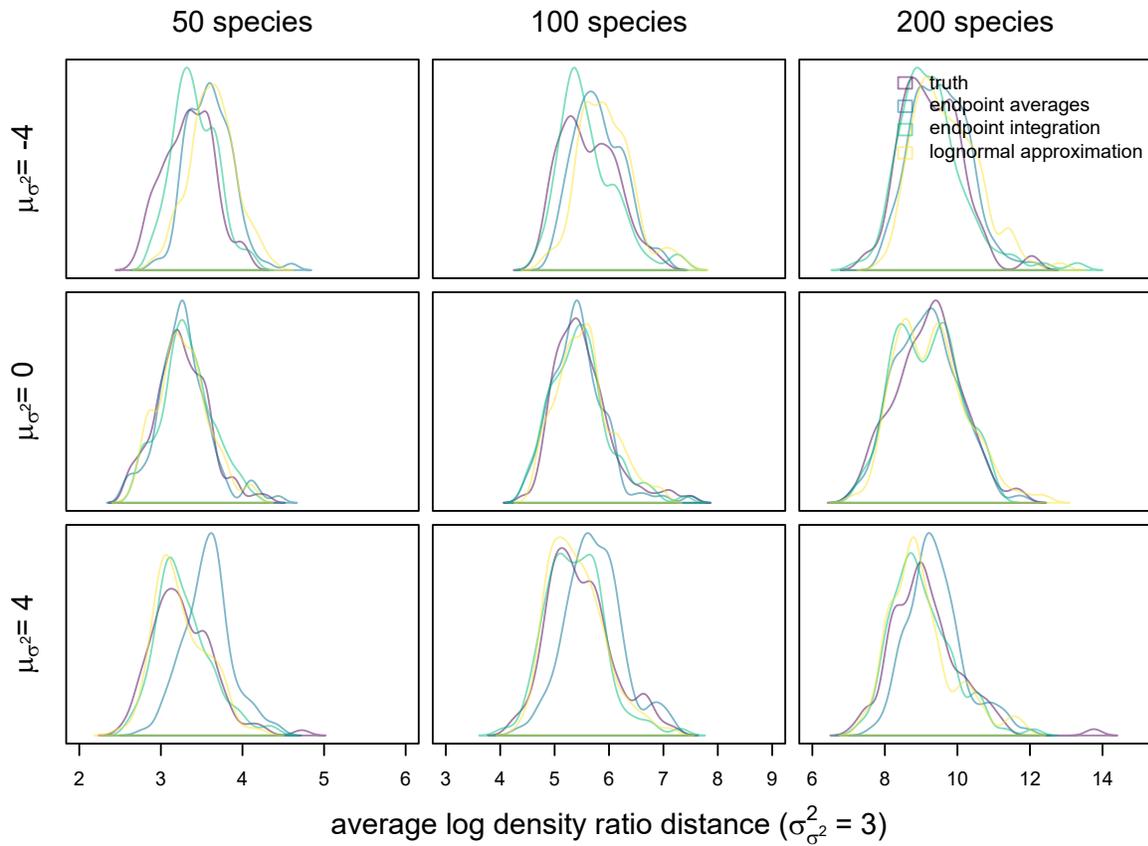


Figure 1B.8 Distributions of average log density ratio distances between simulated branchwise average distributions under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 3. Probability densities were estimated via K nearest neighbors. All simulations were run on pure-birth phylogenies of height 3.

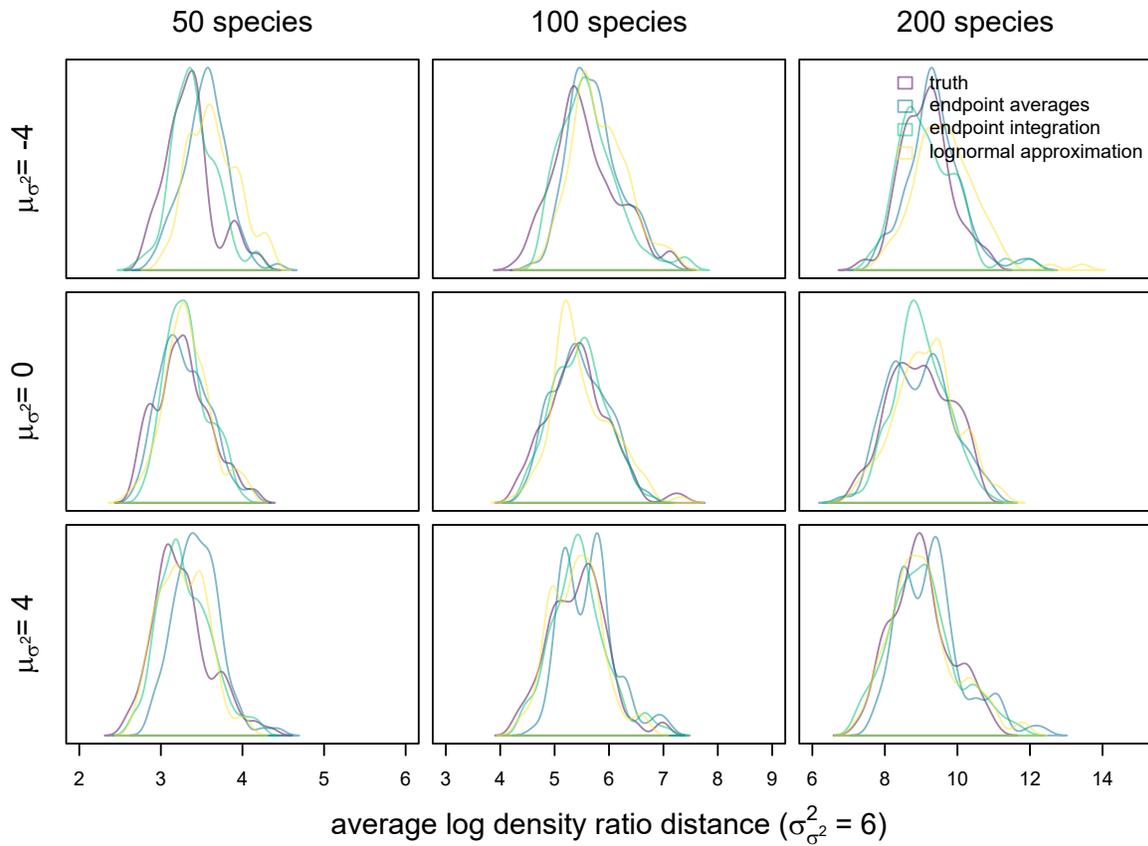


Figure 1B.9 Distributions of average log density ratio distances between simulated branchwise average distributions under different approximation strategies and the true distribution with rate variance ( $\sigma_{\sigma^2}^2$ ) set to 6. Probability densities were estimated via K nearest neighbors. All simulations were run on pure-birth phylogenies of height 1.

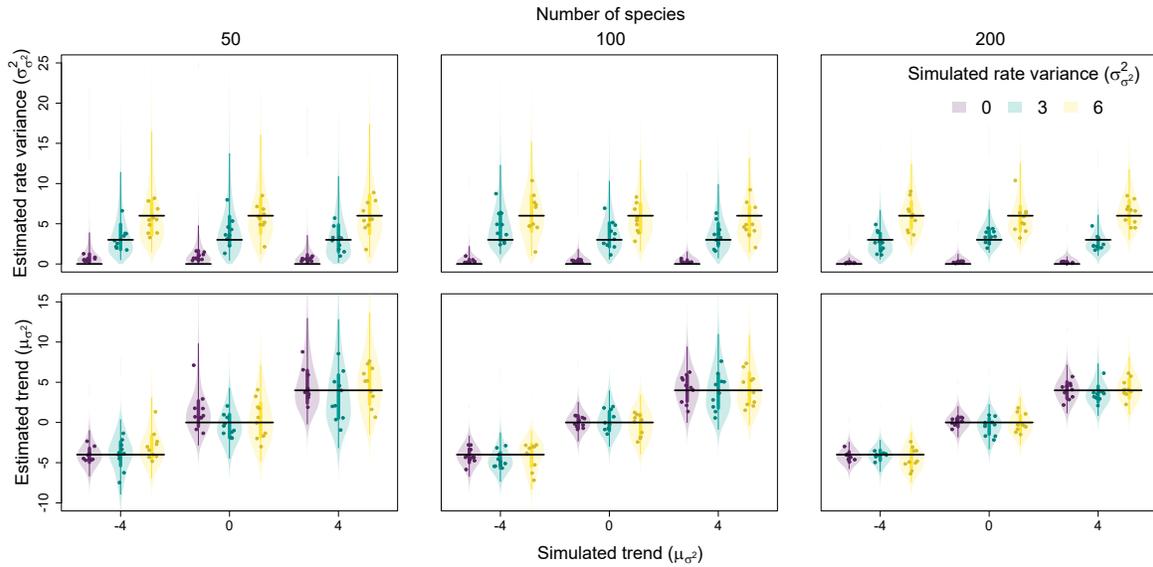


Figure 1B.10 Relationship between simulated and estimated rate variance ( $\sigma_{\sigma^2}^2$ ) and trend ( $\mu_{\sigma^2}$ ) parameters. Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values.

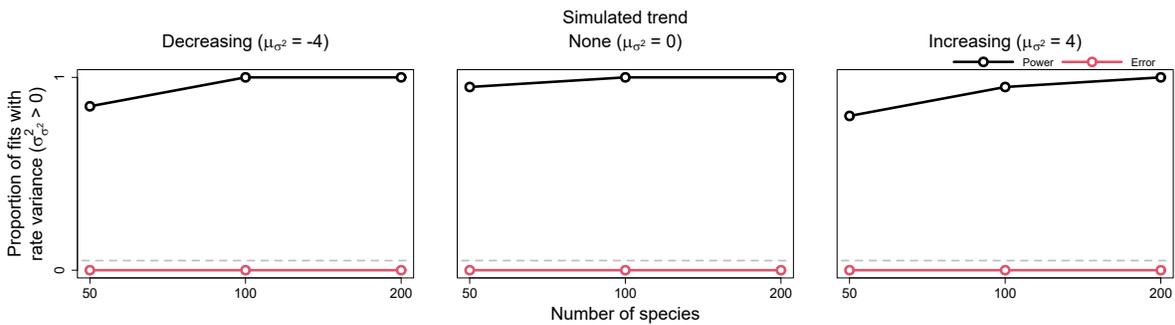


Figure 1B.11 Power and error rates for the rate variance parameter ( $\sigma_{\sigma^2}^2$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for rate variance significantly greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in red) as a function of tree size.

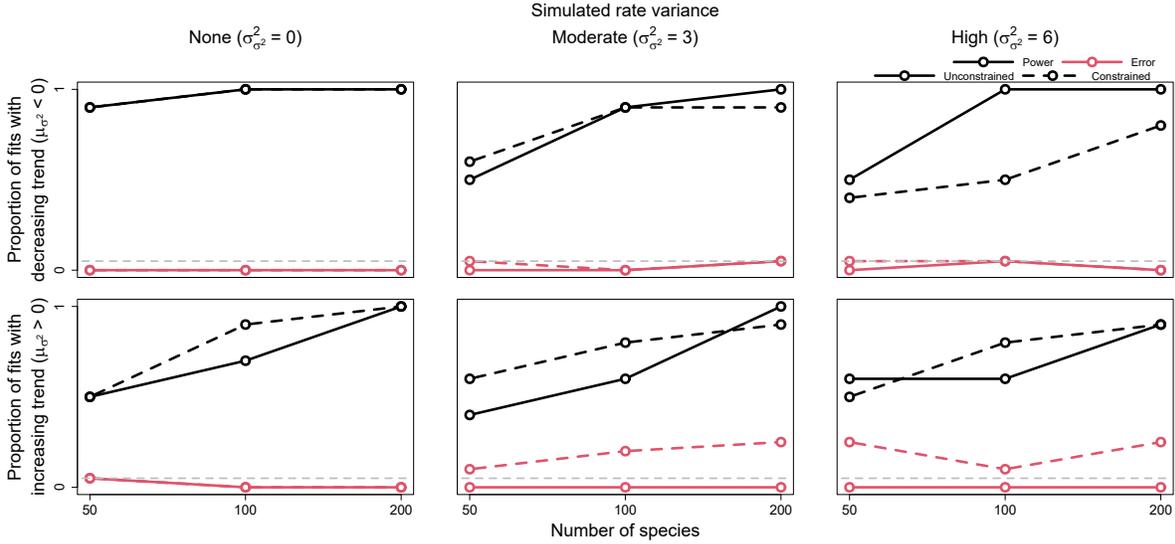


Figure 1B.12 Power and error rates for the trend parameter ( $\mu_{\sigma^2}$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for trends significantly less and greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of tree size. Results are shown for both models allowed to freely estimate rate variance ( $\sigma_{\sigma^2}^2$ ) (i.e., unconstrained models, solid lines) and models with rate variance constrained to 0 (i.e., constrained models, dashed lines). The latter models are identical to conventional early/late burst models.

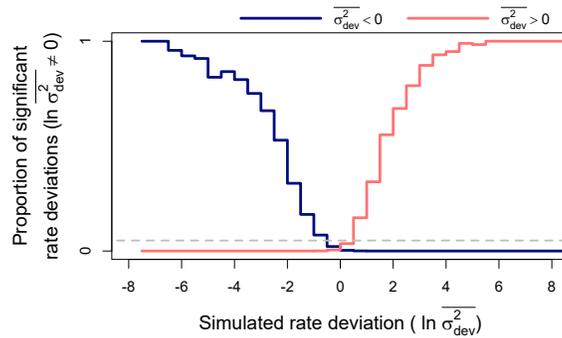


Figure 1B.13 Power and error rates for branchwise rate parameters ( $\ln \overline{\sigma^2}$ ). Lines depict changes in proportions of branchwise rates considered anomalously slow (in dark blue) or fast (in light red) as a function of simulated rate deviations ( $\ln \overline{\sigma_{dev}^2}$ ). These results combine all fits to simulated data that detected rate variance ( $\sigma_{\sigma^2}^2$ ) significantly greater than 0. The proportions are equivalent to power when the detected rate deviation is of the same sign as the true, simulated deviation (left of 0 for anomalously slow rates in dark blue and right for anomalously fast rates in light red), and to error rate when the detected and true rate deviations are of opposite signs. Here, significant rate deviations for simulated rate deviations that are exactly 0 are considered errors regardless of sign.

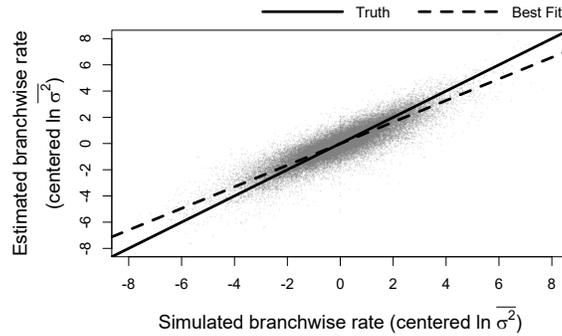


Figure 1B.14 Relationship between simulated and estimated branchwise rate parameters ( $\ln \overline{\sigma^2}$ ). For each simulation and posterior sample, branchwise rates were first centered by subtracting their mean. We estimated centered branchwise rates by taking the median of the centered posterior samples. The solid line represents the position of the true centered branchwise rates, while the shallower, dashed line represents the observed line of best fit for these data.

Table 1B.1 Median absolute errors of rate variance, trend, and branchwise rate posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates			
	0	3	6	0	3	6	0	3	6	
	50 species									
$\mu_{\sigma^2} =$	-4	0.61	1.58	2.26	0.94	1.68	1.78	0.42	0.80	0.96
	0	0.89	1.89	2.23	2.09	1.56	2.22	0.62	0.82	1.04
	4	0.58	1.68	2.41	2.15	2.98	2.62	0.63	0.92	0.98
	100 species									
$\mu_{\sigma^2} =$	-4	0.31	2.11	2.37	0.91	1.22	1.43	0.32	0.77	0.86
	0	0.31	1.59	1.95	0.81	1.26	1.47	0.32	0.82	0.93
	4	0.26	1.49	2.21	1.67	2.16	2.02	0.41	0.85	0.94
	200 species									
$\mu_{\sigma^2} =$	-4	0.14	1.23	1.79	0.62	0.66	1.29	0.23	0.68	0.80
	0	0.21	0.93	1.82	0.65	1.09	1.10	0.24	0.72	0.84
	4	0.18	0.98	1.50	1.09	1.17	1.27	0.28	0.73	0.84

Table 1B.2 Breadths of rate variance, trend, and branchwise rate posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates			
	0	3	6	0	3	6	0	3	6	
	50 species									
$\mu_{\sigma^2} =$	-4	3.67	9.11	12.98	4.66	6.02	6.81	2.28	3.24	3.65
	0	4.38	10.67	12.60	7.28	7.09	8.00	2.60	3.41	3.89
	4	3.35	9.00	13.88	10.34	10.95	12.09	2.81	3.50	4.10
	100 species									
$\mu_{\sigma^2} =$	-4	1.77	7.96	9.58	3.53	4.56	4.72	1.71	3.22	3.46
	0	1.64	6.72	9.15	4.04	5.09	5.67	1.76	3.12	3.42
	4	1.36	6.77	8.13	6.74	8.08	7.86	1.87	3.31	3.55
	200 species									
$\mu_{\sigma^2} =$	-4	0.71	3.97	7.20	2.64	3.58	4.06	1.24	2.50	3.12
	0	1.04	4.26	6.52	3.34	3.98	4.15	1.36	2.77	3.25
	4	0.79	3.62	6.89	4.53	4.88	5.69	1.39	2.70	3.37

Table 1B.3 Coverage of rate variance, trend, and branchwise rate posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

$\sigma_{\sigma^2}^2 =$	rate variance			trend			branchwise rates		
	0	3	6	0	3	6	0	3	6
50 species									
$\mu_{\sigma^2} =$	-4	—	1.00	1.00	0.80	0.90	1.00	0.95	0.94
	0	—	1.00	1.00	0.90	1.00	1.00	0.97	0.98
	4	—	1.00	1.00	1.00	0.80	1.00	0.95	0.94
100 species									
$\mu_{\sigma^2} =$	-4	—	0.70	0.90	0.90	1.00	0.90	1.00	0.96
	0	—	1.00	1.00	1.00	1.00	0.90	1.00	0.94
	4	—	1.00	0.90	0.90	0.90	1.00	0.99	0.95
200 species									
$\mu_{\sigma^2} =$	-4	—	0.90	1.00	1.00	1.00	0.90	0.99	0.93
	0	—	1.00	0.80	1.00	0.90	1.00	1.00	0.95
	4	—	1.00	1.00	0.90	1.00	1.00	0.99	0.93

## APPENDIX 1C

### AVERAGE CHANGES IN TRAIT EVOLUTION RATES

Conventional early/late burst (EB/LB) models of trait evolution assume that rates follow a homogeneous, exponential declines or increases with respect to time (Blomberg et al., 2003). The definition of EBs/LBs under such models is thus straight-forward—any given time slice in a clade’s history is associated with a single trait evolution rate, and these rates can only decrease, increase or stay the same. On the other hand, allowing for rate heterogeneity independent of overall temporal trends means that any given time slice in a clade’s history is associated with a *distribution* of trait evolution rates. Because of this, our new method allows for alternative definitions of EBs/LBs, depending on how one summarizes these distributions. In the current study, we mainly consider a definition based on whether the medians, or geometric means, of these distributions decrease or increase over time (change per unit time given by  $\mu_{\sigma^2}$ , hereafter the “trend” parameter, as in the main text). Alternatively, one could use a definition based on whether the average, or arithmetic means, of these distributions decrease or increase over time (change per unit time given by  $\mu_{\sigma^2} + \sigma_{\sigma^2}^2/2$ , hereafter the “average change” parameter,  $\delta_{\sigma^2}$ ).

We chose to focus on trend over average change estimation and define EBs/LBs based on the trend parameter for a few reasons. First, average change is a composite parameter of both the trend and rate variance parameters, posing some interpretational challenges. In general, it seems more intuitive to consider the magnitude of deterministic changes in trait evolution rates (the trend component) apart from the magnitude of stochastic changes (the rate variance component). Second, because rates evolve in an approximately log-normal manner under our model, medians are a natural, reliable way of summarizing their distributions, corresponding to the exponentiated average of rates on the natural log scale. In contrast, the right skew of log-normal distributions causes raw averages of trait evolution rates to be highly influenced by few, extreme outliers, particularly when rate variance is high. For this reason, our model can produce trait evolution scenarios whereby rates exhibit declines in the majority of lineages (directly related to changes in median rates) while increasing on average (Fig. 1C.1 and 1C.2). Lastly, many macroevolutionary biologists consider

“accounting” for lineages/subclades exhibiting unusual trait evolution rates critical to elucidating and understanding changes in rates over time (Lloyd et al., 2012; Slater and Pennell, 2014; Benson et al., 2014; Hopkins and Smith, 2015; Wright, 2017; Puttick, 2018). This implies that many empiricists intuitively define EBs/LBs based on majority changes in rates rather than changes in average rates. Additionally, by log-transforming traits prior to analysis, many macroevolutionary biologists implicitly use GBM processes to model trait evolution, just as we use a (approximate) GBM process to model rate evolution here. In the context of trait evolution, the analogous trend parameter is widely considered by empiricists and method developers alike to determine whether a clade exhibits a directional “evolutionary trend” in traits, regardless the estimated variance parameter (Hunt, 2006; Raj Pant et al., 2014; Sookias et al., 2012; Gill et al., 2017).

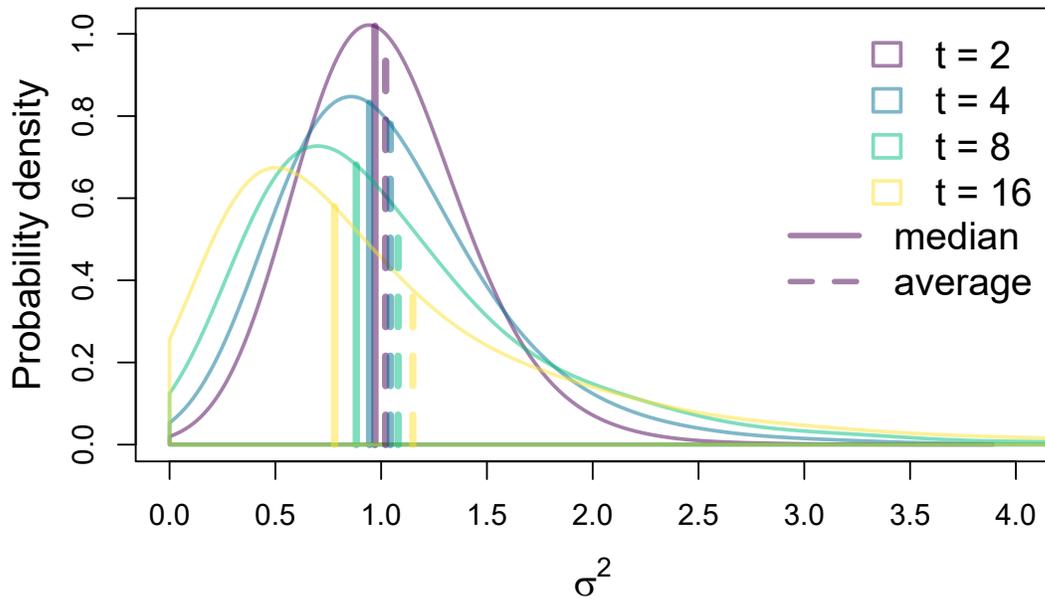


Figure 1C.1 Distributions of 6,000 rates simulated as evolving under a GBM process with trend of -0.015 and rate variance of 0.05 at various time points, with starting rate of 1 at time  $t = 0$ . Parameter values were chosen to clearly illustrate how rates under our model may exhibit majority declines while increasing on average due to the skewed nature of rate change. Solid and dashed vertical lines represent the positions of median and average rate values, respectively, for each time point.

Here, we briefly consider our new method’s performance with respect to estimating and de-

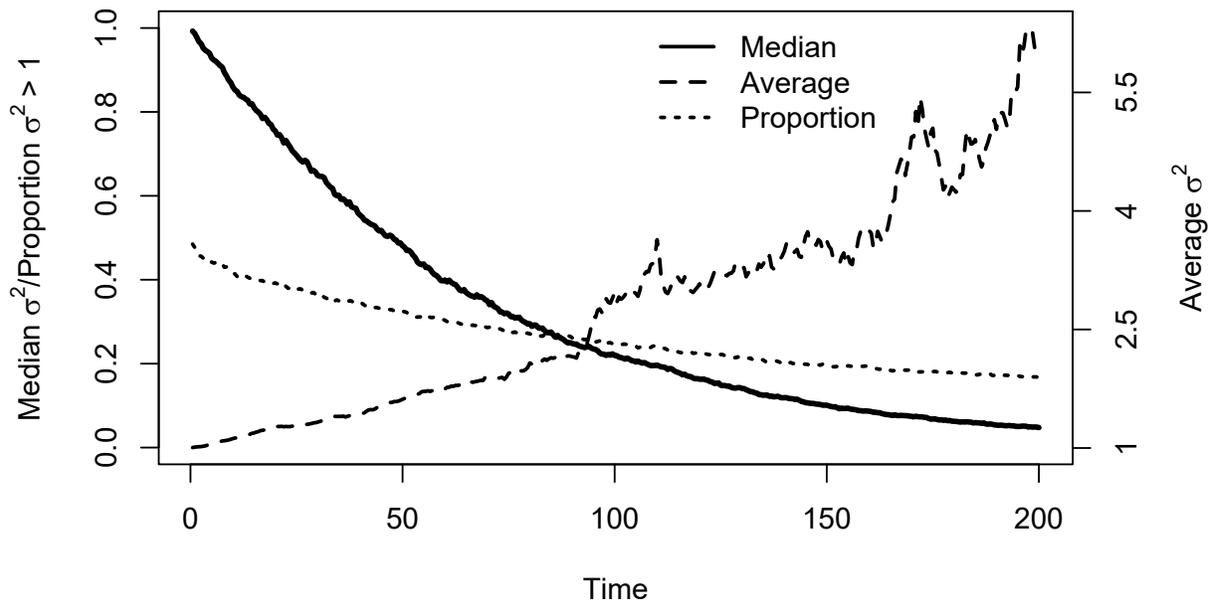


Figure 1C.2 Changes over time in the median and average of 6,000 rates simulated as evolving under a GBM process with trend of  $-0.015$  and rate variance of  $0.05$ , with starting rate of  $1$  at time  $t = 0$ . Parameter values were chosen to clearly illustrate how rates under our model may exhibit majority declines while increasing on average due to the skewed nature of rate change. Solid and dashed lines depict changes in median and average rate values, respectively, while the dotted line depicts changes in the proportion of rates greater than the starting rate of  $1$ .

tecting average changes in trait evolution rates. Interestingly, our simulation study results revealed that, in the presence of time-independent rate heterogeneity, conventional EB/LB models (equivalent to our new models with rate variance constrained to  $0$ ) appear to estimate average change, rather than trend parameters, as defined under our model (Fig. 1C.3 and 1C.4). We are not aware of any previous research explicitly demonstrating this phenomenon. When comparing performance of constrained to unconstrained models with respect to detecting significant average change (i.e., 95% equal-tailed interval lies entirely below or above  $0$ ), we generally see only a modest reduction in error rates and greatly reduced power to detect negative average change under the full, unconstrained model (Fig. 1C.5). Nonetheless, inference of the average change parameter seems substantially improved under unconstrained models (Tables 1C.1–1C.3). In the presence of time-independent rate heterogeneity, constrained models tend to exhibit less accurate, overly-narrow

posterior estimates of average change, particularly when the rate variance and trend parameters are high, resulting in low posterior coverage. This warrants caution in interpreting the results of conventional EB/LB models fitted to comparative data exhibiting substantial time-independent rate heterogeneity, and we recommend estimating rate variance even when one's only goal is to estimate changes in average trait evolution rates over time.

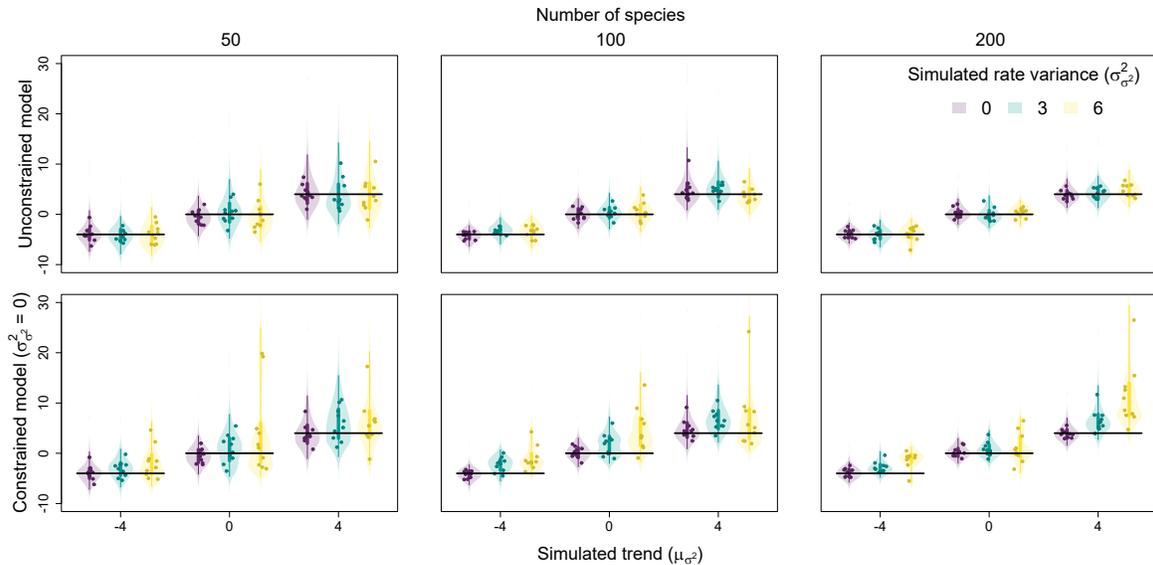


Figure 1C.3 Relationship between simulated rate variance ( $\sigma_{\sigma^2}^2$ )/trend ( $\mu_{\sigma^2}$ ) and estimated trend parameters. Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values. Results for models with estimated rate variance unconstrained and constrained to 0 are shown on top and bottom, respectively.

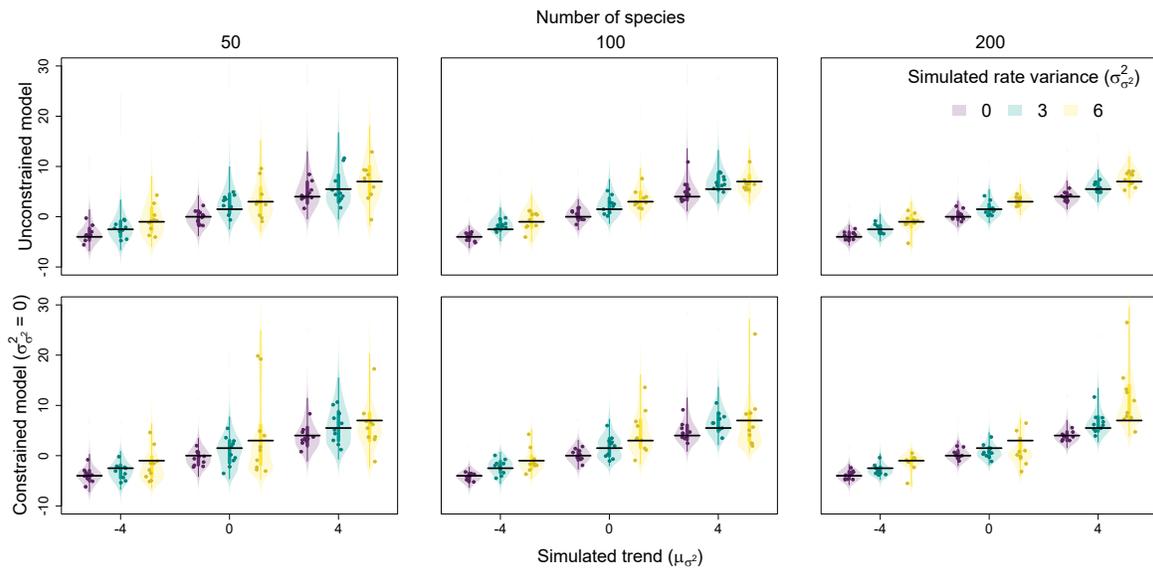


Figure 1C.4 Relationship between simulated rate variance ( $\sigma_{\sigma^2}^2$ )/trend ( $\mu_{\sigma^2}$ ) and estimated average change ( $\delta_{\sigma^2}$ ) parameters. Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values. Results for models with estimated rate variance unconstrained and constrained to 0 are shown on top and bottom, respectively.

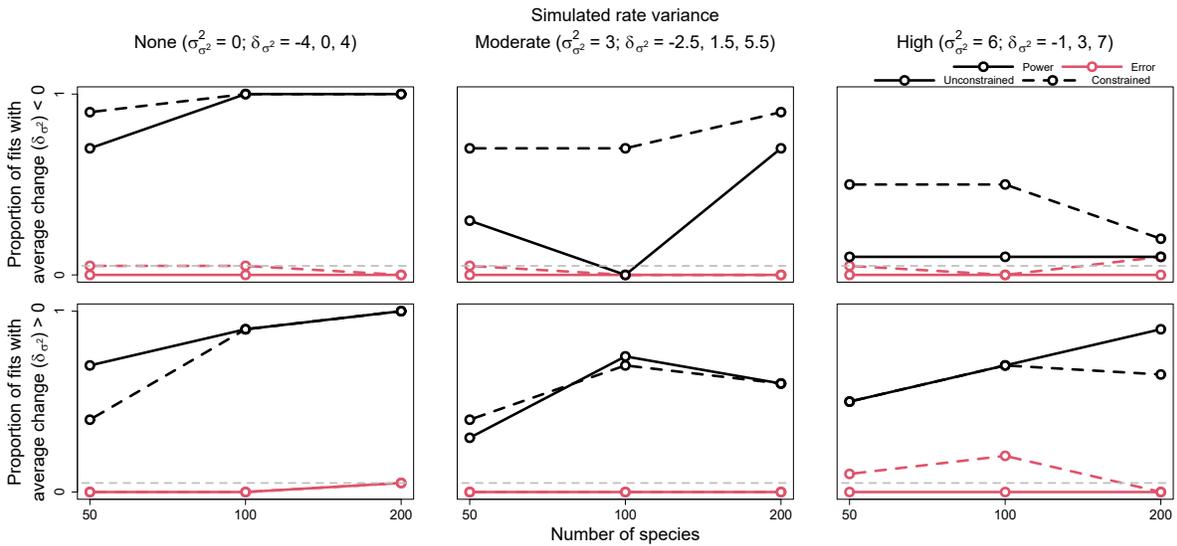


Figure 1C.5 Power and error rates for the average parameter ( $\delta_{\sigma^2}$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for average change significantly less and greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of tree size. Results are shown for both models allowed to freely estimate rate variance ( $\sigma_{\sigma^2}^2$ ) (i.e., unconstrained models, solid lines) and models with rate variance constrained to 0 (i.e., constrained models, dashed lines). The latter models are identical to conventional early/late burst models.

Table 1C.1 Median absolute errors of average change posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy) under models with rate variance unconstrained and constrained to 0, averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

		unconstrained			constrained		
$\sigma_{\sigma^2}^2 =$		0	3	6	0	3	6
50 species							
$\mu_{\sigma^2} =$	-4	1.41	1.61	2.50	1.23	1.50	2.74
	0	1.43	2.10	3.08	1.45	2.45	6.08
	4	2.22	3.04	3.34	2.05	3.05	3.87
100 species							
$\mu_{\sigma^2} =$	-4	0.78	1.27	1.70	0.74	1.28	1.78
	0	1.15	1.65	1.72	1.08	1.88	3.35
	4	1.92	1.98	1.85	1.70	2.08	4.39
200 species							
$\mu_{\sigma^2} =$	-4	0.79	0.92	1.44	0.78	0.96	1.19
	0	0.92	1.21	1.01	0.90	1.35	3.05
	4	0.97	1.15	1.51	0.94	1.80	5.06

Table 1C.2 Breadths of average change posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision) under models with rate variance unconstrained and constrained to 0, averaged across replicates for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

		unconstrained			constrained		
$\sigma_{\sigma^2}^2 =$		0	3	6	0	3	6
50 species							
$\mu_{\sigma^2} =$	-4	5.56	7.95	10.27	4.50	4.65	5.63
	0	6.25	9.89	11.46	5.47	6.82	8.45
	4	11.04	11.69	12.45	9.48	10.81	10.60
100 species							
$\mu_{\sigma^2} =$	-4	3.42	5.48	6.54	3.07	3.49	3.84
	0	4.46	6.27	7.44	3.97	4.58	6.60
	4	7.64	8.97	8.56	7.12	8.41	8.40
200 species							
$\mu_{\sigma^2} =$	-4	2.82	4.14	5.07	2.70	2.87	3.26
	0	3.40	4.50	5.13	3.25	3.30	3.37
	4	4.51	5.45	6.29	4.38	5.78	8.73

Table 1C.3 Coverage of average change posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) under models with rate variance unconstrained and constrained to 0 for each simulated trait evolution scenario and tree size.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively.

		unconstrained			constrained		
$\sigma_{\sigma^2}^2 =$		0	3	6	0	3	6
50 species							
$\mu_{\sigma^2} =$	-4	0.90	1.00	0.90	0.80	0.60	0.50
	0	1.00	1.00	1.00	0.90	0.80	0.40
	4	1.00	1.00	0.90	1.00	1.00	0.80
100 species							
$\mu_{\sigma^2} =$	-4	1.00	1.00	0.90	1.00	0.80	0.60
	0	1.00	0.90	0.90	0.90	0.60	0.60
	4	0.90	1.00	1.00	0.90	0.90	0.60
200 species							
$\mu_{\sigma^2} =$	-4	1.00	1.00	0.90	1.00	0.90	0.80
	0	0.90	1.00	1.00	0.90	0.60	0.10
	4	1.00	1.00	1.00	1.00	0.90	0.50

## APPENDIX 1D

### PRIOR SENSITIVITY STUDY

To see how sensitive our method is to alternate prior specifications, we refit models to our smallest simulations (50 tips) while varying prior settings. We focus on the smallest simulations because the priors are more influential when there is less data. In addition to refitting models with default priors to each simulation (see *Priors* subsection of *Materials and Methods* section in main text), we also refit models with “tight” and “loose” prior settings, whereby the priors for rate variance ( $\sigma_{\sigma^2}$ ), trend ( $\mu_{\sigma^2}$ ), and root rate ( $\sigma_0^2$ ) parameters were made more or less informative, respectively. We did this by either reducing the prior scale parameter (i.e., standard deviation in the case of normal distributions) 5-fold for more precise, informative priors or increasing 3-fold for more relaxed, uninformative priors (i.e., prior scales of  $1/T$  for rate variance,  $2/T$  for trend, and 2 for root rate under the tight settings and  $15/T$ ,  $30/T$ , and 30 under the loose settings, where  $T$  is the height of the phylogeny). Within each of these three prior settings (tight, default, or loose), we additionally shifted the location of the root rate prior by either -3, 0, or 3, yielding a total of 9 prior settings. These shifts correspond to  $\sim 20$ -fold changes in the expected root rate.

Because this simulation study design requires many more model fits compared to the main text’s simulation study (9 trait evolution scenarios with 10 replicates refit under 9 different prior settings, yielding 810 model fits), we only ran 2 Hamiltonian Monte Carlo chains consisting of 1,500 iterations for each model fit and discarded the first 750 iterations as warmup. Chains still mixed relatively well despite the shorter chains (greatest  $\hat{R} \approx 1.021$ ), though effective sample sizes were unsurprisingly lower compared to results in the main text. Nonetheless, bulk effective sample sizes always exceeded the minimum recommended 100 per chain (Vehtari et al., 2021), and all tail effective sizes exceeded 100. Divergent transitions remained relatively rare, with 18 fits exhibiting a single divergent transition and another 4 with 2-5 each. Most low tail effective sample sizes and divergent transitions were associated with loose prior settings, likely reflecting difficulty in sampling the fat tails of posteriors under such priors.

Overall results suggest that *evorates* is robust to alternate prior specifications unless the priors

are overly informative (Figs. 1D.1–1D.3; Tables 1D.1–1D.12). In particular, shifting the root rate prior location had little effect on posterior distributions provided the prior’s scale is larger than the shift magnitude (as in the case of default and loose prior settings). Unsurprisingly, posterior precision generally decreased with more uninformative priors, and loose priors thus tended to yield less accurate posteriors with higher median absolute errors. Counterintuitively, however, default prior settings often resulted in more accurate posteriors than tight prior settings. In the case of branchwise rates, this is likely due to lower estimates of rate variance under tight priors, increasing the shrinkage of branchwise rate estimates (Fig. 1D.4). In the case of trend and root rate inference, this phenomenon mostly occurred when the root rate prior and simulated trend “conflict” by implying different patterns of rate change over time (e.g., a root rate prior shifted by -3 suggests rates must have increased over time to yield the observed trait data, while a decreasing trend implies the opposite). Accordingly, posterior coverage remained essentially constant at ~95% under default and loose prior settings, but dropped significantly—sometimes as low as 10%—under tight prior settings when the root rate prior and simulated trend conflicted in this manner.

Despite the relatively inaccurate inferences of branchwise rate, root rate, and trend parameters under overly informative priors, hypothesis testing was still largely reliable, albeit sometimes underpowered, under all prior settings we considered. Across the board, error rates remained conservative at around 5% or lower, with decreasing trends never mistaken for increasing trends and vice versa. Error rates for detecting significant rate variance may be slightly inflated under tight priors (Fig. 1D.5), perhaps due to tighter constraints on trend estimation forcing the model to instead attribute apparent rate heterogeneity to rate variance. Nonetheless, power to detect significant rate variance appears consistent regardless of prior settings. Notably, the same is true for anomalous rate detection, despite the increasing shrinkage of branchwise rate estimation under tighter priors (Fig. 1D.7). On the other hand, prior settings had considerable influence on power to detect trends (Fig. 1D.6), with generally increasing power under looser priors – particularly when the root rate prior shift and simulated trend both imply similar patterns of rate change over time (e.g., a root rate prior shifted by 3 and decreasing trend).

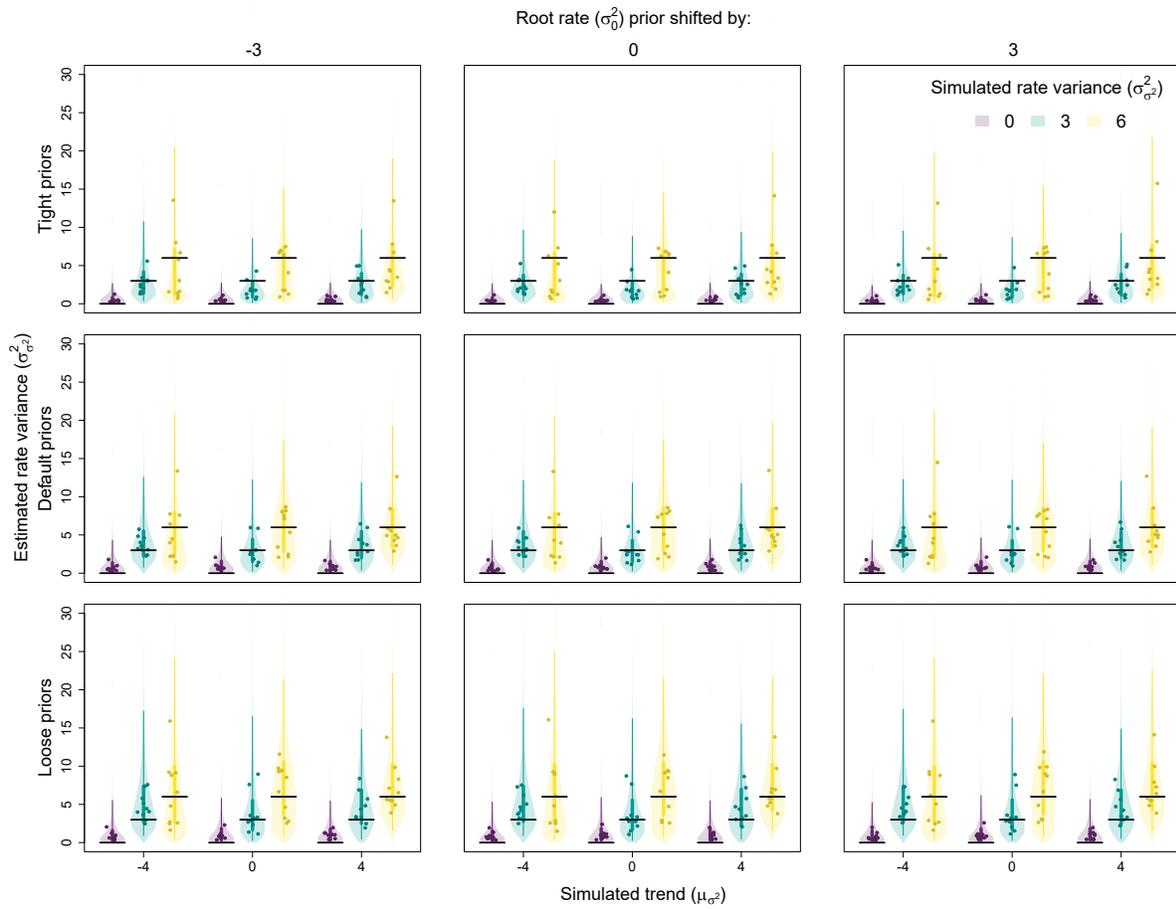


Figure 1D.1 The effect of trait evolution scenario and prior settings on inference of the rate variance parameter ( $\sigma_{\sigma^2}^2$ ). Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario and prior setting. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values.

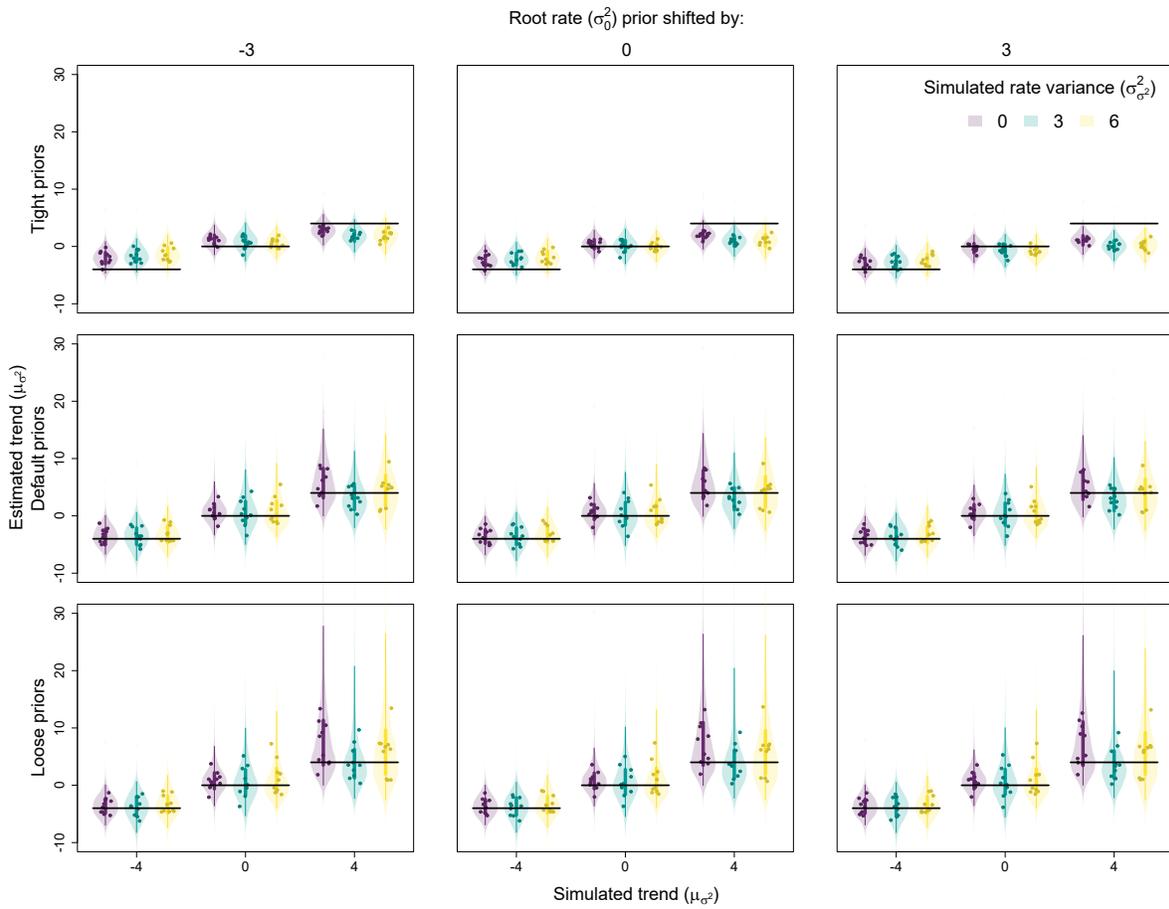


Figure 1D.2 The effect of trait evolution scenario and prior settings on inference of the trend parameter ( $\mu_{\sigma^2}$ ). Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario and prior setting. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values.

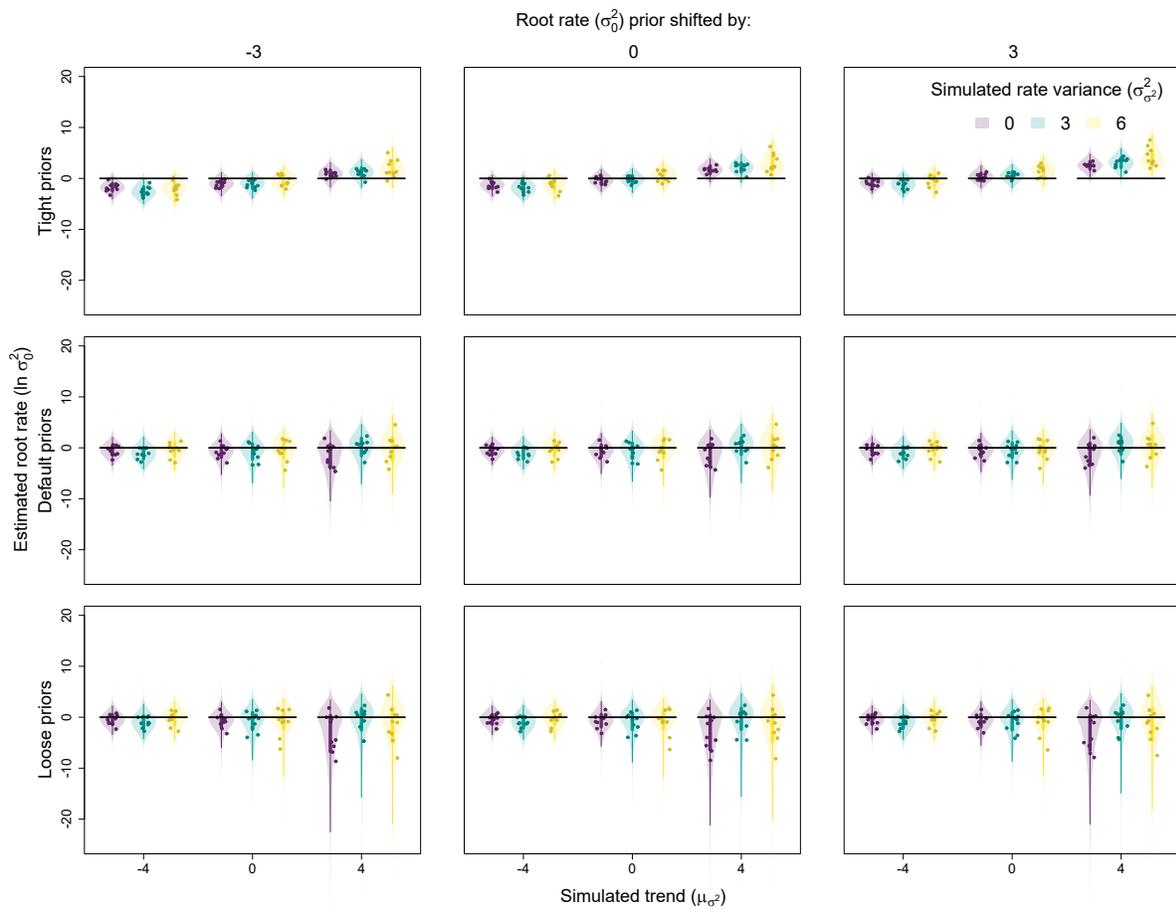


Figure 1D.3 The effect of trait evolution scenario and prior settings on inference of the root rate parameter ( $\sigma_0^2$ ). Each point is the posterior median from a single fit, while the violins are combined posterior distributions from all fits for a given trait evolution scenario and prior setting. Vertical lines represent the 50% (thicker lines) and 95% equal-tailed intervals (thinner lines) of these combined posteriors, while horizontal lines represent positions of true, simulated values.

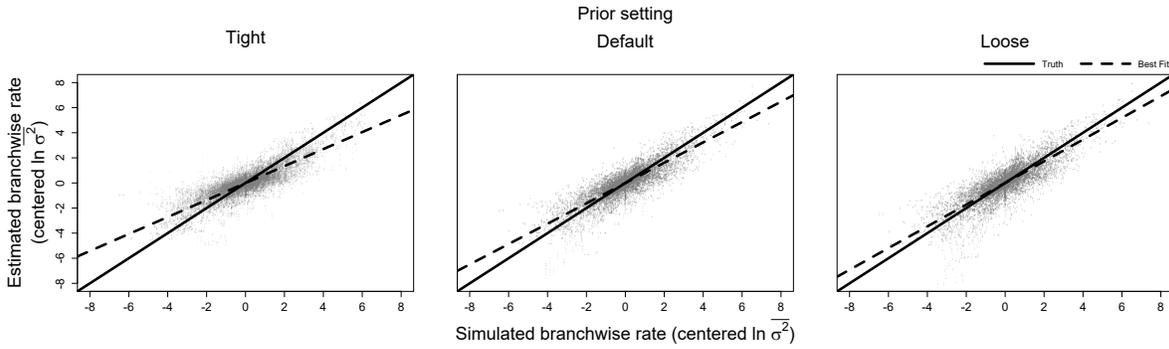


Figure 1D.4 Relationship between simulated and estimated branchwise rate parameters ( $\ln \overline{\sigma^2}$ ) under different prior settings, with tight priors being the most informative and loose priors the least. For each simulation and posterior sample, branchwise rates were first centered by subtracting their mean. We estimated centered branchwise rates by taking the median of the centered posterior samples. The solid line represents the position of the true centered branchwise rates, while the shallower, dashed line represents the observed line of best fit for the data under each prior setting. Note that tighter, more informative priors result in shallower best fit lines due to increased shrinkage of branchwise rate estimates.

Table 1D.1 Median absolute errors of rate variance posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	0.46	1.58	4.14	0.70	1.54	3.49	0.79	2.30	3.97
	0	0.48	1.70	3.07	0.84	1.67	2.81	0.94	2.21	3.40
	4	0.52	1.62	3.39	0.82	1.72	2.79	0.93	2.27	3.08
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	0.43	1.52	3.98	0.68	1.53	3.51	0.79	2.33	4.04
	0	0.45	1.71	3.04	0.81	1.65	2.80	0.95	2.20	3.34
	4	0.51	1.63	3.50	0.83	1.72	2.88	0.94	2.34	3.00
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	0.40	1.52	4.14	0.69	1.53	3.66	0.79	2.28	3.95
	0	0.47	1.74	3.10	0.84	1.69	2.73	0.96	2.20	3.46
	4	0.52	1.65	3.73	0.83	1.72	2.82	0.94	2.23	3.09

Table 1D.2 Median absolute errors of trend posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
$\sigma_0^2$ prior shifted by -3										
$\mu_{\sigma^2} =$	-4	2.03	2.34	2.68	1.32	1.57	1.65	1.35	1.64	1.64
	0	1.28	1.32	1.07	1.55	2.22	2.11	1.65	2.45	2.53
	4	1.34	2.22	2.18	2.75	2.33	2.78	4.24	2.88	3.86
$\sigma_0^2$ prior shifted by 0										
$\mu_{\sigma^2} =$	-4	1.63	1.88	2.16	1.30	1.60	1.61	1.32	1.61	1.66
	0	0.91	1.07	0.94	1.54	2.21	2.05	1.64	2.43	2.51
	4	2.06	3.04	2.97	2.61	2.30	2.77	4.07	2.85	3.82
$\sigma_0^2$ prior shifted by 3										
$\mu_{\sigma^2} =$	-4	1.28	1.50	1.69	1.32	1.58	1.62	1.34	1.62	1.62
	0	0.88	1.16	1.12	1.51	2.14	1.98	1.64	2.50	2.50
	4	2.94	3.91	3.77	2.51	2.35	2.68	4.15	2.78	3.68

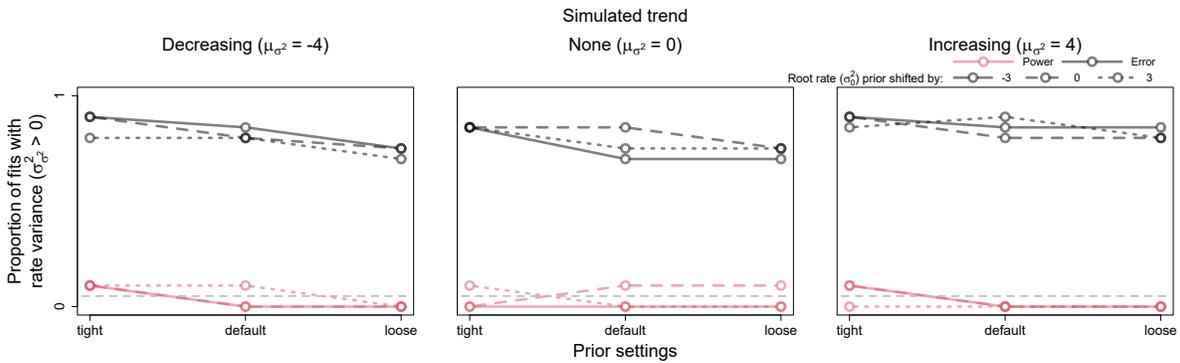


Figure 1D.5 Power and error rates for the rate variance parameter ( $\sigma_{\sigma^2}^2$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for rate variance significantly greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of prior settings, with tight priors being the most informative and loose priors the least. Results are also shown for fits with the location of the root rate ( $\sigma_0^2$ ) prior shifted by -3 (solid lines), 0 (dashed lines), and 3 (dotted lines) from the default setting.

Table 1D.3 Median absolute errors of branchwise rate posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	0.53	0.87	0.98	0.48	0.83	0.90	0.50	0.86	0.91
	0	0.44	0.76	0.94	0.52	0.83	1.01	0.54	0.87	1.07
	4	0.46	0.83	0.93	0.64	0.87	1.02	0.82	0.95	1.16
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	0.47	0.83	0.94	0.48	0.83	0.90	0.49	0.86	0.91
	0	0.40	0.73	0.95	0.51	0.82	1.01	0.53	0.87	1.07
	4	0.52	0.88	0.99	0.63	0.87	1.01	0.80	0.95	1.16
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	0.43	0.79	0.92	0.48	0.82	0.90	0.50	0.86	0.91
	0	0.39	0.73	0.97	0.51	0.82	1.01	0.54	0.88	1.06
	4	0.61	0.95	1.06	0.62	0.87	1.00	0.81	0.94	1.14

Table 1D.4 Median absolute errors of root rate posteriors (i.e., median absolute difference between posterior samples and their true, simulated values, a measure of posterior distribution accuracy), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	1.78	2.54	1.97	1.06	1.41	1.39	1.07	1.45	1.40
	0	1.18	1.24	1.18	1.38	1.66	1.82	1.45	1.84	2.20
	4	1.08	1.36	2.16	2.49	1.83	2.49	3.87	2.39	3.43
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	1.38	1.88	1.40	1.04	1.41	1.36	1.06	1.41	1.41
	0	0.82	0.82	1.18	1.34	1.64	1.79	1.45	1.84	2.17
	4	1.71	2.17	3.05	2.39	1.81	2.50	3.73	2.33	3.36
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	1.02	1.33	1.09	1.06	1.38	1.39	1.06	1.41	1.38
	0	0.79	0.88	1.64	1.33	1.56	1.76	1.44	1.87	2.21
	4	2.54	3.07	4.15	2.27	1.83	2.41	3.80	2.27	3.27

Table 1D.5 Breadths of rate variance posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
				$\sigma_0^2$ prior shifted by -3						
$\mu_{\sigma^2} =$	-4	2.31	8.74	11.61	3.83	10.48	13.13	4.86	14.17	15.74
	0	2.42	6.13	10.76	4.24	8.76	12.73	5.33	11.29	16.48
	4	2.40	7.14	11.94	3.89	9.60	13.33	4.81	12.18	16.82
				$\sigma_0^2$ prior shifted by 0						
$\mu_{\sigma^2} =$	-4	2.20	7.73	11.23	3.84	10.45	13.05	4.60	14.78	15.98
	0	2.23	6.34	10.18	4.21	8.42	13.17	5.14	11.20	16.64
	4	2.40	7.04	11.75	3.93	9.44	13.65	4.78	12.35	16.75
				$\sigma_0^2$ prior shifted by 3						
$\mu_{\sigma^2} =$	-4	2.10	7.82	11.03	3.82	10.31	12.66	4.74	14.44	16.06
	0	2.26	6.10	10.49	4.02	8.34	12.68	5.21	11.50	16.57
	4	2.50	6.90	12.57	4.04	9.28	13.47	4.90	12.18	17.02

Table 1D.6 Breadths of trend posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

		tight priors			default priors			loose priors		
$\sigma_{\sigma^2}^2 =$		0	3	6	0	3	6	0	3	6
					$\sigma_0^2$ prior shifted by -3					
$\mu_{\sigma^2} =$	-4	3.62	4.34	4.61	4.87	6.18	6.68	4.99	6.70	7.08
	0	4.42	4.95	4.80	6.77	8.67	8.60	7.53	9.72	10.23
	4	4.90	5.19	5.38	12.14	10.74	12.73	21.26	15.61	21.57
					$\sigma_0^2$ prior shifted by 0					
$\mu_{\sigma^2} =$	-4	3.63	4.40	4.64	4.81	6.25	6.73	4.97	6.72	7.04
	0	4.23	4.84	4.76	6.77	8.46	8.52	7.44	9.90	10.47
	4	4.64	5.00	5.18	11.57	10.51	12.32	19.99	15.26	20.95
					$\sigma_0^2$ prior shifted by 3					
$\mu_{\sigma^2} =$	-4	3.64	4.37	4.65	4.85	6.23	6.69	4.90	6.64	6.94
	0	4.22	4.66	4.66	6.81	8.23	8.52	7.36	10.13	10.61
	4	4.63	4.90	5.20	11.56	10.28	11.82	19.62	15.68	19.45

Table 1D.7 Breadths of branchwise rate posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	2.01	3.06	3.21	2.33	3.36	3.49	2.41	3.61	3.67
	0	2.05	2.73	3.37	2.52	3.28	3.85	2.65	3.52	4.20
	4	2.15	2.91	3.39	3.10	3.59	4.14	4.32	4.25	5.26
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	1.98	3.00	3.18	2.33	3.36	3.49	2.41	3.61	3.66
	0	2.04	2.72	3.33	2.51	3.26	3.83	2.66	3.54	4.21
	4	2.14	2.89	3.37	3.03	3.57	4.10	4.11	4.24	5.23
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	1.97	2.98	3.17	2.33	3.36	3.49	2.41	3.61	3.66
	0	2.04	2.70	3.36	2.49	3.22	3.83	2.63	3.57	4.21
	4	2.15	2.90	3.41	3.03	3.53	4.06	4.07	4.26	5.03

Table 1D.8 Breadths of root rate posteriors (i.e., the difference between the 97.5% and 2.5% quantiles of posterior samples, a measure of posterior distribution precision), averaged across replicates for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
				$\sigma_0^2$ prior shifted by -3						
$\mu_{\sigma^2} =$	-4	2.91	3.76	3.85	4.03	5.10	5.54	4.13	5.69	5.86
	0	3.86	4.21	4.34	5.94	7.34	7.55	6.52	8.47	9.18
	4	4.33	4.56	4.83	10.91	9.50	11.60	19.65	14.21	19.87
				$\sigma_0^2$ prior shifted by 0						
$\mu_{\sigma^2} =$	-4	2.92	3.65	3.84	4.03	5.22	5.51	4.17	5.52	5.80
	0	3.71	4.10	4.20	5.98	7.21	7.37	6.61	8.51	9.39
	4	4.17	4.43	4.78	10.40	9.38	11.22	18.51	13.81	19.42
				$\sigma_0^2$ prior shifted by 3						
$\mu_{\sigma^2} =$	-4	3.01	3.63	3.94	4.06	5.14	5.59	4.12	5.66	5.76
	0	3.73	4.09	4.29	5.87	7.10	7.40	6.41	8.71	9.39
	4	4.13	4.41	4.80	10.25	9.10	10.72	18.05	14.09	17.82

Table 1D.9 Coverage of rate variance posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	—	1.00	0.70	—	1.00	1.00	—	1.00	0.90
	0	—	0.90	0.80	—	1.00	1.00	—	1.00	1.00
	4	—	1.00	0.90	—	1.00	1.00	—	1.00	0.90
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	—	1.00	0.70	—	1.00	1.00	—	1.00	0.90
	0	—	0.90	0.70	—	1.00	1.00	—	1.00	1.00
	4	—	1.00	0.80	—	1.00	0.90	—	1.00	0.90
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	—	1.00	0.60	—	1.00	1.00	—	1.00	0.90
	0	—	0.90	0.70	—	1.00	1.00	—	1.00	1.00
	4	—	1.00	0.80	—	1.00	1.00	—	1.00	0.90

Table 1D.10 Coverage of trend posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	0.50	0.40	0.40	1.00	0.90	1.00	1.00	1.00	1.00
	0	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	0.90
	4	1.00	0.70	0.80	1.00	1.00	1.00	0.90	1.00	0.90
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	0.50	0.70	0.60	1.00	1.00	1.00	1.00	1.00	1.00
	0	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.90	0.90
	4	0.80	0.10	0.30	1.00	1.00	1.00	1.00	1.00	0.90
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	0.70	0.70	0.80	1.00	0.90	1.00	1.00	1.00	1.00
	0	1.00	1.00	1.00	1.00	0.90	0.90	1.00	0.90	0.90
	4	0.10	0.00	0.10	1.00	1.00	1.00	1.00	1.00	0.90

Table 1D.11 Coverage of branchwise rate posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	0.91	0.89	0.86	0.98	0.95	0.94	0.98	0.96	0.95
	0	0.99	0.93	0.91	1.00	0.96	0.94	1.00	0.97	0.94
	4	1.00	0.91	0.92	0.99	0.96	0.96	0.97	0.97	0.95
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	0.94	0.91	0.87	0.98	0.95	0.94	0.98	0.96	0.95
	0	1.00	0.93	0.90	1.00	0.96	0.93	1.00	0.97	0.95
	4	0.97	0.87	0.89	0.99	0.96	0.95	0.98	0.97	0.96
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	0.97	0.93	0.88	0.98	0.95	0.94	0.98	0.96	0.95
	0	1.00	0.94	0.89	1.00	0.97	0.94	1.00	0.97	0.95
	4	0.89	0.83	0.86	0.99	0.96	0.95	0.98	0.97	0.96

Table 1D.12 Coverage of root rate posteriors (i.e., proportion of times the true, simulated value is greater than the 2.5% posterior distribution quantile and less than the 97.5% quantile) for each simulated trait evolution scenario and prior settings.  $\sigma_{\sigma^2}^2$  and  $\mu_{\sigma^2}$  indicate the true, simulated values of rate variance and trend parameters, respectively, while  $\sigma_0^2$  prior shifts refer to alteration of the root rate prior location.

$\sigma_{\sigma^2}^2 =$	tight priors			default priors			loose priors			
	0	3	6	0	3	6	0	3	6	
	$\sigma_0^2$ prior shifted by -3									
$\mu_{\sigma^2} =$	-4	0.40	0.30	0.60	1.00	0.80	1.00	0.90	0.90	1.00
	0	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00	0.90
	4	1.00	1.00	0.60	1.00	1.00	1.00	1.00	1.00	1.00
	$\sigma_0^2$ prior shifted by 0									
$\mu_{\sigma^2} =$	-4	0.60	0.60	0.80	1.00	0.80	1.00	1.00	1.00	1.00
	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.80	0.40	0.60	1.00	1.00	1.00	1.00	1.00	1.00
	$\sigma_0^2$ prior shifted by 3									
$\mu_{\sigma^2} =$	-4	0.90	0.80	0.90	1.00	0.80	1.00	1.00	0.80	1.00
	0	1.00	1.00	0.60	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.10	0.10	0.10	1.00	1.00	1.00	1.00	1.00	1.00

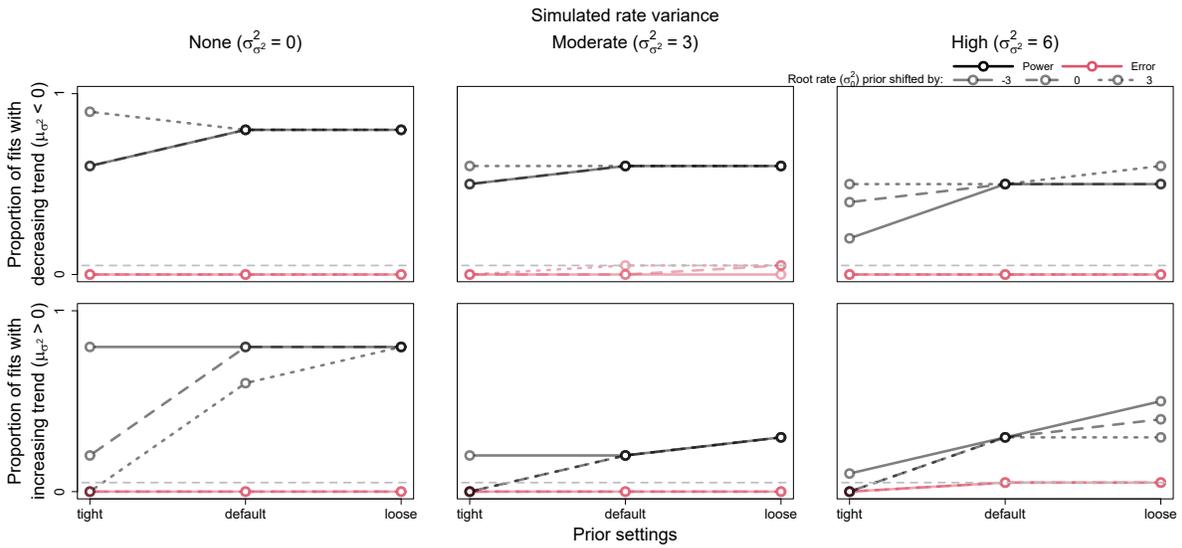


Figure 1D.6 Power and error rates for the trend parameter ( $\mu_{\sigma^2}$ ). Lines depict changes in the proportion of model fits that correctly showed evidence for trends significantly less and greater than 0 (i.e., power, in black) and incorrectly showed evidence (i.e., error, in light red) as a function of prior settings, with tight priors being the most informative and loose priors the least. Results are also shown for fits with the location of the root rate ( $\sigma_0^2$ ) prior shifted by -3 (solid lines), 0 (dashed lines), and 3 (dotted lines) from the default setting.

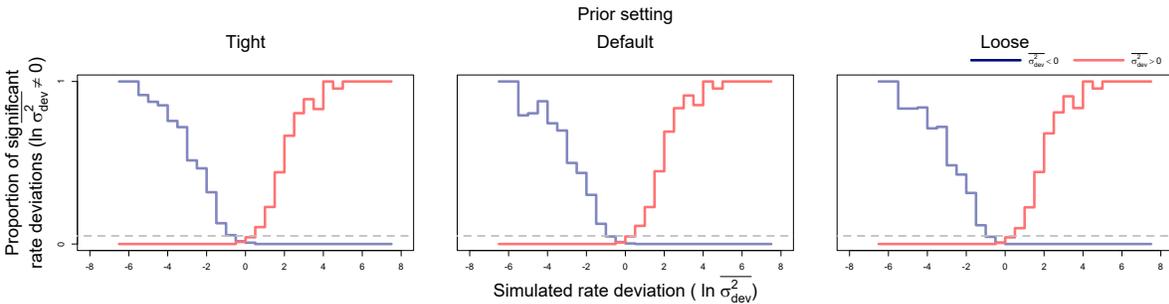


Figure 1D.7 Power and error rates for branchwise rate parameters ( $\ln \overline{\sigma^2}$ ) under different prior settings. Lines depict changes in proportions of branchwise rates considered anomalously slow (in dark blue) or fast (in light red) as a function of simulated rate deviations ( $\ln \overline{\sigma_{dev}^2}$ ). These results combine all fits to simulated data that detected rate variance ( $\sigma_{\sigma^2}^2$ ) significantly greater than 0. The proportions are equivalent to power when the detected rate deviation is of the same sign as the true, simulated deviation (left of 0 for anomalously slow rates in dark blue and right for anomalously fast rates in light red), and to error rate when the detected and true rate deviations are of opposite signs. Here, significant rate deviations for simulated rate deviations that are exactly 0 are considered errors regardless of sign.

## CHAPTER 2

### STOCHASTIC CHARACTER MAPPING OF CONTINUOUS TRAITS ON PHYLOGENIES

#### 2.1 Introduction

A central challenge in macroevolutionary biology is inferring how phenotypes evolve from limited samples of living and/or fossilized organisms. Accordingly, evolutionary biologists have long practiced “ancestral state reconstruction” (ASR)—estimating the unobserved phenotypes of (typically extinct) lineages based on observed phenotypes in their relatives (Dobzhansky and Sturtevant, 1938; Sanger et al., 1955; Pauling et al., 1963; Witmer, 1995; Schultz et al., 1996; Sumrall and Brochu, 2008). With the development of phylogenetic comparative methods that provide a rigorous statistical framework for performing ASR under various trait evolution models, ASR has become ubiquitous in macroevolutionary research (Schluter et al., 1997; Groussin et al., 2016; Joy et al., 2016). One technique for performing ASR, called stochastic character mapping (or “simmapping”), is particularly popular, allowing researchers to randomly sample evolutionary histories (often termed “simmaps”) of a character on a phylogeny according to the probability of a given reconstruction under some trait evolution model (Nielsen, 2002; Huelsenbeck et al., 2003; Bollback, 2006). By sampling hundreds or thousands of simmaps, researchers can generate distributions of possible evolutionary histories which may be used in various macroevolutionary analyses—for example, determining the timing/frequency of evolutionary events (e.g., Baliga and Law, 2016; Tornabene et al., 2016; Freyman and Höhna, 2019; Hughes et al., 2021; Landis et al., 2021; Siqueira et al., 2023) or how past life history/environmental factors affected evolution (e.g., de Alencar et al., 2017; Borstein et al., 2019; Burns and Bloom, 2020; Fabre et al., 2020; Rincon-Sandoval et al., 2020; Drury et al., 2021; Nations et al., 2021; Friedman and Muñoz, 2023). In this way, simmaps allow researchers to flexibly conduct analyses over a range of possible histories and effectively account for the inherently incomplete and uncertain knowledge of the evolutionary past in most macroevolutionary studies.

Unfortunately, while simmaps have revolutionized statistical pipelines for studying macroevo-

lution, current simmapping implementations are limited to sampling histories of discrete variables. The lack of simmapping methods for continuous variables constrains approaches for analyzing macroevolutionary data and leaves a conspicuous methodological gap in the field. For example, testing whether the tempo and/or mode of trait evolution varies according to some “explanatory factor” like habitat or diet is generally straight-forward provided the factor’s entire evolutionary history is explicitly known (e.g., Revell, 2013; Clavel et al., 2015; Beaulieu and O’Meara, 2023). When a factor’s history is unknown (as is generally the case), researchers are currently unable to generate simmaps of continuous factor histories to use in their analyses like they would for discrete factors. Consequently, testing how continuously-varying factors like body size (e.g., Friedman et al., 2019), generation time (e.g., Gingerich, 2001), or climatic niche (e.g., Tribble et al., 2023) affect trait evolution processes is much more difficult, often requiring the development of novel methods and/or analysis pipelines tailored for testing specific hypotheses (Cooper and Purvis, 2009; Hansen et al., 2008; Felsenstein, 2012; Baker et al., 2015; Weir and Lawson, 2015; Clavel and Morlon, 2017; Hansen et al., 2022; Boyko et al., 2023a; Tribble et al., 2023; Uyeda et al., 2021).

Here, we introduce an efficient method for simmapping continuous variables under Brownian motion models of evolution. These continuous simmaps, or “contsimmaps” for short, may be used to directly visualize and analyze the dynamics of phenotypic evolution inferred under trait evolution models, determine the timing and phylogenetic locations of major evolutionary transitions in continuous phenotypes while incorporating uncertainty, and explore how continuous factors may have affected evolutionary processes. First, we present an algorithm for generating contsimmaps by sampling values of continuous variables at arbitrary points on a phylogeny conditional on observed values under a given Brownian motion model. Second, to showcase potential uses of contsimmaps, we outline a general approach for inferring relationships between aspects of continuous trait evolution processes and continuous factors based on contsimmapped factor histories. Using an extensive simulation study, we verify the proposed approach’s ability to accurately infer relationships between continuous factors and rates of trait evolution. We go on to apply this pipeline

to an empirical case study, asking whether divergence in plant height is associated with variation in rates of leaf and flower trait evolution in a clade of *Eucalyptus* trees that range from  $\sim 1$  to 100 meters tall (Brooker et al., 2015; Thornhill et al., 2019; Falster et al., 2021).

## 2.2 Materials and Methods

We designed *contsim* to be a flexible framework for sampling evolutionary histories of continuous variables on a phylogeny under Brownian motion models, analogous to conventional discrete *simmap* methods that sample evolutionary histories of discrete variables under continuous time Markov chain models. An important difference between *contsim* and conventional *simmap* is how evolutionary histories are represented. Conventional *simmap* assign parts of a phylogeny to regimes representing different values of the discrete variable, providing reconstructed evolutionary histories in continuous time. This is impractical for continuous variables evolving under Brownian motion models, which have infinitely many possible values and constantly change over time. Accordingly, *contsim* instead sample values of continuous variables at a finite number of time points evenly distributed across a phylogeny, with the number of time points controlled by a user-specified “resolution”. We implemented this framework in an R package called *contsimmap*, which supports *contsim* under a flexible Brownian motion modeling framework capable of accommodating multiple correlated variables, multiple measurements per tip and/or internal node with associated intraspecific variation/measurement error, missing measurements, and multiple evolutionary trends/rates that differ according to regimes mapped onto a phylogeny. The R package additionally provides tools for transforming, summarizing, and visualizing *contsim* (Fig. 2.1), as well as fitting Brownian motion models with parameters dependent on *contsim* variables.

### 2.2.1 Generating *Contsim* Maps

Here we present a general algorithm for sampling phenotypic values across a dense set of time points on a phylogeny given measurements associated with its nodes, “node error” variances, and a Brownian motion model with potentially regime-dependent evolutionary trends and rates (notably, such regimes must be mapped onto the phylogeny *a priori*). Assume we have  $e$  edges in

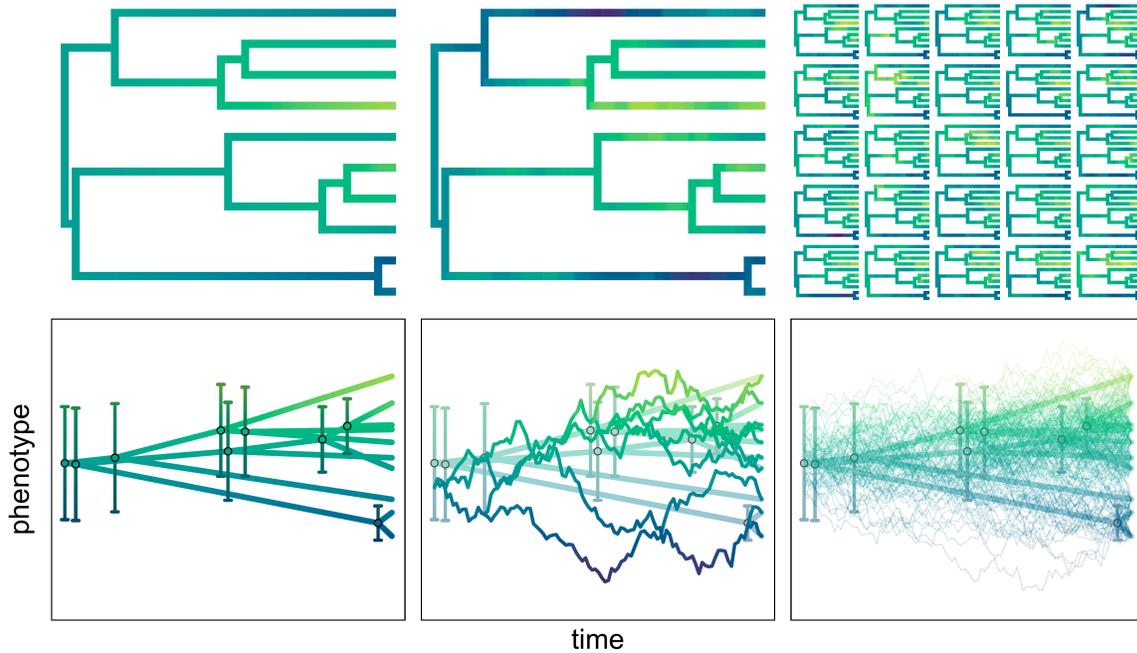


Figure 2.1 Phylogram and phenogram-based visualizations of contsimmaps, colored according to reconstructed phenotypic values. Left: conventional ancestral state reconstructions of a continuously-varying phenotype under a Brownian motion model. The top phylogram is colored according to mean phenotypic estimates, which are also depicted in the phylogram below along with error bars representing 95% confidence intervals on the estimates at each node. Middle: a single contsimmap generated using the same data and Brownian motion model. Right: a sample of 25 contsimmaps representing the overall distribution of generated contsimmaps. The bottom middle and right phylograms also include conventional ancestral state reconstructions for reference.

our phylogeny (including a 0-length root edge indexed 0) with  $s$  mapped regimes and data on  $m$  continuous phenotypic variables. Let  $\tau$  be an  $e$ -length vector of edge lengths assumed to be in units of time. To outline our algorithm, it is useful to focus on the quantities associated with an individual edge denoted  $i$ . First, we define two sets of time points along edge  $i$ :  $n_i$  “interpolant” points at equally-spaced times  $t_i$ , excluding times corresponding to  $i$ ’s ancestral/descendant nodes, and  $n_i^*$  “critical” points at (strictly increasing) times  $t_i^*$ , including times corresponding to  $i$ ’s descendant node (the ancestral node is excluded because it is equivalent to the descendant node of the edge ancestral to  $i$ ) as well as any regime shifts along  $i$ . The number of interpolant points,  $n_i$ , is chosen by rounding  $\frac{\tau_i \xi}{T} - 1$  to the next largest integer, where  $T$  represents the total height of the phylogeny and  $\xi$  a user-specified resolution. Specifically,  $\xi$  corresponds to approximate number of time

points across the height of a phylogeny by defining a “target” time interval for all edges given by  $\frac{T}{\xi}$ . Now let  $X_i$  and  $X_i^*$  be  $n_i \times m$  and  $n_i^* \times m$  matrices, respectively, of phenotypic values at each time point along  $i$ . Further, let  $r_i^*$  be an  $n_i^*$ -length vector of regimes associated with the preceding critical time interval for each entry of  $t_i^*$ . Edge  $i$  may also have measurements associated with its descendant node: let  $Y_i$  be an  $o_i \times m$  matrix of these measurements (in the case that there are no measurements,  $o_i = 0$ ).

Our algorithm samples values of  $X$  and  $X^*$  given the phenotypic measurements/regime mappings for each edge described above ( $Y$  and  $r$ , respectively),  $e$   $m \times m$  variance-covariance matrices describing error in measurements at descendant nodes for each edge ( $\Gamma$ , commonly termed “node error”),  $s$   $m$ -length vectors describing deterministic changes in phenotypes per unit time for each regime ( $\mu$ , commonly termed “evolutionary trends”), and  $s$   $m \times m$  matrices describing stochastic changes in phenotypes per unit time for each regime ( $\Sigma$ , commonly termed “evolutionary rate matrices”). To do this, we additionally keep track of the  $n_i^*$  variance-covariance matrices for each edge  $i$ , denoted  $V_i^*$ , which describe the uncertainty of phenotypic estimates at critical time points. Below, we adopt a general notation of using subscripts to denote edge/regime indices first, followed by time points next (if applicable), and lastly specific phenotypic variables. For example,  $X_i^*$  refers to the matrix of phenotypic values at critical time points along the  $i$ th edge,  $X_{i,j}^*$  to its  $j$ th row (corresponding to the  $j$ th entry of  $t_i^*$  or time  $t_{i,j}^*$ ), and  $X_{i,j,k}^*$  to  $k$ th value in this row (corresponding to the  $k$ th phenotypic variable). Similarly,  $\mu_i$  would refer to trends in evolution for the  $i$ th regime and  $\mu_{i,k}$  to the trend for the  $k$ th phenotypic variable specifically.

Our algorithm, similarly to other rapid algorithms for ASR under Brownian motion models (e.g., Goolsby, 2017; Hassler et al., 2022), consists of three main steps: 1) traversing the phylogeny from tips to root (i.e., in postorder), 2) handling phenotypic values at the root, and 3) traversing the phylogeny from root to tips (i.e., in preorder). At a high level, the first step calculates initial estimates of mean phenotypic values and associated uncertainty at critical time points along all edges (i.e.,  $X^*$  and  $V^*$ ) based *only* on the descendants of each edge. Notably, because the root of the phylogeny only has descendants by definition, this first step already returns final estimates

of mean phenotypic values at the root and associated uncertainty (i.e.,  $X_{0,1}^*$  and  $V_{0,1}^*$ ). Thus, the second step simply consists of sampling phenotypic values at the root based on the multivariate normal distribution defined by  $X_{0,1}^*$  and  $V_{0,1}^*$ . The most complex step is the third and final one, which updates  $X^*$  and  $V^*$  based on the *non*-descendants of each edge, uses the updated  $X^*$  and  $V^*$  to sample phenotypic values at critical time points, and finally samples values at interpolant time points conditional on the sampled values at critical time points. We describe these steps in more detail below:

1) Complete a postorder traversal over all edges. For each edge denoted  $i$  with  $p_i$  immediate descendant edges:

1a) Initialize calculations for edge  $i$  by defining  $Z$  as an  $(o_i + p_i) \times m$  matrix and setting the first  $o_i$  rows of  $Z$  to  $Y_i$ . Similarly, define  $W$  as  $o_i + p_i$   $m \times m$  matrices and set the first  $o_i$  matrices to  $\Gamma_i$ . Any missing entries in  $Z$  are set to 0, and corresponding diagonal and off-diagonal entries of the associated  $W$  matrices are set to  $\infty$  and 0, respectively (Hassler et al., 2022). Accordingly, if edge  $i$  is a tip with no associated measurements (i.e.,  $o_i + p_i = 0$ ), create a “dummy observation” by setting  $o_i$  to 1 and initializing  $Z$  as a  $1 \times m$  matrix of 0s and  $W$  as a single  $m \times m$  diagonal matrix with  $\infty$  along the diagonal. Otherwise, if edge  $i$  has immediate descendant edges (i.e., not a tip with  $p_i > 0$ ), for each descendant edge indexed  $l$  from 1 to  $p_i$ :

$$Z_{o_i+l} = X_{d_l,1}^* - \mu_{r_{d_l,1}}^* (t_{d_l,1}^* - t_{i,n_i}^*) \quad (1)$$

$$W_{o_i+l} = V_{d_l,1}^* + \Sigma_{r_{d_l,1}}^* (t_{d_l,1}^* - t_{i,n_i}^*) \quad (2)$$

Where  $Z_l$  denotes the  $l$ th row of  $Z$ ,  $W_l$  denotes the  $l$ th matrix of  $W$ , and  $d$  is a  $p_i$ -length vector of indices corresponding to the edges immediately descending from  $i$ .

1b) Calculate the uncertainty and mean of the phenotypic values at  $i$ 's descendant node:

$$V_{i,n_i^*}^* = \left( \sum_{l=1}^{o_i+p_i} W_l^{-1} \right)^{-1} \quad (3)$$

$$X_{i,n_i^*}^* = V_{i,n_i^*}^* \left( \sum_{l=1}^{o_i+p_i} Z_l W_l^{-1} \right) \quad (4)$$

Where  $W_l^{-1}$  specifically denotes the “pseudo-inverse” of  $W_l$  (Hassler et al., 2022). Notably, Eq. (4) may be undefined when one or more measurements are missing or assumed to be exact, which can cause entries of  $V_{i,n_i^*}^*$  and  $W^{-1}$ , respectively, to be  $\infty$ . The former case can be solved by simply defining  $\infty * 0 = 0$  (Hassler et al., 2022). For the latter case, we can generalize Eq. (4) by thinking of the expression as a weighted average of phenotypic value vectors  $Z_l$  with weights given by  $V_{i,n_i^*}^* W_l^{-1}$ . From this perspective, infinite entries along the diagonals of  $W^{-1}$  correspond to “infinitely large” weights in this average. Accordingly, we define a “normalization” procedure for matrices in  $W^{-1}$ : let  $W_{.,k,k}^{-1}$  refer to the diagonal entries in the  $k$ th row and column across all  $W^{-1}$  matrices. Then, for each phenotypic variable  $k$ , check whether  $W_{.,k,k}^{-1}$  contains  $\infty$ . If it does, set finite and infinite elements of  $W_{.,k,k}^{-1}$  to 0 and 1, respectively, and change corresponding off-diagonal entries to 0. If we denote these normalized matrices  $\tilde{W}^{-1}$ , we can write a more “robust” version of Eq. (4) which can handle exact trait measurements:

$$X_{i,n_i^*}^* = \left( \sum_{l=1}^{o_i+c_i} \tilde{W}_l^{-1} \right)^{-1} \left( \sum_{l=1}^{o_i+c_i} Z_l \tilde{W}_l^{-1} \right) \quad (5)$$

Intuitively, this means that inexact measurements are “superseded” by exact measurements, which themselves are simply averaged together. We note that this solution is more pragmatic than rigorous when there are two or more exact measurements—paradoxically, averages of multiple exact measurements are still considered exact under this framework. While conceptually problematic, situations where multiple measurements are assumed to be exact may arise in practice due to estimation of 0-length edges

during phylogenetic inference and/or simplifying comparative analyses by assuming no measurement error. In any case, the above expression allows our algorithm to yield predictable, defined results in such cases.

- 1c) If edge  $i$  includes any regime shifts, calculate expected phenotypic values and uncertainty at each preceding critical time point. For each critical point  $j$  from  $n_i^* - 1$  to 1:

$$V_{i,j}^* = V_{i,j+1}^* + \Sigma_{r_{i,j+1}^*} (t_{i,j+1}^* - t_{i,j}^*) \quad (6)$$

$$X_{i,j}^* = X_{i,j+1}^* - \mu_{r_{i,j+1}^*} (t_{i,j+1}^* - t_{i,j}^*) \quad (7)$$

- 2) Simulate phenotypic values at the root of the phylogeny by sampling new values for  $X_{0,1}^*$  from a multivariate normal distribution with mean  $X_{0,1}^*$  and variance-covariance matrix  $V_{0,1}^*$ .
- 3) Complete the preorder traversal over all edges except the root ( $i = 0$ ). For each critical time point  $j$  along edge  $i$ :

- 3a) Initialize calculations by defining  $Z$  as an  $m$ -length row vector and  $W$  as a single  $m \times m$  matrix.  $Z$  and  $W$  represent counterparts to  $X_{i,j}^*$  and  $V_{i,j}^*$ . While  $X_{i,j}^*$  and  $V_{i,j}^*$  define the distribution of phenotypic values at the  $j$ th critical time point along edge  $i$  based *only* on all descendants of edge  $i$ ,  $Z$  and  $W$  are based *only* on all *non*-descendants of  $i$ . Because this is a preorder traversal, phenotypic values at the immediately previous critical time point have already been simulated and  $Z$  and  $W$  are given by:

$$Z = X_{i,j-1}^* + \mu_{r_{i,j}^*} (t_{i,j}^* - t_{i,j-1}^*) \quad (8)$$

$$W = \Sigma_{r_{i,j}^*} (t_{i,j}^* - t_{i,j-1}^*) \quad (9)$$

In the case  $j - 1 = 0$ , we must instead use the trait values and time associated with the last critical time point along the edge directly ancestral to  $i$ , denoted  $a$ . Specifi-

cally,  $X_{i,j-1}^*$  and  $t_{i,j-1}^*$  are substituted with  $X_{a,n_a}^*$  and  $t_{a,n_a}^*$ , respectively, in the expression above.

- 3b) Update the uncertainty and mean of the phenotypic values at the  $j$ th critical time point along edge  $i$ :

$$V_{i,j}^* = \left( W^{-1} + V_{i,j}^{*-1} \right)^{-1} \quad (10)$$

$$X_{i,j}^* = \left( \tilde{W}^{-1} + \tilde{V}_{i,j}^{*-1} \right)^{-1} \left( Z\tilde{W}^{-1} + X_{i,j}^* \tilde{V}_{i,j}^{*-1} \right) \quad (11)$$

Where  $\tilde{W}^{-1}$  and  $\tilde{V}_{i,j}^{*-1}$  again denote normalized versions of these matrices. This normalization procedure is identical to the one used for Eq. (5), with  $W^{-1}$  and  $V_{i,j}^{*-1}$  treated as a set of two matrices, with one key exception. Notably,  $W^{-1}$  will only contain  $\infty$  when one or more phenotypic variables evolve with a rate of 0 (i.e., at least one diagonal entry of  $\Sigma_{r_{i,j}}^*$  is 0). To keep resulting contsimmaps consistent with this condition, in the case that  $W_{k,k}^{-1}$  and  $(V_{i,j}^{*-1})_{k,k}$  are *both*  $\infty$ , *only*  $W_{k,k}^{-1}$  is set to 1 and  $(V_{i,j}^{*-1})_{k,k}$  is instead set to 0. This critically ensures that phenotypic variables with an evolutionary rate of 0 only exhibit changes over time due to trends defined by  $\mu$ .

- 3c) Simulate phenotypic values at the  $j$ th critical time point along edge  $i$  by replacing  $X_{i,j}^*$  with new values sampled from a multivariate normal distribution with mean  $X_{i,j}^*$  and variance-covariance matrix  $V_{i,j}^*$ .
- 3d) Simulate phenotypic values at all interpolant time points along edge  $i$  lying between critical time points  $j-1$  and  $j$ . For each interpolant time point  $l$  for which  $t_{i,j-1}^* < t_{i,l} < t_{i,j}^*$  (if  $j-1 = 0$ , swap  $t_{i,j-1}^*$  for  $t_{a,n_a}^*$  as above), sample  $X_{i,l}$  from a multivariate normal distribution with mean  $Z$  and variance-covariance matrix  $W$  given by:

$$Z = \frac{(X_{i,j}^* - X_{i,l-1})(t_{i,l} - t_{i,l-1})}{t_{i,j}^* - t_{i,l-1}} + X_{i,l-1} \quad (12)$$

$$W = \frac{(t_{i,j}^* - t_{i,l})(t_{i,l} - t_{i,l-1})}{t_{i,j}^* - t_{i,l-1}} \Sigma_{r_{i,j}}^* \quad (13)$$

In the case  $l - 1 = 0$ , we must instead use the phenotypic values and time associated with the immediately previous critical time point  $j - 1$ . Specifically,  $X_{i,l-1}$  and  $t_{i,l-1}$  are swapped with  $X_{i,j-1}^*$  and  $t_{i,j-1}^*$ , respectively, in the expressions above (if  $j - 1 = 0$ , these are in turn swapped with  $X_{a,n_a}^*$  and  $t_{a,n_a}^*$ , respectively, as above).

## 2.2.2 Using Contsimmaps to Analyze Factor-Dependent Trait Evolution

Simmaps are particularly useful for “sequential inference” pipelines, whereby the impact of some “explanatory factor” on an evolutionary process is inferred by first generating simmaps of factor histories, then fitting factor-dependent evolutionary models conditional on the simmaps. For conventional discrete simmaps, arbitrary evolutionary models can be rendered factor-dependent by simply allowing the parameters of the evolutionary model—such as evolutionary rates or trends in the case of Brownian motion models—to vary among lineages in different discrete states (e.g., lineages in mountain versus lowland habitats, annual versus perennial plant lineages; see Revell, 2013 for further description of this approach). For contsimmaps, with an infinite continuum of states, models must instead use “parameter functions” to map factor values to associated parameter values (see FitzJohn, 2010 for an example of this approach). For example, a researcher could test whether increased temperatures are associated with accelerated body size evolution by fitting a Brownian motion model assuming an exponential relationship,  $\sigma^2 = e^{\beta_0 + \beta_1 T}$ , between contsimmapped thermal niche values ( $T$ ) and rates of change in body size over time ( $\sigma^2$ ). Here,  $\beta_0$  and  $\beta_1$  are free parameters inferred by fitting the model to data, with  $\beta_1$  directly quantifying the relationship between temperature and rates.

We implemented tools in our R package for fitting factor-dependent Brownian motion models to trait data conditional on samples of contsimmapped factor histories. Specifically, our implementation constructs a likelihood function for a given trait dataset based on a sample of  $n$  contsimmaps and user-defined parameter functions that map contsimmapped values to evolutionary rates, trends, and/or node error variances. Given estimates for the free parameters in all parameter functions, the outputted likelihood function automatically transforms contsimmapped values into parameter values and uses the pruning algorithm outlined by Hassler et al., 2022 to calculate likelihoods

conditional on each contsimmap. To derive a single overall likelihood,  $\tilde{L}$ , from the  $n$  conditional likelihoods,  $L$ , the likelihood function marginalizes over the contsimmaps by assuming either a “flat” or “nuisance” prior. Under the flat prior, each contsimmap is assumed equally likely, with the overall likelihood given by the average of all conditional likelihoods:

$$\tilde{L} = \frac{1}{n} \sum_{i=1}^n L_i \quad (14)$$

Under the nuisance prior, the conditional likelihoods for each contsimmap are weighted by the probability they gave rise to the trait data among all other contsimmaps and summed (often termed the “FitzJohn root prior” in the context of marginalizing over root states; see FitzJohn et al., 2009):

$$\tilde{L} = \frac{1}{\sum_{i=1}^n L_i} \sum_{i=1}^n L_i^2 \quad (15)$$

Intuitively, the nuisance prior allows the trait data and model to influence which contsimmapped factor histories are considered most likely by taking a weighted average of the conditional likelihoods, with higher weights assigned to contsimmaps that explain the observed data relatively well under a given model. The rough contribution of each contsimmap’s conditional likelihood to the overall likelihood, which we term “normalized conditional likelihoods”, can be quantified as  $L/\sum_{i=1}^n L$  and  $L^2/\sum_{i=1}^n L^2$  in the cases of flat and nuisance priors, respectively.

Notably, the conditional likelihoods under each contsimmap also depend on the trait values at the root of the phylogeny, which are often inferred as additional free parameter in other phylogenetic comparative methods for modeling continuous trait evolution (e.g., Revell, 2012; Pennell et al., 2014; Boucher et al., 2018). However, because conditional likelihoods under any single root trait value will vary drastically across different contsimmaps (e.g., see Figure 4 in Boyko et al., 2023a), we calculate conditional likelihoods for each contsimmap while marginalizing over root trait values by assuming either a flat or nuisance prior over the root trait values as well. Note that the likelihood of observed data under any (potentially multivariate) Brownian motion model conditional on a vector of root trait values  $x$  is given by  $r\Phi(x; \hat{x}, V)$ , where  $r$  is a proportionality

constant (called the “remainder” in Hassler et al., 2022),  $\hat{x}$  and  $V$  denote the expected root trait values and associated uncertainty (in the form of a variance-covariance matrix), and  $\Phi(x; \mu, \Sigma)$  represents the probability density function of a multivariate normal distribution with mean  $\mu$  and variance-covariance matrix  $\Sigma$  evaluated at  $x$ . Because the integral of any probability density function is 1 by definition, the integral of this expression is  $r$ —thus, the overall likelihood for a given contsimmap is  $r$  under a flat prior. Under a nuisance prior, the overall likelihood is equal to the proportionality constant  $r$  multiplied by the integral of the squared multivariate normal probability density function  $(r/\sqrt{|V|}(4\pi)^k)$ , where  $k$  denotes the number of traits/dimensions of the multivariate normal distribution). Assuming a nuisance prior will deflate overall likelihoods conditional on any given contsimmap when root trait values are highly uncertain (i.e.,  $|V|$  is large) and vice versa.

To fit these factor-dependent Brownian motion models to data, optimization or Bayesian inference algorithms may be used to infer values of the free parameters defining parameter functions. Our R package currently provides tools for using the C++ library NLOPT (Johnson, 2021), interfaced through the R package *nloptr* (Ypma et al., 2022), to find free parameter values that maximize the likelihood of an outputted likelihood function. In general, we found that likelihood surfaces under continuous factor-dependent Brownian motion models may be quite complex, exhibiting multiple optima and/or relatively flat “ridges” that present challenges for numerical optimization. While our implementation allows users to apply any algorithm available in the NLOPT library and customize optimizer settings as they see fit, we developed a convenient default optimization procedure, consisting of three phases that leverage the complementary strengths/weaknesses of three distinct algorithms. The initial “warmup” phase uses NLOPT’s gradient-based truncated Newton algorithm (Dembo and Steihaug, 1983) with preconditioning and random restarts (gradients are stochastically approximated using finite differences; see APPENDIX 2B for details), which rapidly improves model fit but often gets stuck on suboptimal peaks of the likelihood surface. The subsequent “exploratory” phase uses NLOPT’s *sbplx* (based on subplex; see Rowan, 1990) to search for higher-likelihood regions in the vicinity of this initial estimate. However, because the *sbplx* algorithm tends to terminate in relatively flat areas and/or saddle points of the likelihood sur-

face rather than true maxima, a final “polish” phase uses NLOPT’s principal axis algorithm (Brent, 2013) to find a local maximum either within or close to the high-likelihood region. Ultimately, this procedure appears to offer a good compromise between robustness and speed, finding and converging on peaks associated with relatively high likelihoods in a practical amount of time. Notably, we found that the algorithms used in the warmup and polish phases do not work well in (and are largely unnecessary for) searching low-dimensional parameter spaces, so our implementation skips these phases by default when fitting models with two parameters or less.

### 2.2.3 Simulation Study

To assess the performance of our approach for inferring relationships between trait evolution processes and continuous factors, we tested whether our method could reliably detect and quantify factor-dependent rates of continuous trait evolution from simulated data. Broadly speaking, we simulated the evolution of a single continuous trait,  $Y$ , with rates depending on a simulated continuous factor, either an “observed” factor  $X_o$  or “unobserved” factor  $X_h$ , and applied our approach to the simulated data to infer relationships between rates of  $Y$  evolution and  $X_o$ . In an empirical context,  $X_h$  represents an unobserved or “hidden” factor that nonetheless affected rates of trait evolution and may thus mislead hypothesis testing for factor-dependent rates (May and Moore, 2020; see also Beaulieu and O’Meara, 2016; Boyko and Beaulieu, 2023; Boyko et al., 2023b; Tribble et al., 2023). After outlining our simulation procedure below, we outline our analysis procedure, which includes a pragmatic technique for constructing additional null models that help account for hidden factors and thereby mitigate their potential effects on hypothesis testing.

We allowed rates of  $Y$  evolution,  $\sigma^2$ , to vary with factor values (denoted  $X$  below) according to one of four parameter functions: 1) a “simple” function whereby rates exponentially increase/decrease, 2) a “threshold” function whereby rates shift between some minimum and maximum value, 3) a “sweetspot” function whereby rates peak or dip around a particular factor value, and 4) a “null” function where rates stay constant. We define the simple function as:

$$\sigma^2 = e^{\beta_0 + \beta_1 X} \tag{16}$$

Where  $\beta_0$  and  $\beta_1$  represent the intercept and slope of the factor-rate relationship on a natural logarithmic scale. This function provides a simple means to test hypotheses that only claim rates tend to increase or decrease in association with some factor like temperature (e.g., Clavel and Morlon, 2017; Slater et al., 2017) or generation time (e.g., Gingerich, 2001). However, it also allows rates of trait evolution to become arbitrarily close to 0 and grow without bound, motivating the threshold function, a logistic curve bounded between strictly positive minimum and maximum rate values:

$$\sigma^2 = e^\alpha \left( \tanh \delta \left( 2 \left( 1 + e^{-2\pi\sqrt{3}\frac{X-\theta}{\omega}} \right)^{-1} - 1 \right) + 1 \right) \quad (17)$$

Here,  $\alpha$  alters the overall scale of rates, specifically corresponding to the natural log of the rate halfway between the minimum and maximum possible rates or “mid rate”, while  $\theta$  determines the factor value or “location” at which rates shift (i.e., the value of  $X$  at which  $\sigma^2 = e^\alpha$ ). We denote  $\delta$ , which controls the direction and magnitude of the rate shift, the “rate deviation”. The fold-difference between minimum and maximum rates is explicitly given by  $e^{2|\delta|}$ , and positive and negative values of  $\delta$  yield upward and downward shifts with increasing factor values, respectively. Lastly,  $\omega$  adjusts the “width” of the shift, with rates roughly reaching their minimum/maximum values at factor values of  $\theta \pm \omega/2$ . Notably, both the simple and threshold functions only allow rates to monotonically increase/decrease with increasing factor values, yet some empirical evidence suggest evolutionary rates may exhibit more complex modal relationships with factors like body size, peaking or dipping at intermediate factor values (Cooper and Purvis, 2009; FitzJohn et al., 2009; Feldman et al., 2016; Amado et al., 2021). Thus, we defined the sweetspot function, whereby factor-rate relationships follow a Gaussian curve of arbitrary height and orientation which takes on strictly positive values:

$$\sigma^2 = e^\alpha \left( \tanh \delta \left( 2e^{-18\left(\frac{X-\theta}{\omega}\right)^2} - 1 \right) + 1 \right) \quad (18)$$

Where  $\alpha$ ,  $\theta$ ,  $\delta$ , and  $\omega$  largely have the same effects and interpretations as they do for the threshold function. Now, however, if  $\delta$  is positive, rates will peak to their maximum at  $\theta$  and

roughly reach their minimum at factor values of  $\theta \pm e^{\omega}/2$ . Conversely, if  $\delta$  is negative, rates will instead dip to their minimum at  $\theta$  and roughly reach their maximum at  $\theta \pm e^{\omega}/2$ . We use these unconventional parameterizations of logistic and Gaussian curves to limit interdependence among parameters in controlling the overall shape of parameter functions, which improves the behavior of numerical optimization routines during model fitting, while also conveniently allowing parameters to take on any value and still form valid factor-rate relationships whereby rates never take on negative values (in practice, we still found it necessary to impose boundaries on parameters to improve model fitting behavior; see below). Lastly, we defined the null parameter function as  $\sigma^2 = e^{\alpha}$  for consistency with threshold and sweetspot functions.

For each simulation, we used *phytools* (Revell, 2012) to simulate an ultrametric, pure-birth phylogeny of height 1 with either 50, 100, or 200 tips. To simulate the continuous factors  $X_o$  and  $X_h$ , we generated two densely-sampled continuous factor histories (with a resolution or  $\xi$  value of 500) evolving under Brownian motion processes with root trait values of 0, no trends, and constant rates of 4. For the trait  $Y$ , we simulated an additional Brownian motion process with starting value 0 and no trends, but with rates varying according one of 12 possible factor-rate relationships which differed both in the overall magnitude (i.e., “strength”) of rate variation as well as how gradually rates changed with respect to factor values (i.e., “width”). Specifically, we used one null parameter function; six “strong” versions of the simple, threshold, or sweetspot functions depending on either  $X_o$  or  $X_h$  whereby rates varied  $\sim 20$ -fold between factor values of -2 and 2; three “weak” versions of the simple, threshold, or sweetspot functions depending on  $X_o$  only whereby rates only varied  $\sim 5$ -fold on the same interval; and two “wide” versions of the threshold or sweetspot functions depending on  $X_o$  only whereby rates again varied  $\sim 20$ -fold but between factor values of -4 and 4. See Table 2A.1 for the specific parameter values used.

Notably, our approach assumes a given factor perfectly predicts rates of trait evolution, which is unlikely for empirical data. To test the robustness of our approach to imperfect correspondences between factors and rates, we additionally modified some simulations by adding “noise” to factor-rate relationships. We added noise by multiplying rate values at each time point with a random

variable sampled from a gamma distribution with shape and rate  $\frac{t}{v}$ , where  $t$  is the length of the time interval preceding a time point. These multipliers represent the average value of a white noise gamma process with mean 1 and unit variance  $v$  over a time interval of length  $t$ . By multiplying the rates over each time increment with these random variables, the simulated trait evolution process is transformed from Brownian motion to variance-gamma (a type of Lévy or “pulsed” trait evolution process explored in some prior work; see Landis et al., 2013; Landis and Schraiber, 2017). This procedure ensures that random noise around rates tends to “average out” over sufficiently long periods of time, such that rates converge to what would be expected under the simulated factor-rate relationship given enough data. Thus, this kind of rate noise only weakens the factor-rate relationship rather than completely altering it. For our simulations, we set  $v$  to 0.05, corresponding to rates ranging between roughly 10%-300% their expected value over a time interval spanning a tenth of the phylogeny’s height, 40%-200% for a third of the height, and 60%-150% for the entire height.

Ultimately, the three phylogeny sizes (50, 100, or 200 tips), 12 possible factor-rate relationships, and presence/absence of noise around factor-rate relationships yield  $3 \times 12 \times 2 = 72$  simulation conditions. We repeated the simulation process 100 times for each condition, yielding a grand total of 7,200 simulations. For the analysis procedure, we retained the phylogeny and tip factor/trait values of  $X_o$  and  $Y$  for each simulation (assuming no node error in  $X_o/Y$  values to render the simulation study more manageable). We analyzed the factor/trait datasets by first fitting Brownian motion models to the  $X_o$  data (assuming no node error or evolutionary trends) and using the fitted model to generate 100 contsimmaps of  $X_o$  with a resolution or  $\xi$  value of 100. We then fit four Brownian motion models to the  $Y$  data conditioned on these contsimmaps (again assuming no node error or evolutionary trends)—three non-null models assuming rates depend on mapped  $X_o$  values through either a simple, threshold, or sweetspot function with unknown free parameters, plus a null model assuming an unknown, constant rate (given by the mid rate/ $\alpha$  parameter). Trait data simulated with rates varying due to noise and/or the hidden factor,  $X_h$ , are likely to exhibit a poor fit to the null model and provide spurious support for non-null models that at least allow rates

to vary across the phylogeny (May and Moore, 2020; see also Beaulieu and O’Meara, 2016; Boyko and Beaulieu, 2023; Boyko et al., 2023b; Tribble et al., 2023). Thus, we fit three additional null Brownian motion models to the  $Y$  data assuming rates depend on a mapped “dummy factor”  $D$ , which is simulated under the Brownian motion model fitted to the  $X_o$  data, through either a simple, threshold, or sweetspot relationship. Because  $D$  exhibits the same evolutionary dynamics as  $X_o$  but with random tip values, support for  $D$ -dependent over  $X_o$ -dependent models provide evidence that rates of  $Y$  evolution vary, but in a way not necessarily related to the observed factor  $X_o$  (see also Tribble et al., 2023, which employs a similar approach in testing for relationships between discrete trait evolution and continuous factors).

We fit all Brownian motion models to simulated data by running the default optimization procedure (see previous section) 10 times from initial parameter values sampled from a uniform distribution spanning from -5 to 5 and retaining the inferred parameter values from whichever optimization run found the highest likelihood. To save on the time needed to run the simulation study, we only allowed the warmup and exploratory/polish phases to run for a maximum of 1,000 and 10,000 iterations, respectively. We always assumed flat priors over root trait values. On the other hand, we assumed nuisance priors over contsimmaps for  $D$ -dependent models and flat priors for all other models. Intuitively, this forces non-null models to “fairly” integrate over probable histories of observed factors while allowing null models to assign higher weights to simulated dummy factor contsimmaps which are particularly likely to explain the trait data. Preliminary tests of our approach showed that this “mixed prior” technique renders null models more competitive with non-null models and greatly reduces false positive errors at a modest cost to statistical power. Preliminary tests also demonstrated that the default optimization procedure sometimes gets stuck in likelihood ridges when fitting threshold and sweetspot function-based models, which may collapse to effectively constant rate models as the location ( $\theta$ ) and/or width ( $\omega$ ) parameters become arbitrarily small/large. To prevent optimization algorithms from spending too much time exploring these regions of parameter space, we imposed boundaries on these parameters by defining a “factor interval” spanning from  $\min X - (\max X - \min X)/2$  to  $\max X + (\max X - \min X)/2$ , where

$X$  represents whatever factor a given parameter function depends on (either the observed factor  $X_o$  or dummy factor  $D$  depending on the model). We constrained the location of inferred shifts and peaks/dips to lie within the factor interval and the corresponding widths (given by  $e^{\omega}$ ) to be between  $1/100^{\text{th}}$  and 5 times the range of the factor interval. We also limited the rate deviation ( $\delta$ ) parameter to be between -10 and 10, as rate deviations with an absolute value of 10 or greater cause minimum and maximum rates to effectively equal 0 and  $2e^{\alpha}$  (i.e., the maximum allowable rate for a given mid rate or  $\alpha$  value), respectively.

For each simulation, we selected the model with lowest Bayesian Information Criterion (BIC; see Schwarz, 1978; Dziak et al., 2020) as the best-fitting model. We chose to use BIC over the more widely used small sample size corrected Akaike Information Criterion (AICc) after preliminary investigations demonstrated that AICc-based model selection exhibits somewhat elevated error rates among simulations with noisy rates on large phylogenies (Tables 2A.2 and 2A.3). BIC penalizes model complexity more harshly than AICc—particularly for large sample sizes—and substantially reduced the error rates of our method at little cost to overall power and accuracy. We calculated error (“false positive”) and power (“true positive”) rates as the percent of simulations for which  $X_o$ -dependent models were selected as the best-fitting model for simulations with constant/ $X_{h_i}$ -dependent rates and  $X_o$ -dependent rates, respectively. Among the simulations yielding true positive results, we additionally calculated “differentiation rates” as the percent for which the best-fitting model assumed the same  $X_o$ -rate relationship used to simulate the data. To further explore how accurately our approach can estimate the relationship between  $X_o$  and rates of  $Y$  evolution, we also calculated BIC weights and generated model-averaged rate estimates across values of  $X_o$  for all simulations with either constant or  $X_o$ -dependent rates, ignoring simulations with  $X_{h_i}$ -dependent rates because they lack a straight-forward expected  $X_o$ -rate relationship to compare with rate estimates. We measured the overall accuracy, bias, and precision of model-averaged rates estimates across values of  $X_o$  by calculating the inverse of the median absolute fold difference between estimated and simulated rates (i.e., the median absolute difference between estimated and simulated rates on the log scale, negated such that higher values correspond to increased accuracy), the per-

cent of estimated rates greater than simulated rates, and the fold difference between the 2.5% and 97.5% quantiles of estimated rates, respectively.

#### 2.2.4 Empirical Example

We applied our approach for inferring continuous factor-dependent rates of trait evolution to test whether rates of phenotypic evolution are associated with size in the *Eucalyptus* subgenus *Eucalyptus*, sometimes called “Monocalypts”, a clade of  $\sim 124$  Australian woody plant species ranging from shrubby “mallees” only reaching a few meters in height (e.g., *E. acies* Brooker, *E. cunninghamii* Sweet, *E. erectifolia* Brooker & Hopper) to gargantuan trees growing to nearly 100 meters in height (e.g., *E. obliqua* L’Hér, *E. regnans* F.Muell.) (Ladiges et al., 2010; Brooker et al., 2015; Thornhill et al., 2019; Falster et al., 2021; Nicolle, 2022). Body size tends to be associated with many aspects of life history, as larger organisms generally live in smaller populations with slower generational turnover (Niklas and Enquist, 2001; White et al., 2007; Sibly and Brown, 2007; Adler et al., 2014; Salguero-Gómez et al., 2016; Bakewell et al., 2020), leading to the hypothesis that evolutionary rates are slower among lineages of larger organisms compared to their smaller relatives. Nonetheless, it remains unclear how rates of phenotypic evolution generally scale with size in many groups, particularly plants (Cooper and Purvis, 2009; Baker et al., 2015; Chira et al., 2018; Friedman et al., 2019; Zimova et al., 2023; see also Lanfear et al., 2013). We obtained a phylogeny of 108 Monocalypt species by pruning the clade out from a recently inferred, time-calibrated phylogeny of over 700 species in *Eucalyptus* and related genera (the “Maximum Likelihood 1” tree in Thornhill et al., 2019). We used the online *Eucalyptus* encyclopedia EUCLID (Brooker et al., 2015) and the AusTraits database (Falster et al., 2021) to gather data on 10 continuous traits for these species, which we categorized into four modules: 1) juvenile leaves consisting of juvenile leaf widths/lengths, 2) adult leaves consisting of petiole lengths and adult leaf widths/lengths, 3) inflorescences consisting peduncle and inflorescence pedicel lengths, and 4) flowers consisting of bud, fruit, and seed lengths. We used the same sources to also aggregate data on height for these species, which acts as a proxy for size and plays the role of a factor in our analyses. All of these traits exhibit substantial within-species variation. Accordingly, trait values for a

given species were generally reported as ranges rather than average values, though only maximum values were reported for some 0-29 species depending on the trait (e.g., “Grows up to 8 meters in height”). The only average trait values reported were bud lengths for 13 species. Ultimately, for each trait, we obtained data for between 102 and 105 out of the 108 species in the phylogeny, with full value ranges (i.e., both minimum and maximum values) known for 73-105 species.

We assumed trait values were log-normally distributed within species with provided minimum and maximum trait values corresponding to the 2.5% and 97.5% quantiles of this distribution, such that averages and associated intraspecific variances (on the natural log scale) are given by the mid-points and  $1/4^2$  the squared differences, respectively, of log-transformed minimum and maximum trait values. Notably, some value ranges for petiole, peduncle, and pedicel lengths had maximum and/or minimum values of 0 in some species, which is undefined on a log scale. Because these traits were measured to a precision of 1 mm (Brooker et al., 2015), we rounded up maximum trait values of 0 to 0.05 mm (i.e., the average assuming a flat prior over measurements from 0 to 1 mm). For minimum trait values of 0, we either rounded them up to 0.05 mm if paired with a non-zero maximum value or treated them as missing if paired with a maximum value also equal to 0. For all cases where only the maximum trait value for a given species was known, we developed a crude but simple procedure for imputing minimum trait values to mitigate bias resulting from comparing average trait values in some species to maximum trait values in others. Specifically, for each trait, we regressed log-transformed minimum trait values on maximum values for all species with fully-known value ranges, and used the fitted relationship to predict minimum trait values for species with known maximum values only. For species with only average bud length measurements available, we simply log-transformed the given averages and treated associated intraspecific variance as an unknown parameter to be estimated during model fitting.

Ultimately, we generated 349 contsimmaps (resolution/ $\xi = 100$ ) of Monocalypt height under an evolving rates or “evorates” model whereby rates of Monocalypt height evolution were allowed to gradually change over time and across lineages (Martin et al., 2023; see APPENDIX 2C for further details). We elected to use this relatively complicated model of height evolution as a constant-rate

Brownian motion model exhibited a rather poor fit to our *Monocalypt* height data by comparison, reflecting substantial variation in rates of height evolution among *Monocalypts*. We then generated 349 corresponding dummy factor contsimmaps by simulating trait evolution with starting values and rates identical to those of the height contsimmaps. Using these contsimmaps, we fit seven multivariate Brownian motion models to each trait module—six models assuming rates depend either on height or the dummy factor through a simple, threshold, or sweetspot relationship, plus a null model assuming constant rates. All models allowed each trait to have independent mid rate ( $\alpha$ ) or intercept ( $\beta_0$ ) parameters to account for overall differences in rates among traits within a given module, though we assumed all other free parameters were constant across traits for statistical tractability (i.e., a single parameter function “shape” for each module which is rescaled among different traits). We fixed node error variances based on calculated intraspecific variances (excepting bud lengths in some species for which only average trait values were known). Because our data were largely derived from minimum-maximum ranges within species, we lacked any meaningful signal of trait correlations within species and thus set all intraspecific correlations among traits to 0 in our models, though we did estimate evolutionary correlations among traits. We used the correlation matrix transform outlined in Lewandowski et al., 2009 and Stan Development Team, 2019 to avoid issues with exploring correlation parameters that form invalid correlation matrices. To fit these models, we used the same optimization procedure and parameter boundaries which were used in the simulation study, but ran the optimization procedure 20 rather than 10 times.

## **2.3 Results**

### **2.3.1 Simulation Study**

To illustrate potential uses of contsimmaps, we performed an extensive simulation study to investigate whether contsimmaps of continuous variables can be used to infer relationships between rates of continuous trait evolution and continuously-varying factors, in analogy with popular approaches for inferring relationships between rates and discretely-varying factors (e.g., Revell, 2013). Overall, our contsimmap-based pipeline exhibited appropriate error and modest power rates when applied to simulated data (Tables 2.1 and 2.2; Fig. 2.2). Across all simulation conditions,

error rates remained consistent and conservative at around 2-4%, only exceeding 5% in 2 out of the 24 null simulation conditions. On the other hand, power rates varied dramatically across simulation conditions, ranging from only ~5-25% for simulations exhibiting subtler patterns of rate variation (i.e., weak and wide factor-rate relationships) on phylogenies with 50 tips, to ~75-90% and ~90-100% for those exhibiting strong rate variation on phylogenies with 100 and 200 tips, respectively. Unsurprisingly, higher power rates were generally associated with more accurate, unbiased, and precise model-averaged estimates of factor-rate relationships, while lower power rates were associated with increased bias towards inference of overly conservative and “flat” relationships (Fig. 2.3; Figs. 2A.1–2A.3). Differentiation rates—that is, the ability to distinguish among different kinds of factor-rate relationships—largely depended on the kind of factor-rate relationship used to simulate data. Specifically, our method could easily distinguish sweetspot (i.e., modal) and especially simple (i.e., exponential) factor-rate relationships from other kinds of relationships, yet often mistook threshold (i.e., logistic) relationships for either simple or sweetspot ones.

Table 2.1 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given kind of factor-rate relationship (i.e., simple, threshold, or sweetspot) as the best-fitting one (based on having the lowest Bayesian Information Criterion) across all simulation conditions without random variation around factor-rate relationships (“noise”). All models assuming either constant or dummy factor-dependent rates were considered “null” models.

	constant	hidden factor-dependent			simple		threshold			sweetspot		
		simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
		50 tips										
null	0.98	1.00	0.99	0.98	0.51	0.80	0.60	0.83	0.68	0.74	0.94	0.90
simple	0.00	0.00	0.01	0.00	0.48	0.18	0.34	0.15	0.29	0.06	0.02	0.06
threshold	0.02	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.02	0.02	0.00	0.00
sweetspot	0.00	0.00	0.00	0.02	0.00	0.01	0.04	0.01	0.01	0.18	0.04	0.04
		100 tips										
null	0.98	0.98	1.00	0.96	0.11	0.56	0.19	0.60	0.43	0.10	0.68	0.60
simple	0.02	0.01	0.00	0.00	0.87	0.44	0.49	0.34	0.51	0.04	0.10	0.13
threshold	0.00	0.00	0.00	0.01	0.02	0.00	0.13	0.02	0.01	0.09	0.00	0.04
sweetspot	0.00	0.01	0.00	0.03	0.00	0.00	0.19	0.04	0.05	0.77	0.22	0.23
		200 tips										
null	0.98	0.96	0.98	0.96	0.00	0.13	0.07	0.16	0.10	0.00	0.29	0.21
simple	0.02	0.01	0.01	0.01	0.99	0.87	0.17	0.57	0.65	0.00	0.07	0.08
threshold	0.00	0.00	0.00	0.01	0.00	0.00	0.39	0.09	0.13	0.04	0.01	0.08
sweetspot	0.00	0.03	0.01	0.02	0.01	0.00	0.37	0.18	0.12	0.96	0.63	0.63

Table 2.2 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given kind of factor-rate relationship (i.e., simple, threshold, or sweetspot) as the best-fitting one (based on having the lowest Bayesian Information Criterion) across all simulation conditions with random variation around factor-rate relationships (“noise”). All models assuming either constant or dummy factor-dependent rates were considered “null” models.

	constant	hidden factor-dependent			simple		threshold			sweetspot		
		simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
		50 tips										
null	0.99	0.99	0.97	0.99	0.71	0.87	0.62	0.92	0.83	0.76	0.95	0.82
simple	0.00	0.00	0.01	0.00	0.26	0.11	0.24	0.07	0.12	0.05	0.01	0.08
threshold	0.01	0.01	0.02	0.01	0.02	0.01	0.06	0.00	0.01	0.02	0.01	0.01
sweetspot	0.00	0.00	0.00	0.00	0.01	0.01	0.08	0.01	0.04	0.17	0.03	0.09
		100 tips										
null	0.96	0.96	0.98	0.94	0.24	0.64	0.33	0.69	0.48	0.25	0.73	0.67
simple	0.00	0.00	0.00	0.01	0.71	0.27	0.24	0.15	0.38	0.00	0.05	0.07
threshold	0.02	0.02	0.01	0.02	0.00	0.03	0.21	0.09	0.06	0.07	0.03	0.06
sweetspot	0.02	0.02	0.01	0.03	0.05	0.06	0.22	0.07	0.08	0.68	0.19	0.20
		200 tips										
null	0.93	0.97	0.96	0.98	0.07	0.37	0.06	0.34	0.23	0.02	0.44	0.22
simple	0.00	0.00	0.02	0.00	0.86	0.55	0.17	0.36	0.51	0.00	0.04	0.05
threshold	0.04	0.02	0.00	0.00	0.05	0.04	0.44	0.11	0.14	0.07	0.02	0.13
sweetspot	0.03	0.01	0.02	0.02	0.02	0.04	0.33	0.19	0.12	0.91	0.50	0.60

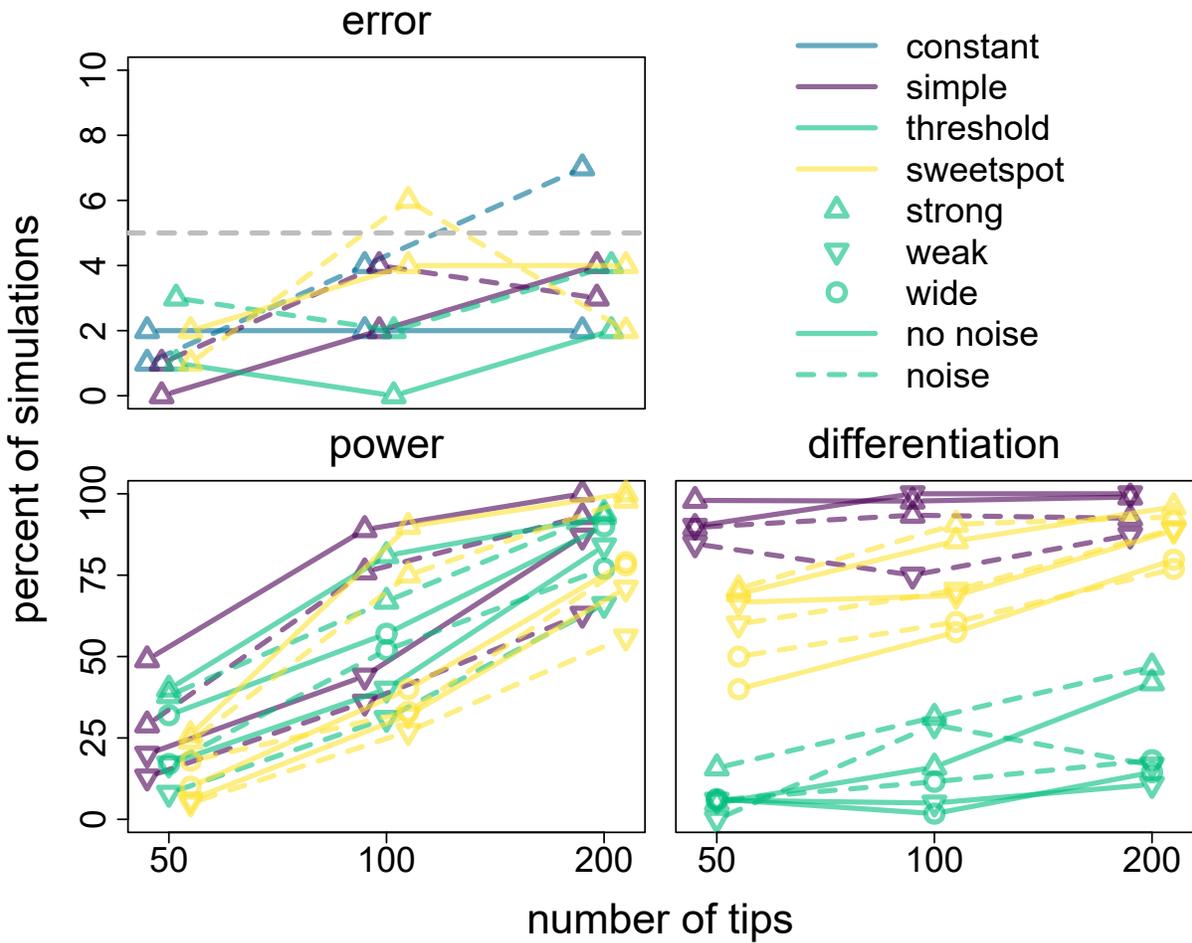


Figure 2.2 Error, power, and differentiation rates of the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution. Different colors correspond to simulations with different factor-rate relationships; different symbols to simulations with differing relationship strength and, in the case of threshold and sweetspot models, width; and dashed versus solid lines to simulations with versus without random variation in rates (“noise”) around simulated factor-rate relationships. Top left: percent of simulations with either constant or hidden factor-dependent rates for which the best-fitting model (i.e., lowest Bayesian Information Criterion) was an observed factor-dependent model (i.e., error rates). Bottom left: percent of simulations with observed factor-dependent rates for which the best-fitting model was also an observed factor-dependent model (i.e., power rates). Bottom right: percent of observed-factor dependent simulations for which the best fitting model assumed the very same kind factor-rate relationship (i.e., simple, threshold, or sweetspot) used to simulate the data (i.e., differentiation rates).

Random variation around simulated factor-rate relationships, or “noise”, had minor yet noticeable effects on our method’s performance. Specifically, noise was associated with slightly increased error rates, albeit inconsistently, as well as ~5-15% decreases to power rates across all

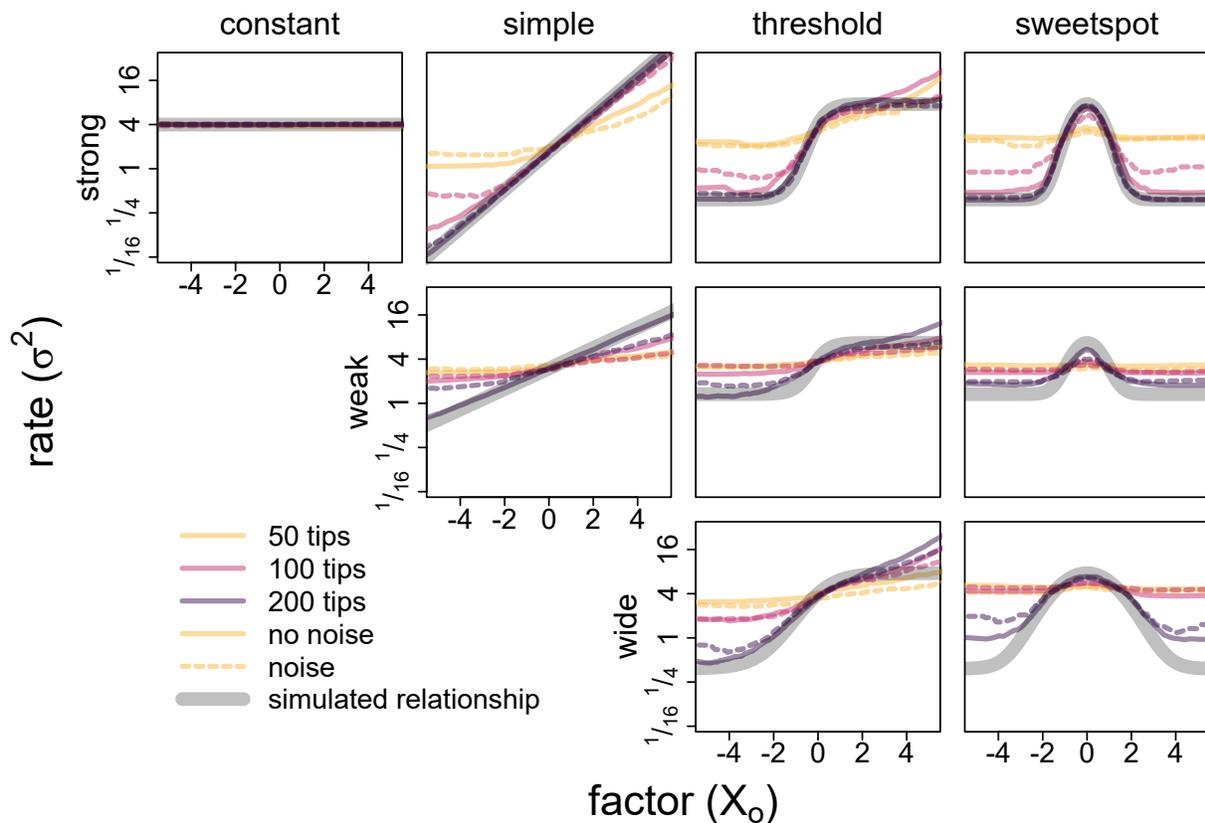


Figure 2.3 Mean model-averaged factor-rate relationships inferred using the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution in comparison to simulated factor-rate relationships under all simulation conditions with either constant or observed factor-dependent rates. Different colors correspond to different phylogeny sizes (i.e., number of tips) and dashed versus solid lines to simulation conditions with versus without random variation in rates (“noise”) around simulated factor-rate relationships. Simulated factor-rate relationships are represented by thick gray lines for reference.

simulation conditions. In terms of differentiation rates, noise appeared to increase support for more complex models assuming threshold or sweetspot factor-rate relationships at the expense of those assuming simple relationships. Overall, however, phylogeny size and/or the factor-relationship used to simulate data had a much stronger effect on both power and differentiation rates compared to noise. Additionally, low error rates across the board indicate that our method is fairly robust to random rate variation due to noise and/or unobserved factors. Interestingly, noise generally had the least severe effects on our method’s performance for simulations on both small and large phylogenies. This pattern is particularly apparent in the estimated factor-rate relationships for simulations

with strong relationships in Fig. 2.3. While noise had virtually no effect on estimated relationships for simulations on phylogenies with 100 and 200 tips, it tended to conservatively bias estimated relationships for simulations with 100 tips. Roughly speaking, we hypothesize that simulations on smaller phylogenies often failed to yield enough signal of factor-dependent rates to support factor-dependent models regardless of noise, while simulations on larger phylogenies yielded enough signal to “overcome” the effect of noise on inference. Indeed, for simulation conditions with weak or wide factor-rate relationships—which generated weaker signals of factor-dependent rates—noise’s effect on estimated relationships was most pronounced for phylogenies with 200 rather than 100 tips, presumably reflecting greater amount of data needed to reliably infer subtler factor-rate relationships.

### 2.3.2 Empirical Example

In addition to the simulation study, we also applied our contsimmap-based pipeline for inferring factor-dependent rates of continuous trait evolution to test whether rates of leaf and flower trait evolution are associated with height variation in the *Eucalyptus* subgenus *Eucalyptus*, or Monocalypts. Ultimately, we found no evidence of height-rate associations for any of the trait modules investigated, instead finding overwhelming support for associations between rates and simulated dummy factors. These results strongly suggest rates of leaf and flower trait evolution in Monocalypts are variable but not closely related to height (Fig. 2.4). To further investigate these results, we calculated marginal rate estimates for each trait module across the Monocalypt phylogeny by first taking the weighted average of estimated rates across all contsimmaps under each fitted model, with weights given by the normalized conditional likelihoods for each contsimmap (i.e., prior-weighted conditional likelihoods for each contsimmap rescaled to have a sum to 1), then model-averaging the resulting rate estimates across all models using BIC weights. Notably, marginal rate variation patterns for each trait module appear similar to those for height evolution as inferred under an evorates model (see APPENDIX 2C for further details on this model)—particularly for juvenile leaves and flowers—suggesting that overall rates of phenotypic evolution among Monocalypts vary according to some common yet currently unknown factor.

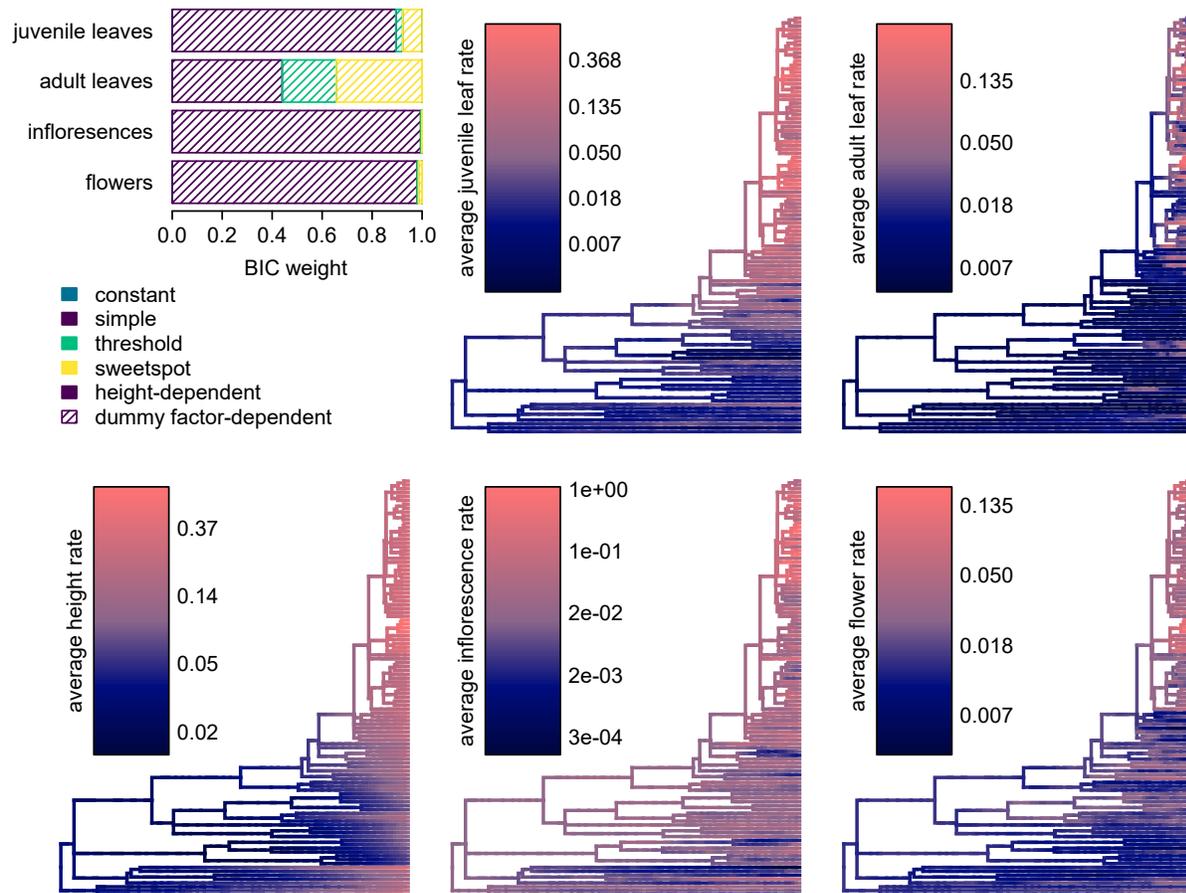


Figure 2.4 Variation in rates of leaf and flower trait evolution in the *Eucalyptus* subgenus *Eucalyptus* (“Monocalypts”). Top left: bar graph displaying Bayesian Information Criterion weights for the contsimmap-based models fit to each trait module. Different colors correspond to different factor-rate relationships, while solid and hatched bars correspond to models assuming rates depend on height and simulated dummy factors, respectively. Bottom left: Phylogram colored according to the average of 349 rate contsimmaps sampled from the posterior of an evolving rates model fit to the Monocalypt height data, with dark blue and light red corresponding to slow and fast rates, respectively. Right: phylograms depicting marginal rates estimated under contsimmap-based models for each trait module, with dark blue and light red once again corresponding to slow and fast rates, respectively.

## 2.4 Discussion

Here, we introduced *contsimmapping*, a flexible and efficient method for sampling histories of continuous variables on phylogenies under Brownian motion models of trait evolution. *Contsimmapping* provides an intuitive framework for reconstructing and analyzing the evolution of continuous phenotypes while accounting for our inherently incomplete knowledge of the evolutionary past. Further, like conventional *simmapping* for discrete variables, *contsimmapping* provides a way to generate distributions of probable evolutionary histories for some variable, enabling straight-forward development of rigorous statistical pipelines for determining how evolutionary processes are affected by continuous variables like body size (e.g., Friedman et al., 2019), generation time (e.g., Gingerich, 2001), or climatic niche (e.g., Tribble et al., 2023). To this end, we implemented a *contsimmap*-based pipeline capable of robustly detecting and accurately quantifying relationships between continuously-varying factors and rates of trait evolution. Using this novel approach, we show that rates of leaf and flower trait evolution are generally unrelated to height in a clade of *Eucalyptus* trees spanning nearly two orders of magnitude in height variation (Brooker et al., 2015; Thornhill et al., 2019; Falster et al., 2021). Nonetheless, our analysis uncovered striking similarities in patterns of evolutionary rate variation across different *Eucalyptus* traits (Figure 2.4), demonstrating how *contsimmaps* provide a powerful toolkit for flexibly interrogating the evolutionary dynamics of continuous traits.

Our implementation of *contsimmapping* is already quite comprehensive, allowing for *contsimmapping* of multiple correlated continuous variables while incorporating within-species variation/measurement error and even regime-dependent rates and trends. Additionally, the *contsimmap* R package provides a variety of tools for transforming, summarizing, and visualizing *contsimmaps*, which may be used to conveniently explore patterns in phylogenetic comparative data on continuous traits while accounting for uncertainty as well as produce publication-quality figures (e.g., Figs. 2.1,2.4; see also Figs. 2C.3 and 2C.5). While we largely focused on using *contsimmaps* to infer associations between rates of continuous trait evolution and continuously-varying factors in the current study, our pipeline could be easily adapted to infer factor-dependent evolutionary

trends and node error variances using tools already available in our R package. Looking forward, contsimmapping could be extended to sample evolutionary histories under other popular models of continuous trait evolution such as Ornstein-Uhlenbeck and Lévy processes, which are generally interpreted as models of adaptive and pulsed evolution, respectively. Another promising future direction would be the development of methods for fitting Ornstein-Uhlenbeck models, discrete trait evolution models, or lineage diversification models with factor-dependent parameters conditional on contsimmaps. Such methods would further enhance macroevolutionary biologists' ability to explore and answer questions regarding the interplay of evolutionary processes with continuous variables.

#### **2.4.1 Using contsimmaps to infer associations between continuous variables and rates of trait evolution**

Overall, our contsimmap-based pipeline provides an effective and practical way to detect and quantify relationships among rates of trait evolution and continuous variables—at least given a sufficiently large effect size and phylogeny. Our pipeline is not the first statistical method developed for inferring relationships between rates of trait evolution and continuous factors, though it is currently the most flexible and thoroughly-tested one to our knowledge. Indeed, many previous methods were developed for particular empirical studies to circumvent the lack of a standard, general-purpose approach. Accordingly, the general statistical performance of these methods often remains largely unknown (e.g., Cooper and Purvis, 2009; Uyeda et al., 2021). The *evorag* method developed by Weir and Lawson, 2015 constitutes a notable exception as a well-studied method suitable for a variety of purposes, though it only uses sister pair contrasts to infer factor-rate relationships. By using entire phylogenies, our method theoretically more fully utilizes the information available in phylogenetic comparative data to infer factor-rate relationships. More recently, Hansen et al., 2022 developed an elegant regression-based approach for inferring correlations between rates and continuous variables, though this method is only capable of inferring linear or exponential relationships between a single continuous factor and rates for a single continuous trait (or node error variances, which are interpreted as “microevolutionary rates” by the authors). In comparison to this

method, our contsimmap-based pipeline sacrifices computational simplicity for flexibility, allowing researchers to incorporate explicit reconstructions of discrete and multiple continuous factors into analyses, fit multivariate models of continuous trait evolution, and/or simultaneously link multiple trait evolution process parameters (i.e., rates, node error variances, and/or evolutionary trends) with reconstructed factors via arbitrary parameter functions (computational/statistical tractability notwithstanding). Overall, our new pipeline is a substantial generalization of previous methods, greatly expanding the hypothesis-testing capabilities of macroevolutionary biologists and allowing researchers to shed light on many famous long-standing questions regarding how phenotypic evolutionary processes are affected by continuous variables.

While our pipeline for inferring associations between rates of trait evolution and continuous factors is generally quite accurate and robust, it notably struggles to detect relatively weak factor-rate relationships from smaller phylogenetic comparative datasets. Based on the simulation conditions examined here, our method generally seems to require phylogenies with around 100 tips or more to reliably detect  $\sim 20$ -fold differences in rates associated with a continuous factor, and at least 200 tips for 5-fold differences. Nonetheless, our approach is still able to reliably detect and reject factor-rate associations even in the face of rate variation due to unobserved factors and/or general “noise” around relationships. Thus, empirical applications of this approach may fail to detect weak factor-rate associations in some cases yet are importantly unlikely to infer spurious associations (though, as with any correlative statistical method, associations may in fact reflect unobserved factors which correlated/confounded with observed factors). Interestingly, while the error rates of our method remained low across all simulation conditions, we do note subtly elevated error among simulations on larger phylogenies involving noise and/or sweetspot (i.e., modal) relationships between rates and unobserved factors. We hypothesize these patterns ultimately reflect “poorly replicated bursts” of trait evolution (*sensu* Maddison and FitzJohn, 2015; Uyeda et al., 2018), whereby apparent yet transient rate fluctuations occur in a few subclades that happen to exhibit similar observed factor values by chance. In practice, one could investigate marginalized rate estimates (see Figure 2.4) to detect such scenarios and/or incorporate node error into analyses to re-

duce the impact of recent, transient rate fluctuations on the inference of trait evolutionary processes (Landis and Schraiber, 2017). Notably, support for constant rate models was rather low across all simulation conditions, including those with truly constant rates unaffected by noise and/or unobserved factors (Tables 2A.4 and 2A.5). We suspect the highly stochastic nature of trait divergence under Brownian motion processes (i.e., subclades exhibiting somewhat anomalous levels of trait disparity by chance), along with the phylogenetic autocorrelation of factor data, may lead to phylogenetic pseudoreplication of spurious factor-rate associations with surprising frequency. Such phenomena would be naturally exacerbated by actual rate variation driven by unobserved factors and/or within-species variation/measurement error creating the illusion of unexpectedly large disparity among recently-diverged lineages (Felsenstein, 2008; Landis and Schraiber, 2017). In any case, dummy factor-dependent models were absolutely critical for controlling our method’s error rates by providing more “competitive” null models that effectively account for such spurious correlations generated either by chance or unobserved factors. Our work adds to a growing body of literature demonstrating that accounting for “background” or “residual” heterogeneity in evolutionary processes is absolutely critical for robust phylogenetic comparative inference and hypothesis testing (Maddison and FitzJohn, 2015; Beaulieu and O’Meara, 2016; Uyeda et al., 2018; May and Moore, 2020; Boyko and Beaulieu, 2023; Boyko et al., 2023b; Tribble et al., 2023).

Beyond the magnitude of rate variation and phylogeny size, differences in the shapes of factor-rate relationships had a profound influence on our method’s power and accuracy. Perhaps unsurprisingly, “wider” factor-rate relationships that featured more gradual changes in rates over factor values were generally harder to both detect and accurately infer compared to narrower relationships with the same effect size. Thus, our method’s ability to infer factor-rate relationships importantly depends not only on the overall magnitude of rate differences, but also how frequently apparent rate shifts tend to occur across a phylogeny, which is primarily determined by how abruptly rates change with respect to factor values. Additionally, sweetspot factor-rate relationships, whereby rates peak or dip at intermediate factor values, appear to require particularly strong effect sizes and/or large phylogenies to reliably detect, likely due to difficulties in correlating the transient rate

fluctuations that occur under such relationships with factor values only observed at the tips of a phylogeny. By comparison, simple (i.e., exponential) and threshold (i.e., logistic) factor-rate relationships, whereby rates strictly decrease or increase with respect factor values, appear to generate more persistent rate shifts and thus yield stronger signals of factor-rate associations. However, similarly to methods for detecting relationships between continuous factors and lineage diversification rates (FitzJohn, 2010), despite consistently detecting factor-dependent rates from data simulated under more complex threshold relationships, our approach largely failed to distinguish such relationships from sweetspot and especially simple relationships—even with large phylogenies including 200 tips.

Notably, some recent work suggests detecting correlations between rates and factors from phylogenetic comparative data is generally quite challenging (Hansen et al., 2022). Thus, the low power documented in our simulation study may well reflect fundamental limits to phylogenetic comparative inference of factor-rate relationships rather than shortcomings particular to our approach. Shrinkage of estimated rates, whereby high and low rates tend to be under- and over-estimated, respectively, has also been found in previous studies investigating methods for inferring variation in rates of trait evolution (Revell, 2013; Martin et al., 2023). Notably, the shrinkage of factor-rate relationships inferred from larger phylogenies was generally mild compared to the shrinkage of relationships inferred from smaller phylogenies—particularly when the simulated factor-rate relationship was weak—suggesting the shrinkage observed in our simulation study primarily arises from our method’s limited power with small effect and sample sizes. Nonetheless, as noted by Revell, 2013, uncertainty in factor histories may cause factor values associated with low rates to be inferred in lineages evolving at relatively high rates and vice versa, weakening apparent factor-rate associations and further exacerbating shrinkage of estimated rates. A joint inference approach, whereby the factor history and its effect on trait evolution processes are inferred simultaneously rather than sequentially, could help mitigate these effects and enable more accurate inference of factor-rate associations. However, such methods entail significant mathematical, statistical, and computational challenges due to their complexity (e.g., FitzJohn, 2010, May and

Moore, 2020, Boyko et al., 2023b).

#### **2.4.2 Size is unrelated to rates of trait evolution in *Monocalypts***

Using our contsimmap-based pipeline for inferring associations between rates of trait evolution and continuous factors, we show that rates and size appear unrelated in the *Eucalyptus* subgenus *Eucalyptus* (i.e., *Monocalypts*), despite numerous theoretical predictions and empirical findings suggesting body size is intertwined with many aspects life history and thus evolutionary rates (Niklas and Enquist, 2001; White et al., 2007; Sibly and Brown, 2007; Adler et al., 2014; Salguero-Gómez et al., 2016; Bakewell et al., 2020). Broadly, lineages of larger organisms are thought to accumulate genetic and phenotypic variation more slowly due to their relatively small population sizes and long generation times. Indeed, larger animals and plants do generally exhibit slower rates of molecular evolution (Gillooly et al., 2005; Fontanillas et al., 2007; Bromham, 2011; Lanfear et al., 2013; Weber et al., 2014; but see Thomas et al., 2006; Wright et al., 2011; Lourenço et al., 2013). However, smaller populations can also evolve rapidly due to genetic drift, and empirical relationships between body size and rates of phenotypic evolution/lineage diversification have notably proven inconsistent compared to relationships between size and molecular rates. In line with our results, a recent analysis suggests size and lineage diversification rates are unrelated within the genus *Eucalyptus* (Vasconcelos et al., 2022), agreeing with patterns found in several animal groups (Cardillo et al., 2003; Feldman et al., 2016; Rainford et al., 2016; but see Amado et al., 2021) but contrasting with a broader pattern of slower lineage diversification among larger plants (Boucher et al., 2017; Igea et al., 2017; see also Wollenberg et al., 2011; Tedesco et al., 2017). Previous work on phenotypic rates predominantly focuses on vertebrates and has also yielded mixed results, suggesting that body shape evolution is slower among larger fish and birds (Friedman et al., 2019; Zimova et al., 2023) while body size and cranial shape evolution is faster among larger mammals (Cooper and Purvis, 2009; Baker et al., 2015).

Ultimately, the lack of general scaling relationships between body size and rates of phenotypic evolution (or lineage diversification rates) likely reflects idiosyncratic adaptations in some clades weakening or completely altering expected relationships among body size, life history traits, and

evolutionary rates. For example, *Eucalyptus* species may exhibit an inverse relationship between size and lifespan, as smaller, shrubby species, termed “mallees”, tend to be highly fire-tolerant and live for many centuries through countless cycles of burning followed by resprouting. On the other hand, the tallest *Eucalyptus* species generally live in more mesic environments and rely on banks of serotinous seed capsules to sprout and replace the previous generation following intense, mature tree-killing fires (Nicolle, 2006). These peculiarities of *Eucalyptus* biology may account for differences in size-lineage diversification rate correlations between *Eucalyptus* and other plant clades (Boucher et al., 2017; Igea et al., 2017; Vasconcelos et al., 2022), though further investigations of size-phenotypic rate correlations are needed to gauge whether the results of our study are truly anomalous or reflect broader patterns across plants.

While rates of phenotypic evolution among Monocalypts showed no relationship to size, they are highly variable based on our analysis’ strong support for models assuming rates are associated with simulated dummy factors. By further interrogating the patterns of rate variation inferred under the dummy factor-dependent models, we show that variation in rates of phenotypic evolution across the Monocalypt phylogeny are relatively consistent across traits. We hypothesize these patterns reflect some common as-of-yet unknown factor modulating the apparent pace of phenotypic evolution among Monocalypts. Given that the highest rates consistently occur in the subclade roughly corresponding to the section *Eucalyptus*—a relatively young radiation concentrated in southeast Australia (Ladiges et al., 2010; Nicolle, 2022)—such rate variation may reflect a genuine shift in evolutionary dynamics among this subclade, an overall acceleration in rates of phenotypic evolution over time, or phenotypic evolution among Monocalypts generally following Ornstein-Uhlenbeck-like processes (Blomberg et al., 2003). Alternatively, such results may also reflect inaccurate estimation of phylogenetic topology and/or branch lengths, as accurate phylogenetic inference is challenging for *Eucalyptus* due to frequent hybridization and incomplete lineage sorting (Rutherford et al., 2016; Thornhill et al., 2019; McLay et al., 2023). In any case, cotsim mapping provides an effective tool for more thoroughly exploring the interplay of body size and rates of phenotypic evolution, potentially helping researchers develop more cohesive, general-

ized theories explaining how the rates of different evolutionary processes interrelate to one another and various life history attributes across the tree of life.

### **2.4.3 Conclusion**

Here, we developed a novel method for sampling the histories of continuous variables on phylogenies under Brownian motion models of trait evolution, generalizing conventional discrete stochastic character mapping or “simmapping” methods to work with continuous variables. We further show how these continuous stochastic character maps or “contsimmaps” may be used to accurately and robustly infer relationships between evolutionary processes and continuously-varying factors such as body size, generation time, or climatic niche. In the process, we notably developed pragmatic techniques to account for the influence of unobserved factors on evolutionary processes in macroevolutionary hypothesis testing. Lastly, we used contsimmaps to test whether height is associated with rates of leaf and flower trait evolution in a clade of eucalyptus trees. Despite finding no evidence for height-rate associations, our empirical case study nonetheless demonstrates the empirical utility of contsimmaps in characterizing general patterns of variation in evolutionary processes across clades. Ultimately, contsimmapping will empower researchers with new and innovative strategies for analyzing the evolutionary dynamics of continuous phenotypes and testing macroevolutionary hypotheses that require knowing how continuous variables have changed over evolutionary time.

## BIBLIOGRAPHY

- Adler P.B., Salguero-Gómez R., Compagnoni A., Hsu J.S., Ray-Mukherjee J., Mbeau-Ache C., and Franco M. 2014. Functional traits explain variation in plant life history strategies. *Proc Natl Acad Sci USA* 111:740–745.
- Amado T.F., Martinez P.A., Pincheira-Donoso D., and Olalla-Tárraga M.Á. 2021. Body size distributions of anurans are explained by diversification rates and the environment. *Glob Ecol Biogeogr* 30:154–164.
- Baker J., Meade A., Pagel M., and Venditti C. 2015. Adaptive evolution toward larger size in mammals. *Proc Natl Acad Sci USA* 112:5093–5098.
- Bakewell A.T., Davis K.E., Freckleton R.P., Isaac N.J.B., and Mayhew P.J. 2020. Comparing life histories across taxonomic groups in multiple dimensions: How mammal-like are insects? *Am Nat* 195:70–81.
- Baliga V.B. and Law C.J. 2016. Cleaners among wrasses: Phylogenetics and evolutionary patterns of cleaning behavior within Labridae. *Mol Phylogenet Evol* 94:424–435.
- Beaulieu J. and O’Meara B. 2023. OUwie: Analysis of Evolutionary Rates in an OU Framework. R package version 2.10.
- Beaulieu J.M. and O’Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst Biol* 65:583–601.
- Beaulieu J.M., O’Meara B.C., and Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. *Syst Biol* 62:725–737.
- Betancourt M. and Girolami M. 2019. Hamiltonian Monte Carlo for hierarchical models. Pages 79–97 *in* *Current Trends in Bayesian Methodology with Applications* (S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan, eds.). Chapman and Hall/CRC Press, Boca Raton, FL.
- Blomberg S.P., Garland T. Jr, and Ives A.R. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57:717–745.
- Bollback J.P. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7:88.
- Borstein S.R., Fordyce J.A., O’Meara B.C., Wainwright P.C., and McGee M.D. 2019. Reef fish functional traits evolve fastest at trophic extremes. *Nat Ecol Evol* 3:191–199.
- Bottou L., Curtis F.E., and Nocedal J. 2018. Optimization methods for large-scale machine learning. *SIAM Rev* 60:223–311.

- Boucher F.C., Démetry V., Conti E., Harmon L.J., and Uyeda J. 2018. A general model for estimating macroevolutionary landscapes. *Syst Biol* 67:304–319.
- Boucher F.C., Verboom G.A., Musker S., and Ellis A.G. 2017. Plant size: A key determinant of diversification? *New Phytol* 216:24–31.
- Boyko J.D. and Beaulieu J.M. 2023. Reducing the biases in false correlations between discrete characters. *Syst Biol* 72:476–488.
- Boyko J.D., Hagen E.R., Beaulieu J.M., and Vasconcelos T. 2023a. The evolutionary responses of life-history strategies to climatic variability in flowering plants. *New Phytol* 240:1587–1600.
- Boyko J.D., O’Meara B.C., and Beaulieu J.M. 2023b. A novel method for jointly modeling the evolution of discrete and continuous traits. *Evolution* 77:836–851.
- Brent R.P. 2013. *Algorithms for Minimization Without Derivatives*. Dover Publications, Mineola, NY.
- Bromham L. 2011. The genome as a life-history character: Why rate of molecular evolution varies between mammal species. *Philos Trans R Soc B* 366:2503–2513.
- Brooker I., Slee A., Connor J., Duffy S., and West J. 2015. *EUCLID Eucalypts of Australia*. 4 ed. Identic Pty Ltd, Brisbane, QLD.
- Burns M.D. and Bloom D.D. 2020. Migratory lineages rapidly evolve larger body sizes than non-migratory relatives in ray-finned fishes. *Proc R Soc B* 287:20192615.
- Cardillo M., Huxtable J.S., and Bromham L. 2003. Geographic range size, life history and rates of diversification in Australian mammals. *J Evol Biol* 16:282–288.
- Chira A.M., Cooney C.R., Bright J.A., Capp E.J.R., Hughes E.C., Moody C.J.A., Nouri L.O., Varley Z.K., and Thomas G.H. 2018. Correlates of rate heterogeneity in avian ecomorphological traits. *Ecol Lett* 21:1505–1514.
- Clavel J., Escarguel G., and Merceron G. 2015. mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data. *Methods Ecol Evol* 6:1311–1319.
- Clavel J. and Morlon H. 2017. Accelerated body size evolution during cold climatic periods in the Cenozoic. *Proc Natl Acad Sci USA* 114:4183–4188.
- Cooper N. and Purvis A. 2009. What factors shape rates of phenotypic evolution? A comparative study of cranial morphology of four mammalian clades. *J Evol Biol* 22:1024–1035.
- de Alencar L.R.V., Martins M., Burin G., and Quental T.B. 2017. Arboreality constrains morphological evolution but not species diversification in vipers. *Proc R Soc B* 284:20171775.

- Dembo R.S. and Steihaug T. 1983. Truncated-Newton algorithms for large-scale unconstrained optimization. *Math Program* 26:190–212.
- Dobzhansky T. and Sturtevant A.H. 1938. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* 23:28–64.
- Drury J.P., Clavel J., Tobias J.A., Rolland J., Sheard C., and Morlon H. 2021. Tempo and mode of morphological evolution are decoupled from latitude in birds. *PLoS Biol* 19:e3001270.
- Dufresne D. 2004. The log-normal approximation in financial and other computations. *Adv Appl Probab* 36:747–773.
- Dziak J.J., Coffman D.L., Lanza S.T., Li R., and Jermini L.S. 2020. Sensitivity and specificity of information criteria. *Brief Bioinform* 21:553–565.
- Fabre A.C., Bardua C., Bon M., Clavel J., Felice R.N., Streicher J.W., Bonnel J., Stanley E.L., Blackburn D.C., and Goswami A. 2020. Metamorphosis shapes cranial diversity and rate of evolution in salamanders. *Nat Ecol Evol* 4:1129–1140.
- Falster D., Gallagher R., Wenk E.H., Wright I.J., Indiarto D., Andrew S.C., Baxter C., Lawson J., Allen S., Fuchs A., Monro A., Kar F., Adams M.A., Ahrens C.W., Alfonzetti M., Angevin T., Apgaua D.M.G., Arndt S., Atkin O.K., Atkinson J., Auld T., Baker A., von Balthazar M., Bean A., Blackman C.J., Bloomfield K., Bowman D.M.J.S., Bragg J., Brodrigg T.J., Buckton G., Burrows G., Caldwell E., Camac J., Carpenter R., Catford J.A., Cawthray G.R., Cernusak L.A., Chandler G., Chapman A.R., Cheal D., Cheesman A.W., Chen S.C., Choat B., Clinton B., Clode P.L., Coleman H., Cornwell W.K., Cosgrove M., Crisp M., Cross E., Crous K.Y., Cunningham S., Curran T., Curtis E., Daws M.I., DeGabriel J.L., Denton M.D., Dong N., Du P., Duan H., Duncan D.H., Duncan R.P., Duretto M., Dwyer J.M., Edwards C., Esperon-Rodriguez M., Evans J.R., Everingham S.E., Farrell C., Firn J., Fonseca C.R., French B.J., Frood D., Funk J.L., Geange S.R., Ghannoum O., Gleason S.M., Gosper C.R., Gray E., Groom P.K., Grootemaat S., Gross C., Guerin G., Guja L., Hahs A.K., Harrison M.T., Hayes P.E., Henery M., Hochuli D., Howell J., Huang G., Hughes L., Huisman J., Ilic J., Jagdish A., Jin D., Jordan G., Jurado E., Kanowski J., Kasel S., Kellermann J., Kenny B., Kohout M., Kooyman R.M., Kotowska M.M., Lai H.R., Laliberté E., Lambers H., Lamont B.B., Lanfear R., van Langevelde F., Laughlin D.C., Laugier-Kitchener B.A., Laurance S., Lehmann C.E.R., Leigh A., Leishman M.R., Lenz T., Lepschi B., Lewis J.D., Lim F., Liu U., Lord J., Lusk C.H., Macinnis-Ng C., McPherson H., Magallón S., Manea A., López-Martinez A., Mayfield M., McCarthy J.K., Meers T., van der Merwe M., Metcalfe D.J., Milberg P., Mokany K., Moles A.T., Moore B.D., Moore N., Morgan J.W., Morris W., Muir A., Munroe S., Nicholson Á., Nicolle D., Nicotra A.B., Niinemets Ü., North T., O'Reilly-Nugent A., O'Sullivan O.S., Oberle B., Onoda Y., Ooi M.K.J., Osborne C.P., Paczkowska G., Pekin B., Guilherme Pereira C., Pickering C., Pickup M., Pollock L.J., Poot P., Powell J.R., Power S.A., Prentice I.C., Prior L., Prober S.M., Read J., Reynolds V., Richards A.E., Richardson B., Roderick M.L., Rosell J.A., Rossetto M., Rye B., Rymer P.D., Sams M.A., Sanson G., Sauquet H., Schmidt S., Schönenberger J., Schulze E.D., Sendall K., Sinclair S., Smith B., Smith R., Soper F., Sparrow B., Standish R.J., Staples T.L., Stephens R.,

- Szota C., Taseski G., Tasker E., Thomas F., Tissue D.T., Tjoelker M.G., Tng D.Y.P., de Tombeur F., Tomlinson K., Turner N.C., Veneklaas E.J., Venn S., Vesik P., Vlasveld C., Vorontsova M.S., Warren C.A., Warwick N., Weerasinghe L.K., Wells J., Westoby M., White M., Williams N.S.G., Wills J., Wilson P.G., Yates C., Zanne A.E., Zemunik G., and Ziemińska K. 2021. AusTraits, a curated plant trait database for the Australian flora. *Sci Data* 8:254.
- Feldman A., Sabath N., Pyron R.A., Mayrose I., and Meiri S. 2016. Body sizes and diversification rates of lizards, snakes, amphisbaenians and the tuatara. *Glob Ecol Biogeogr* 25:187–197.
- Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *Am Nat* 171:713–725.
- Felsenstein J. 2012. A comparative method for both discrete and continuous characters using the threshold model. *Am Nat* 179:145–156.
- FitzJohn R.G. 2010. Quantitative traits and diversification. *Syst Biol* 59:619–633.
- FitzJohn R.G., Maddison W.P., and Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst Biol* 58:595–611.
- Fontanillas E., Welch J.J., Thomas J.A., and Bromham L. 2007. The influence of body size and net diversification rate on molecular evolution during the radiation of animal phyla. *BMC Evol Biol* 7:95.
- Freyman W.A. and Höhna S. 2019. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible Onagraceae lineages. *Syst Biol* 68:505–519.
- Friedman S.T., Martinez C.M., Price S.A., and Wainwright P.C. 2019. The influence of size on body shape diversification across Indo-Pacific shore fishes. *Evolution* 73:1873–1884.
- Friedman S.T. and Muñoz M.M. 2023. A latitudinal gradient of deep-sea invasions for marine fishes. *Nat Commun* 14:773.
- Geyer C. 2011. Introduction to Markov chain Monte Carlo. *in* Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G. L. Jones, and M. Xiao-Li, eds.). Chapman and Hall/CRC, Boca Raton, FL.
- Gillooly J.F., Allen A.P., West G.B., and Brown J.H. 2005. The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci USA* 102:140–145.
- Gingerich P.D. 2001. Rates of evolution on the time scale of the evolutionary process. *Genetica* 112-113:127–144.
- Goolsby E.W. 2017. Rapid maximum likelihood ancestral state reconstruction of continuous char-

- acters: A rerooting-free algorithm. *Ecol Evol* 7:2791–2797.
- Groussin M., Daubin V., Gouy M., and Tannier E. 2016. Ancestral reconstruction: Theory and practice. Pages 70–77 in *Encyclopedia of Evolutionary Biology* (R. M. Kliman, ed.). Academic Press, Waltham, MA.
- Hansen T.F., Bolstad G.H., and Tsuboi M. 2022. Analyzing disparity and rates of morphological evolution with model-based phylogenetic comparative methods. *Syst Biol* 71:1054–1072.
- Hansen T.F., Pienaar J., and Orzack S.H. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62:1965–1977.
- Hassler G., Tolkoﬀ M.R., Allen W.L., Ho L.S.T., Lemey P., and Suchard M.A. 2022. Inferring phenotypic trait evolution on large trees with many incomplete measurements. *J Am Stat Assoc* 117:678–692.
- Huelsenbeck J.P., Nielsen R., and Bollback J.P. 2003. Stochastic mapping of morphological characters. *Syst Biol* 52:131–158.
- Hughes J.J., Berv J.S., Chester S.G.B., Sargis E.J., and Field D.J. 2021. Ecological selectivity and the evolution of mammalian substrate preference across the K-Pg boundary. *Ecol Evol* 11:14540–14554.
- Igea J., Miller E.F., Papadopoulos A.S.T., and Tanentzap A.J. 2017. Seed size and its rate of evolution correlate with species diversification across angiosperms. *PLoS Biol* 15:e2002792.
- Johnson S.G. 2021. The NLOpt nonlinear-optimization package. Version 2.7.1.
- Joy J.B., Liang R.H., McCloskey R.M., Nguyen T., and Poon A.F.Y. 2016. Ancestral reconstruction. *PLoS Comput Biol* 12:e1004763.
- Ladiges P.Y., Bayly M.J., and Nelson G.J. 2010. East-west continental vicariance in: *Eucalyptus*: Subgenus: *Eucalyptus*. Pages 267–302 in *Beyond Cladistics: The Branching of a Paradigm* (D. M. Williams and S. Knapp, eds.). University of California Press, Oakland, CA.
- Landis M.J., Eaton D.A.R., Clement W.L., Park B., Spriggs E.L., Sweeney P.W., Edwards E.J., and Donoghue M.J. 2021. Joint phylogenetic estimation of geographic movements and biome shifts during the global diversification of *Viburnum*. *Syst Biol* 70:67–85.
- Landis M.J. and Schraiber J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes. *Proc Natl Acad Sci USA* 114:13224–13229.
- Landis M.J., Schraiber J.G., and Liang M. 2013. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Syst Biol* 62:193–204.

- Lanfear R., Ho S.Y.W., Jonathan Davies T., Moles A.T., Aarssen L., Swenson N.G., Warman L., Zanne A.E., and Allen A.P. 2013. Taller plants have lower rates of molecular evolution. *Nat Commun* 4:1879.
- Lepage T., Bryant D., Philippe H., and Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol* 24:2669–2680.
- Lewandowski D., Kurowicka D., and Joe H. 2009. Generating random correlation matrices based on vines and extended onion method. *J Multivar Anal* 100:1989–2001.
- Lourenço J.M., Glémin S., Chiari Y., and Galtier N. 2013. The determinants of the molecular substitution process in turtles. *J Evol Biol* 26:38–50.
- Maddison W.P. and FitzJohn R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst Biol* 64:127–136.
- Martin B.S., Bradburd G.S., Harmon L.J., and Weber M.G. 2023. Modeling the evolution of rates of continuous trait evolution. *Syst Biol* 72:590–605.
- May M.R. and Moore B.R. 2020. A Bayesian approach for inferring the impact of a discrete character on rates of continuous-character evolution in the presence of background-rate variation. *Syst Biol* 69:530–544.
- McLay T.G.B., Fowler R.M., Fahey P.S., Murphy D.J., Udovicic F., Cantrill D.J., and Bayly M.J. 2023. Phylogenomics reveals extreme gene tree discordance in a lineage of dominant trees: Hybridization, introgression, and incomplete lineage sorting blur deep evolutionary relationships despite clear species groupings in *Eucalyptus* subgenus *Eudesmia*. *Mol Phylogenet Evol* 187:107869.
- Nations J.A., Mount G.G., Morere S.M., Achmadi A.S., Rowe K.C., and Esselstyn J.A. 2021. Locomotory mode transitions alter phenotypic evolution and lineage diversification in an ecologically rich clade of mammals. *Evolution* 75:376–393.
- Nelder J.A. and Mead R. 1965. A simplex method for function minimization. *Comput J* 7:308–313.
- Nicolle D. 2006. A classification and census of regenerative strategies in the eucalypts (*Angophora*, *Corymbia* and *Eucalyptus*–Myrtaceae), with special reference to the obligate seeders. *Aust J Bot* 54:391–407.
- Nicolle D. 2022. Classification of the eucalypts (*Angophora*, *Corymbia* and *Eucalyptus*) version 6. <https://www.dn.com.au/Classification-Of-The-Eucalypts.pdf> accessed: 2023-11-29.
- Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol* 51:729–739.

- Niklas K.J. and Enquist B.J. 2001. Invariant scaling relationships for interspecific plant biomass production rates and body size. *Proc Natl Acad Sci USA* 98:2922–2927.
- Pauling L., Zuckerkandl E., Henriksen T., and Löfstad R. 1963. Chemical paleogenetics: Molecular “restoration studies” of extinct forms of life. *Acta Chem Scand* 17:9–16.
- Pennell M.W., Eastman J.M., Slater G.J., Brown J.W., Uyeda J.C., FitzJohn R.G., Alfaro M.E., and Harmon L.J. 2014. geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30:2216–2218.
- Rainford J.L., Hofreiter M., and Mayhew P.J. 2016. Phylogenetic analyses suggest that diversification and body size evolution are independent in insects. *BMC Evol Biol* 16:8.
- Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223.
- Revell L.J. 2013. A comment on the use of stochastic character maps to estimate evolutionary rate variation in a continuously valued trait. *Syst Biol* 62:339–345.
- Rincon-Sandoval M., Duarte-Ribeiro E., Davis A.M., Santaquiteria A., Hughes L.C., Baldwin C.C., Soto-Torres L., Acero P A., Walker H.J. Jr, Carpenter K.E., Sheaves M., Ortí G., Arcila D., and Betancur-R R. 2020. Evolutionary determinism and convergence associated with water-column transitions in marine fishes. *Proc Natl Acad Sci USA* 117:33396–33403.
- Rowan T.H. 1990. Functional stability analysis of numerical algorithms. Ph.D. thesis Department of Computer Science, University of Texas at Austin, TX.
- Rutherford S., Wilson P.G., Rossetto M., and Bonser S.P. 2016. Phylogenomics of the green ash eucalypts (Myrtaceae): A tale of reticulate evolution and misidentification. *Aust Syst Bot* 28:326–354.
- Salguero-Gómez R., Jones O.R., Jongejans E., Blomberg S.P., Hodgson D.J., Mbeau-Ache C., Zuidema P.A., de Kroon H., and Buckley Y.M. 2016. Fast-slow continuum and reproductive strategies structure plant life-history variation worldwide. *Proc Natl Acad Sci USA* 113:230–235.
- Sanger F., Thompson E.O., and Kitai R. 1955. The amide groups of insulin. *Biochem J* 59:509–518.
- Sauer T. 2013. *Numerical Analysis: Pearson New International Edition*. Pearson Education Limited, Harlow, UK.
- Schluter D., Price T., Mooers A.Ø., and Ludwig D. 1997. Likelihood of ancestor state in adaptive radiation. *Evolution* 51:1699–1711.

- Schultz T.R., Cocroft R.B., and Churchill G.A. 1996. The reconstruction of ancestral character states. *Evolution* 50:504–511.
- Schwarz G. 1978. Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Sibly R.M. and Brown J.H. 2007. Effects of body size and lifestyle on evolution of mammal life histories. *Proc Natl Acad Sci USA* 104:17707–17712.
- Siqueira A.C., Muruga P., and Bellwood D.R. 2023. On the evolution of fish-coral interactions. *Ecol Lett* 26:1348–1358.
- Slater G.J., Goldbogen J.A., and Pyenson N.D. 2017. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proc R Soc B* 284:20170546.
- Stan Development Team . 2019. Stan Modeling Language Users Guide and Reference Manual. Version 2.21.0.
- Sumrall C.D. and Brochu C.A. 2008. Viewing paleobiology through the lens of phylogeny. *Paleontol Soc Papers* 14:165–183.
- Tedesco P.A., Paradis E., Lévêque C., and Hugueny B. 2017. Explaining global-scale diversification patterns in actinopterygian fishes. *J Biogeogr* 44:773–783.
- Thomas J.A., Welch J.J., Woolfit M., and Bromham L. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. *Proc Natl Acad Sci USA* 103:7366–7371.
- Thornhill A.H., Crisp M.D., Külheim C., Lam K.E., Nelson L.A., Yeates D.K., and Miller J.T. 2019. A dated molecular perspective of eucalypt taxonomy, evolution and diversification. *Aust Syst Bot* 32:29–48.
- Tornabene L., Van Tassell J.L., Robertson D.R., and Baldwin C.C. 2016. Repeated invasions into the twilight zone: Evolutionary origins of a novel assemblage of fishes from deep Caribbean reefs. *Mol Ecol* 25:3662–3682.
- Tribble C.M., May M.R., Jackson-Gain A., Zenil-Ferguson R., Specht C.D., and Rothfels C.J. 2023. Unearthing modes of climatic adaptation in underground storage organs across Liliales. *Syst Biol* 72:198–212.
- Uyeda J.C., Bone N., McHugh S., Rolland J., and Pennell M.W. 2021. How should functional relationships be evaluated using phylogenetic comparative methods? A case study using metabolic rate and body temperature. *Evolution* 75:1097–1105.
- Uyeda J.C., Zenil-Ferguson R., and Pennell M.W. 2018. Rethinking phylogenetic comparative methods. *Syst Biol* 67:1091–1109.

- Vasconcelos T., O'Meara B.C., and Beaulieu J.M. 2022. A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. *Evolution* 76:1420–1433.
- Weber C.C., Nabholz B., Romiguier J., and Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol.* 15:542.
- Weir J.T. and Lawson A. 2015. Evolutionary rates across gradients. *Methods Ecol Evol* 6:1278–1286.
- Welch J.J. and Waxman D. 2008. Calculating independent contrasts for the comparative study of substitution rates. *J Theor Biol* 251:667–678.
- White E.P., Ernest S.K.M., Kerkhoff A.J., and Enquist B.J. 2007. Relationships between body size and abundance in ecology. *Trends Ecol Evol* 22:323–330.
- Witmer L. 1995. The extant phylogenetic bracket and the importance of reconstructing soft tissues in fossils. Pages 19–33 *in* *Functional Morphology in Vertebrate Paleontology* (J. J. Thomason, ed.). Cambridge University Press, New York, NY.
- Wollenberg K.C., Vieites D.R., Glaw F., and Vences M. 2011. Speciation in little: The role of range and body size in the diversification of Malagasy mantellid frogs. *BMC Evol Biol* 11:217.
- Wright S.D., Ross H.A., Jeanette Keeling D., McBride P., and Gillman L.N. 2011. Thermal energy and the rate of genetic evolution in marine fishes. *Evol Ecol* 25:525–530.
- Ypma J., Johnson S.G., Stamm A., Borchers H.W., Eddelbuettel D., Ripley B., Hornik K., Chiquet J., Adler A., Dai X., and Ooms J. 2022. nlotpr: R Interface to NLOPT. R package version 2.0.3.
- Zimova M., Weeks B.C., Willard D.E., Giery S.T., Jirinec V., Burner R.C., and Winger B.M. 2023. Body size predicts the rate of contemporary morphological change in birds. *Proc Natl Acad Sci USA* 120:e2206971120.

APPENDIX 2A

SUPPLEMENTAL TABLES AND FIGURES

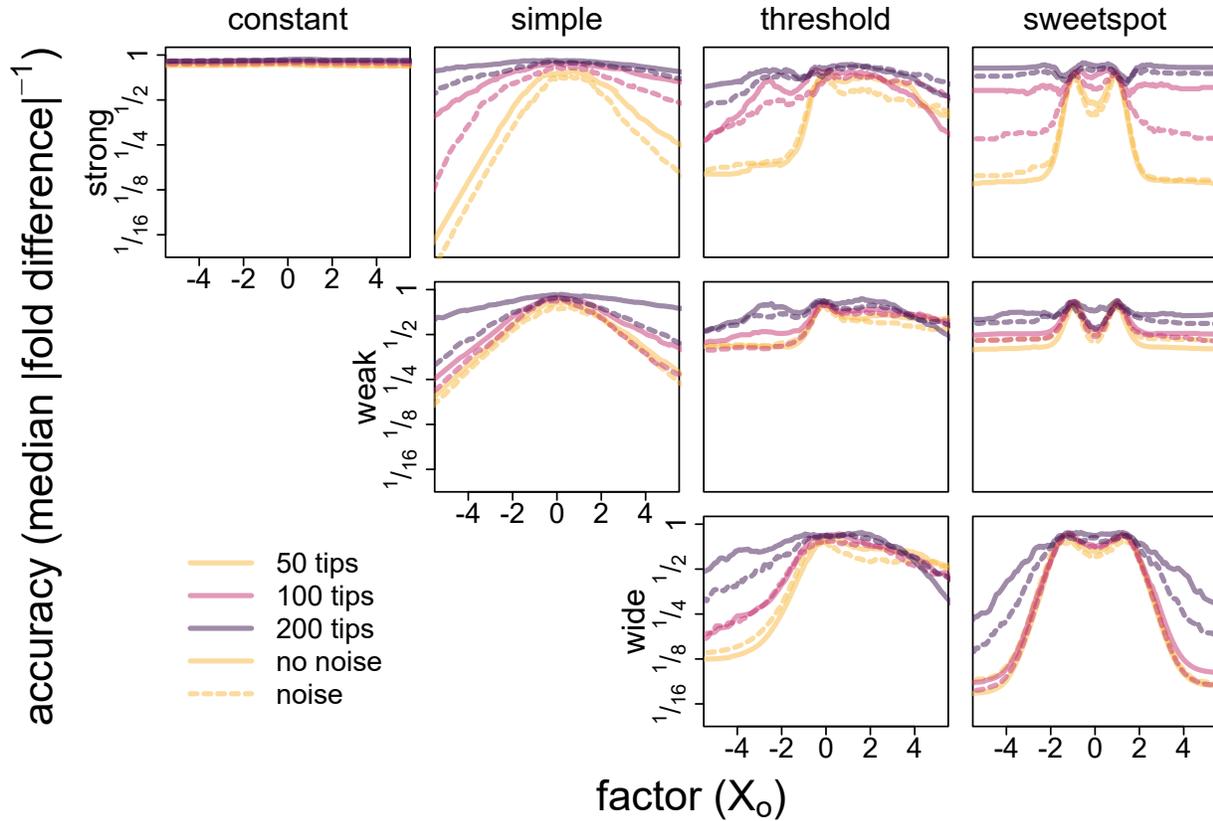


Figure 2A.1 Accuracy of model-averaged factor-rate relationships inferred using the contsimmap-based method for detecting continuous factor-dependent rates of trait evolution for all simulations with either constant or observed factor-dependent rates (i.e., median absolute differences between estimated and simulated relationships on log scale, negated such that higher values correspond to greater accuracy). Different colors correspond to different sample sizes (i.e., number of tips in simulated phylogeny) and solid versus dashed lines to simulations without versus with random variation in rates (“noise”) around inferred relationships.

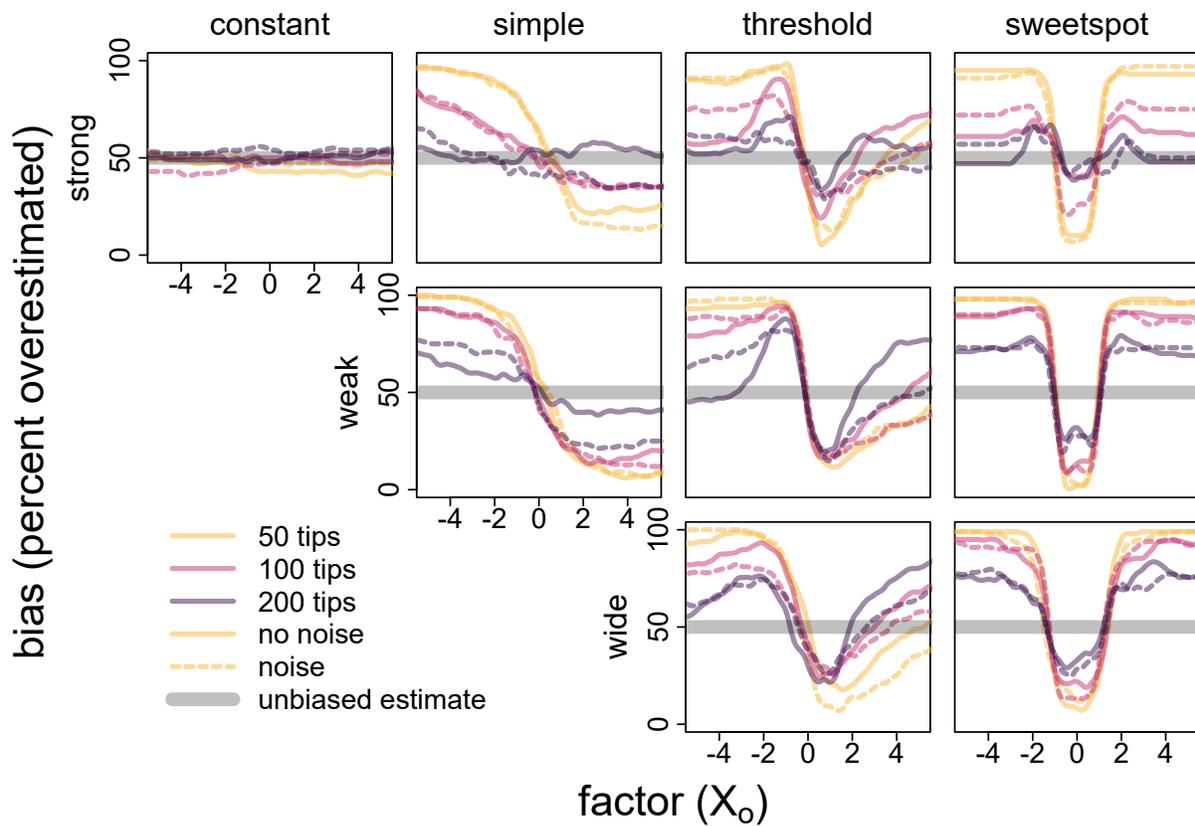


Figure 2A.2 Bias of model-averaged factor-rate relationships inferred using the contsimmap-based method for detecting continuous factor-dependent rates of trait evolution for all simulations with either constant or observed factor-dependent rates (i.e., percent of overestimated model-averaged rates). Different colors correspond to different sample sizes (i.e., number of tips in simulated phylogeny) and solid versus dashed lines to simulations without versus with random variation in rates (“noise”) around inferred relationships. Position of unbiased estimation depicted with thick gray line.

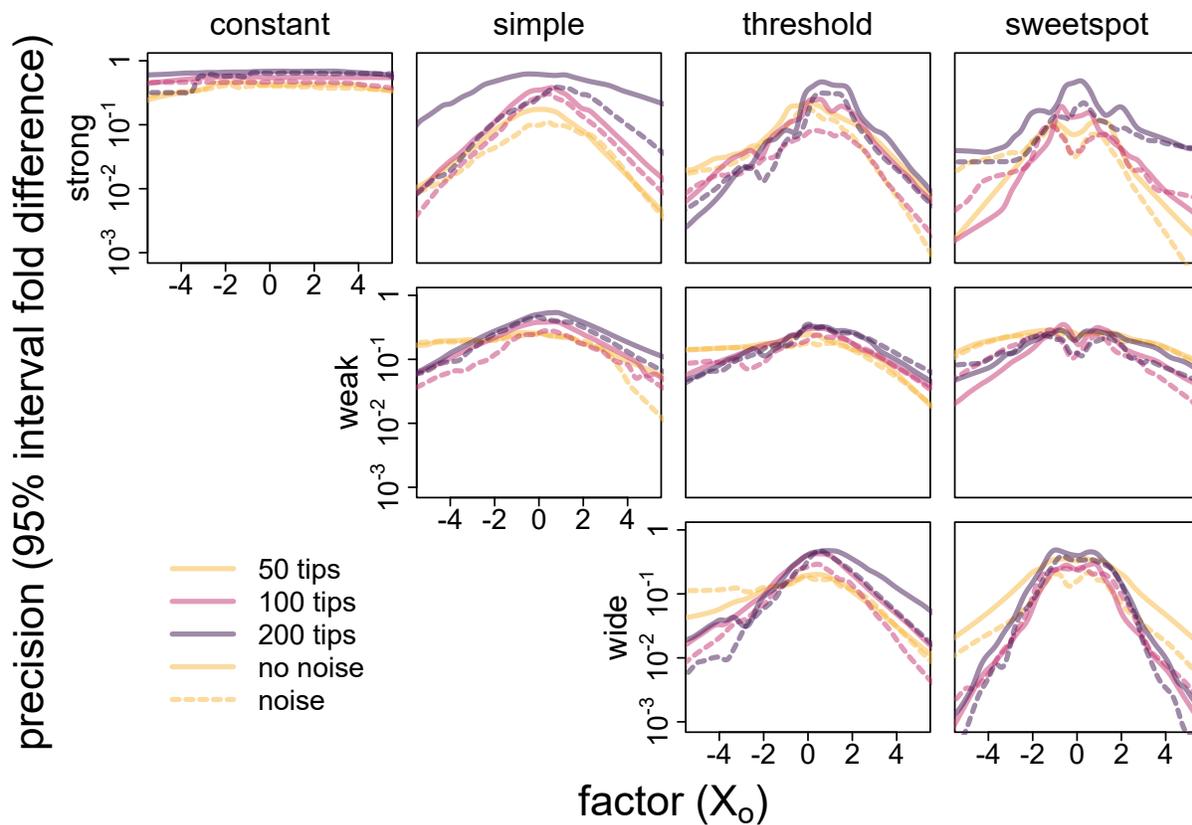


Figure 2A.3 Precision of model-averaged factor-rate relationships inferred using the contsimmap-based method for detecting continuous factor-dependent rates of trait evolution for all simulations with either constant or observed factor-dependent rates (i.e., the lower bound of 95% interval of model-averaged rates divided by the corresponding upper bound). Different colors correspond to different sample sizes (i.e., number of tips in simulated phylogeny) and solid versus dashed lines to simulations without versus with random variation in rates (“noise”) around inferred relationships.

Table 2A.1 Table of parameter values for each version of simple, threshold, sweetspot, and null parameter functions used to generate data for the simulation study of the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution. In general, we consider the “strong” versions of each function the default, which are standardized to yield an overall rate around 4 that varies  $\sim 20$ -fold as factor values range from -2 to 2. “Weak” functions are largely identical to strong functions but instead only cause rates to vary  $\sim 5$ -fold over the -2 to 2 interval. On the other hand, “wide” functions still cause rates to vary  $\sim 20$ -fold but over a wider range of factor values from -4 to 4 (note that there is no wide version of the simple function). Intercept and slope parameters were chosen to ensure simple functions reach the minimum/maximum values of corresponding threshold/sweetspot functions at factor values of -2.5 and 2.5.

function version	parameter					
	intercept ( $\beta_0$ )	slope ( $\beta_1$ )	mid-rate ( $\alpha$ )	location ( $\theta$ )	rate deviation ( $\delta$ )	width ( $\omega$ )
strong	$\ln 4 - \ln \cosh\left(\frac{\ln 20}{2}\right)$	$\frac{\ln 20}{5}$	$\ln 4$	0	$\frac{\ln 20}{2}$	$\ln 4$
weak	$\ln 4 - \ln \cosh\left(\frac{\ln 5}{2}\right)$	$\frac{\ln 5}{5}$	$\ln 4$	0	$\frac{\ln 5}{2}$	$\ln 4$
wide	—	—	$\ln 4$	0	$\frac{\ln 20}{2}$	$\ln 8$

Table 2A.2 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given factor-rate relationship as the best-fitting one (based on having the lowest sample size corrected Akaike Information Criterion) across all simulation conditions without random variation around factor-rate relationships (“noise”). Models assuming either constant or dummy factor-dependent rates were considered “null” models.

	constant	hidden factor-dependent			simple		threshold			sweetspot		
		simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
		50 tips										
null	0.96	0.98	0.99	0.97	0.53	0.83	0.59	0.87	0.67	0.65	0.91	0.82
simple	0.02	0.00	0.01	0.00	0.38	0.11	0.21	0.07	0.19	0.01	0.01	0.06
threshold	0.02	0.01	0.00	0.01	0.04	0.04	0.06	0.02	0.04	0.07	0.02	0.03
sweetspot	0.00	0.01	0.00	0.02	0.05	0.02	0.14	0.04	0.10	0.27	0.06	0.09
		100 tips										
null	0.96	0.98	0.99	0.95	0.12	0.62	0.17	0.58	0.45	0.07	0.64	0.42
simple	0.02	0.00	0.00	0.00	0.74	0.30	0.15	0.18	0.25	0.00	0.03	0.05
threshold	0.01	0.00	0.00	0.01	0.07	0.02	0.34	0.10	0.13	0.13	0.00	0.10
sweetspot	0.01	0.02	0.01	0.04	0.07	0.06	0.34	0.14	0.17	0.80	0.33	0.43
		200 tips										
null	0.93	0.95	0.97	0.96	0.00	0.18	0.04	0.19	0.09	0.00	0.19	0.13
simple	0.03	0.00	0.00	0.01	0.92	0.57	0.03	0.11	0.25	0.00	0.01	0.01
threshold	0.00	0.00	0.00	0.01	0.01	0.10	0.44	0.28	0.24	0.04	0.04	0.11
sweetspot	0.04	0.05	0.03	0.02	0.07	0.15	0.49	0.42	0.42	0.96	0.76	0.75

Table 2A.3 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given factor-rate relationship as the best-fitting one (based on having the lowest sample size corrected Akaike Information Criterion) across all simulation conditions with random variation around factor-rate relationships (“noise”) are given without and within parentheses, respectively. Models assuming either constant or dummy factor-dependent rates were considered “null” models.

	constant	hidden factor-dependent			simple		threshold			sweetspot		
		simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
		50 tips										
null	0.98	0.98	0.96	0.94	0.74	0.86	0.64	0.90	0.82	0.70	0.90	0.81
simple	0.00	0.00	0.01	0.00	0.19	0.07	0.17	0.04	0.09	0.01	0.00	0.04
threshold	0.01	0.01	0.03	0.02	0.02	0.02	0.08	0.01	0.05	0.03	0.01	0.02
sweetspot	0.01	0.01	0.00	0.04	0.05	0.05	0.11	0.05	0.04	0.26	0.09	0.13
		100 tips										
null	0.93	0.95	0.97	0.92	0.32	0.68	0.29	0.66	0.47	0.20	0.64	0.65
simple	0.00	0.00	0.00	0.01	0.39	0.13	0.10	0.04	0.18	0.00	0.03	0.03
threshold	0.04	0.02	0.01	0.04	0.04	0.05	0.32	0.12	0.13	0.07	0.06	0.08
sweetspot	0.03	0.03	0.02	0.03	0.25	0.14	0.29	0.18	0.22	0.73	0.27	0.24
		200 tips										
null	0.87	0.95	0.92	0.97	0.07	0.52	0.04	0.31	0.23	0.02	0.37	0.17
simple	0.00	0.00	0.00	0.00	0.64	0.28	0.05	0.08	0.16	0.00	0.01	0.02
threshold	0.04	0.02	0.01	0.00	0.12	0.07	0.52	0.29	0.25	0.07	0.06	0.16
sweetspot	0.09	0.03	0.07	0.03	0.17	0.13	0.39	0.32	0.36	0.91	0.56	0.65

Table 2A.4 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given factor-rate relationship as the best-fitting one (based on having the lowest Bayesian Information Criterion; d. denotes models assuming relationships with simulated “dummy” factors, while o. denotes models assuming relationships with observed factors) across all simulation conditions without random variation around factor-rate relationships (“noise”).

		constant	hidden factor-dependent			simple		threshold			sweetspot		
			simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
50 tips	constant	0.38	0.05	0.02	0.02	0.01	0.07	0.03	0.10	0.09	0.03	0.13	0.10
	d. simple	0.49	0.78	0.80	0.69	0.42	0.67	0.46	0.64	0.44	0.54	0.75	0.67
	d. threshold	0.08	0.12	0.08	0.12	0.04	0.04	0.03	0.06	0.09	0.10	0.04	0.08
	d. sweetspot	0.03	0.05	0.09	0.15	0.04	0.02	0.08	0.03	0.06	0.07	0.02	0.05
	o. simple	0.00	0.00	0.01	0.00	0.48	0.18	0.34	0.15	0.29	0.06	0.02	0.06
	o. threshold	0.02	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.02	0.02	0.00	0.00
	o. sweetspot	0.00	0.00	0.00	0.02	0.00	0.01	0.04	0.01	0.01	0.18	0.04	0.04
100 tips	constant	0.35	0.01	0.01	0.00	0.00	0.05	0.01	0.04	0.03	0.01	0.00	0.04
	d. simple	0.49	0.71	0.67	0.70	0.10	0.45	0.14	0.51	0.37	0.08	0.55	0.44
	d. threshold	0.07	0.12	0.13	0.06	0.01	0.05	0.02	0.02	0.01	0.01	0.05	0.05
	d. sweetspot	0.07	0.14	0.19	0.20	0.00	0.01	0.02	0.03	0.02	0.00	0.08	0.07
	o. simple	0.02	0.01	0.00	0.00	0.87	0.44	0.49	0.34	0.51	0.04	0.10	0.13
	o. threshold	0.00	0.00	0.00	0.01	0.02	0.00	0.13	0.02	0.01	0.09	0.00	0.04
	o. sweetspot	0.00	0.01	0.00	0.03	0.00	0.00	0.19	0.04	0.05	0.77	0.22	0.23
200 tips	constant	0.40	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00
	d. simple	0.54	0.65	0.69	0.51	0.00	0.12	0.05	0.13	0.06	0.00	0.26	0.18
	d. threshold	0.01	0.21	0.12	0.21	0.00	0.01	0.00	0.00	0.02	0.00	0.01	0.01
	d. sweetspot	0.03	0.10	0.17	0.23	0.00	0.00	0.01	0.01	0.01	0.00	0.02	0.02
	o. simple	0.02	0.01	0.01	0.01	0.99	0.87	0.17	0.57	0.65	0.00	0.07	0.08
	o. threshold	0.00	0.00	0.00	0.01	0.00	0.00	0.39	0.09	0.13	0.04	0.01	0.08
	o. sweetspot	0.00	0.03	0.01	0.02	0.01	0.00	0.37	0.18	0.12	0.96	0.63	0.63

Table 2A.5 Proportions of times the contsimmap-based pipeline for detecting continuous factor-dependent rates of trait evolution selected a model assuming a given factor-rate relationship as the best-fitting one (based on having the lowest Bayesian Information Criterion; d. denotes models assuming relationships with simulated “dummy” factors, while o. denotes models assuming relationships with observed factors) across all simulation conditions with random variation around factor-rate relationships (“noise”).

		constant	hidden factor-dependent			simple		threshold			sweetspot		
			simple	threshold	sweetspot	strong	weak	strong	weak	wide	strong	weak	wide
50 tips	constant	0.14	0.01	0.01	0.03	0.00	0.02	0.01	0.00	0.01	0.03	0.01	0.03
	d. simple	0.57	0.67	0.59	0.59	0.42	0.53	0.33	0.59	0.54	0.46	0.56	0.49
	d. threshold	0.24	0.19	0.14	0.15	0.12	0.20	0.14	0.18	0.17	0.14	0.24	0.16
	d. sweetspot	0.04	0.12	0.23	0.22	0.17	0.12	0.14	0.15	0.11	0.13	0.14	0.14
	o. simple	0.00	0.00	0.01	0.00	0.26	0.11	0.24	0.07	0.12	0.05	0.01	0.08
	o. threshold	0.01	0.01	0.02	0.01	0.02	0.01	0.06	0.00	0.01	0.02	0.01	0.01
	o. sweetspot	0.00	0.00	0.00	0.00	0.01	0.01	0.08	0.01	0.04	0.17	0.03	0.09
100 tips	constant	0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.01
	d. simple	0.46	0.64	0.58	0.43	0.14	0.39	0.21	0.37	0.26	0.10	0.42	0.35
	d. threshold	0.26	0.09	0.19	0.23	0.06	0.12	0.04	0.12	0.13	0.09	0.17	0.17
	d. sweetspot	0.15	0.23	0.21	0.28	0.04	0.12	0.08	0.18	0.09	0.06	0.13	0.14
	o. simple	0.00	0.00	0.00	0.01	0.71	0.27	0.24	0.15	0.38	0.00	0.05	0.07
	o. threshold	0.02	0.02	0.01	0.02	0.00	0.03	0.21	0.09	0.06	0.07	0.03	0.06
	o. sweetspot	0.02	0.02	0.01	0.03	0.05	0.06	0.22	0.07	0.08	0.68	0.19	0.20
200 tips	constant	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	d. simple	0.54	0.62	0.45	0.49	0.02	0.24	0.04	0.27	0.13	0.00	0.23	0.12
	d. threshold	0.19	0.23	0.32	0.24	0.02	0.06	0.02	0.05	0.02	0.00	0.07	0.06
	d. sweetspot	0.12	0.12	0.19	0.25	0.03	0.07	0.00	0.02	0.08	0.02	0.13	0.04
	o. simple	0.00	0.00	0.02	0.00	0.86	0.55	0.17	0.36	0.51	0.00	0.04	0.05
	o. threshold	0.04	0.02	0.00	0.00	0.05	0.04	0.44	0.11	0.14	0.07	0.02	0.13
	o. sweetspot	0.03	0.01	0.02	0.02	0.02	0.04	0.33	0.19	0.12	0.91	0.50	0.60

## APPENDIX 2B

### STOCHASTIC APPROXIMATION OF LIKELIHOOD FUNCTION GRADIENTS

As discussed in subsection 2.2.2, likelihood functions for continuous factor-dependent Brownian motion models (at least when conditioned on contsimmapped—as opposed to jointly inferred—factor histories) appear to often feature multiple local optima and relatively flat ridges, which can pose severe challenges for numerical optimization routines. Accordingly, while simplex-based optimization algorithms like Nelder-Mead (Nelder and Mead, 1965) and Subplex (Rowan, 1990) are commonly used to fit models to phylogenetic comparative data (e.g., Beaulieu et al., 2013), initial tests revealed that simplex-based algorithms tend to be very sensitive to initial parameter values in the case of our method. Specifically, given poor initial estimates of parameter values, simplex-based algorithms tended to terminate in suboptimal peaks or ridges of the likelihood surface and/or take impractically long amounts of time to converge. Thus, we sought to use gradient-based optimization algorithms, which leverage information on how likelihoods change around a given point on the likelihood surface (i.e., gradients) to more rapidly and reliably converge on high-likelihood regions of parameter space given an arbitrary starting point. Unfortunately, there is no simple, closed-form solution for computing gradients of factor-dependent Brownian motion likelihood functions (barring the full-blown implementation of an automatic differentiation engine), so we instead focused on approximating gradients using finite differences.

Let  $x$  denote a  $n$ -length vector of parameter values,  $f$  the corresponding likelihood function, and  $g(x)$  the (unknown) gradient of  $f$  about  $x$ . Note that  $g(x)$ , unlike  $f(x)$ , does not represent a single number but instead an  $n$ -length vector of “slopes” of the likelihood surface ( $f$ ) along each of its input dimensions, corresponding to each of the  $n$  different parameters. Thus, the  $i$ th entry of  $g(x)$  can be approximated by calculating a central finite difference about  $x_i$ :

$$g(x)_i \approx \frac{f(x + h_i \mathbb{1}_i) - f(x - h_i \mathbb{1}_i)}{2h_i} \quad (1)$$

Where  $h_i$  is a small step size to take along the dimension corresponding to the  $i$ th parameter and  $\mathbb{1}_i$  an  $n$ -length indicator vector with 1 in its  $i$ th entry and 0s in all other entries. In practice,  $h_i$

should be small enough to accurately approximate the instantaneous slope/partial derivative of  $f$  with respect to the  $i$ th parameter value, but large enough to avoid numerical errors stemming from rounding during floating point arithmetic operations (Sauer, 2013). To this end, in our implementation, we define  $h_i$  as:

$$h_i = \begin{cases} \sqrt[3]{\varepsilon}/2 & \text{if } |x_i| \leq 1 \\ |x_i|\sqrt[3]{\varepsilon}/2 & \text{if } |x_i| > 1 \end{cases} \quad (2)$$

Where  $\varepsilon$  denotes machine epsilon ( $\sim 1e-16$  for R on a typical computer).

Unfortunately, Eq. (1) breaks down when  $f(x + h_i \mathbf{1}_i)$  and/or  $f(x - h_i \mathbf{1}_i)$  are undefined, which occurs frequently under some circumstances due to the wide variety of parameter functions and boundaries that may be specified using our implementation. Notably, undefined values of  $f$  can also result from the fact that our pruning algorithm rounds rates/variances to 0 when they fall below a certain threshold ( $\sqrt{\varepsilon}$  by default), as we found “pseudoinversion” of matrices including 0s (Hasler et al., 2022) generally produced more numerically stable likelihood calculations compared to direct inversion of matrices with arbitrarily small values. In rare cases, such rounding implies distinct trait measurements must be exactly identical, yielding undefined likelihoods due to unexpected contradictions in observed trait data. Fortunately, we only use  $g(x)$  to help numerical optimization algorithms explore parameter spaces efficiently, rather than precisely identify peaks in the likelihood surface (somewhat analogously to stochastic gradient descent algorithms commonly employed in machine learning; Bottou et al., 2018). Thus, we implemented a crude but pragmatic procedure to roughly approximate  $g(x)$  in cases involving undefined evaluations of  $f$ :

If  $L = f(x)$  is undefined, randomly sample entries of  $g(x)$  by drawing either  $-0.1$  or  $0.1$  with equal probability. Otherwise, for each parameter  $i$  from 1 to  $n$ :

- 1) Complete the “forward” likelihood evaluation,  $L^+ = f(x + h_i \mathbf{1}_i)$ , as well as the “backward” evaluation,  $L^- = f(x - h_i \mathbf{1}_i)$ .

- 2) If both  $L^+$  and  $L^-$  are undefined, randomly sample  $g(x)_i$  by drawing either  $-0.1$  or  $0.1$  with equal probability.
- 3) If only  $L^+$  is defined, set  $g(x)_i$  to the forward difference,  $\frac{L^+ - L}{h_i}$ . If only  $L^-$  is defined, set  $g(x)_i$  to the backward difference,  $\frac{L - L^-}{h_i}$ .
- 4) If both  $L^+$  and  $L^-$  are defined, set  $g(x)_i$  to the central difference,  $\frac{L^+ - L^-}{2h_i}$ .

While switching between the central, forward, or backward finite difference approximations based on which likelihood function evaluations are defined is quite intuitive, there is no theoretical basis to justify sampling random slopes when finite differences are undefined. We settled on this particular procedure through trial and error by implementing a few different strategies for handling undefined gradients and investigating the resulting performance of NLOPT’s truncated Newton algorithm (Dembo and Steihaug, 1983) in fitting factor-dependent Brownian motion models to several of the simulated and empirical datasets described in subsections sections 2.2.3 and 2.2.4. Generally, we found that leaving entries of  $g(x)$  undefined frequently caused errors that lead to premature termination of the algorithm, while setting undefined entries of  $g(x)$  to 0 (or sampling  $g(x)$  from a narrow normal distribution centered at 0) resulted in the algorithm getting “stuck” in complex boundary regions of parameter space. Accordingly, our final strategy is meant to consistently sample slopes substantially different from 0 to prevent the algorithm from getting stuck, while not sampling slopes so large as to make the algorithm “fly off” in some random direction. Nonetheless, in the context of our method, the former condition proved much more vital to the performance of gradient-based optimization than the latter. Should these sampled slopes send the algorithm to effectively random regions of parameter space, subsequent iterations and/or follow-up optimization algorithms (e.g., simplex-based, principal axis) will in any case continue to improve on these estimated parameter values. Ultimately, the specific slopes sampled to replace undefined entries of  $g(x)$  should have very little impact on final parameter estimates as long as the slopes substantially differ from 0, and our admittedly non-systematic investigation of model fitting performance largely supported this conclusion.

Unfortunately, this gradient approximation strategy has one more key problem: its computational complexity scales with the number of parameters. Because the likelihood function must be reevaluated  $2n$  times to approximate a single gradient, fitting models with many parameters via gradient-based optimization can become prohibitively slow under this approach, yet the goal of implementing this gradient approximation procedure is precisely to more efficiently and robustly fit complex, parameter-rich models in the first place. To mitigate this problem, we exploit the fact that our model fitting procedure works by taking weighted averages of likelihoods conditional on each contsimmap. In agreement with typical distributions of likelihoods under factor-dependent continuous trait evolution models conditional on discrete simmaps (Boyko et al., 2023b), contsimmap-based models generally yield only a few substantially high conditional likelihoods for any given vector of parameter estimates, with the majority of conditional likelihoods being low to negligible. In other words, relatively few contsimmaps will often contribute to nearly all variation in  $f$  about a certain point, rendering full reevaluation of the likelihood function largely unnecessary. Instead, one may reevaluate the likelihood much more quickly based on a relatively small subsample of contsimmaps, similar in spirit to how stochastic gradient descent algorithms subsample data before computing gradients to speed up calculations (Bottou et al., 2018).

To outline our contsimmap subsampling procedure more explicitly, let  $M$  denote the total number of contsimmaps and  $L$  an  $M$ -length vector of their associated normalized conditional likelihoods (i.e.,  $L$  is normalized to have a sum of 1) sorted in decreasing order. To define an appropriate subsample size for gradient calculations,  $m$ , we find the position,  $j$ , of the first entry of  $L$  to cumulatively sum to greater than a user-specified quality parameter between 0 and 1,  $q$  (defaulting to 0.9), and set  $m = 2j$ . Then,  $m$  contsimmaps are randomly sampled without replacement in proportion to their normalized conditional likelihoods,  $L$ , generating stochastic subsamples that consist of contsimmaps with especially high conditional likelihoods. We chose to set  $m = 2j$  because it resulted in the subsampled normalized conditional likelihoods summing to around  $q$  or greater rather consistently. Notably, there are some edge cases which may force  $m$  to be lower than  $2j$ : in particular, any normalized conditional likelihoods that underflow to 0 are never subsampled, and, to prevent

excessively large subsamples,  $m$  is never allowed to exceed  $qM$  rounded to the nearest integer. Generally speaking, this procedure yielded gradients accurate to at least a couple of decimal digits while only using around a tenth of all the contsimmaps used in a given factor-dependent Brownian motion model—at least when applied to fitting factor-dependent Brownian Motiopn models to several of the simulated and empirical datasets described in subsections sections 2.2.3 and 2.2.4.

Ultimately, while the crudeness of our gradient approximation procedure is likely inappropriate for fully gradient-based optimization, it nonetheless works well enough to find high likelihood regions of parameter space much more efficiently and consistently than simplex-based optimization.

## APPENDIX 2C

### GENERATING CONTSIMMAPS UNDER EVORATES MODELS

Empirical data on height variation across the *Eucalyptus* subgenus *Eucalyptus* (hereafter Monocalypts) exhibits strong evidence of evolutionary rate variation. Overall, Monocalypts have accumulated a little over 4 units of log height variation (corresponding to around a 70-fold difference) over roughly 30 million years of evolution, yet several groups of closely-related species in the section *Eucalyptus* span 1.5 to 3 units of log height variation (i.e., ~5 to 20-fold differences) despite originating less than 2 million years ago. For example, *E. obliqua* L'Hér is around 40 m tall on average, yet shares a 1.5 million-year-old common ancestor with the 3-4 m tall *E. gregsoniana* L.A.S.Johnson & Blaxell and *E. kybeanensis* Maiden & Camabage. Likewise, the most recent common ancestor of the 40 m tall *E. delegatensis* R.T.Baker and the 2 m tall *E. cunninghamii* Sweet dates to a mere 650,000 years ago. While our data do suggest Monocalypts exhibit substantial levels of within-species height variation (usually ranging between some 50 and 200% the species mean), these disparities among recently-diverged lineages appear too extreme and frequent within the section *Eucalyptus* to result from errors in estimating mean species heights alone. Such rate variation violates the assumptions of a constant-rate Brownian motion model, resulting in inflated estimates of rates of height evolution and highly uncertain ancestral state reconstructions. Additionally, across-species differences in observed height data tends to be attributed to within-species variation rather than evolutionary divergence under constant-rate models, causing an overall “shrinkage” of inferred mean heights at the tips. To mitigate these issues, we decided to take advantage of our contsimmapping algorithm’s ability to handle evolutionary rate variation according to mapped regimes, fitting models that allow rates to vary to the Monocalypt height data using the *evorates* package (Martin et al., 2023) and implementing tools to generate contsimmaps under such models.

More specifically, we fit four models to the Monocalypt height data: 1) a “trend” model whereby rates of height evolution exponentially decrease or increase over time (equivalent to an “early/late burst” model), 2) a “rate variance” model whereby different subclades gradually di-

verge in rates of height evolution, 3) a “full” model combining both the trend and rate variance models, and 4) a “null” model with constant rates (equivalent to a constant-rate Brownian motion model). We fit all models using four independent Hamiltonian Monte Carlo chains consisting of 2,000 iterations, discarding the first 1,000 as warmup for a total of 4,000 posterior samples. All chains adequately converged ( $\hat{R} < 1.01$ ) and achieved sufficient effective sample sizes (effective sample sizes  $> 100$  per chain) (Stan Development Team, 2019). Ultimately, we found high support for increasing rates through time under the trend model (posterior probability  $> 0.99$ ) and substantial rate heterogeneity among clades under the rate variance model (Savage-Dickey ratio  $< 0.01$ ). However, the full model yielded equivocal support for both increasing (posterior probability = 0.75) and heterogeneous rates (Savage-Dickey ratio = 0.36), suggesting that apparent variation in rates of Monocalypt height evolution could be due to accelerating rates, differences in rates among subclades, or both. Indeed, there was a strong negative correlation between posterior samples of the trend and rate variance parameters, such that the full model’s posterior consisted of both trend-like and rate variance-like models (Fig. 2C.1). This conclusion is further supported by a cursory look at the posterior samples of likelihoods under each model, as the posterior likelihoods under the full model overlap with those under both the trend and rate variance models (Fig. 2C.2) We thus chose to integrate over this uncertainty in the underlying height evolution model, generating 4,000 contsimmaps (resolution/ $\xi = 100$ ) of Monocalypt height for each posterior sample under the full model. We ultimately found that generating contsimmaps of Monocalypt height evolution under the null model yielded unrealistically wide distributions of sampled height values while simultaneously “rounding out” observed across-species differences in height. On the other hand, contsimmapping under the full model yielded inferred ancestral states and tip means more concordant with the observed height data (Figs. 2C.3 and 2C.4).

While contsimmaps under the full model generally exhibited a narrower distribution of height values, we nonetheless chose to filter out 516 ( $\sim 13\%$ ) of the 4,000 contsimmaps that included biologically unrealistic height values less than 10 cm or greater than 150 meters (by comparison, 822 or  $\sim 21\%$  of contsimmaps under the null model included height values outside this range).

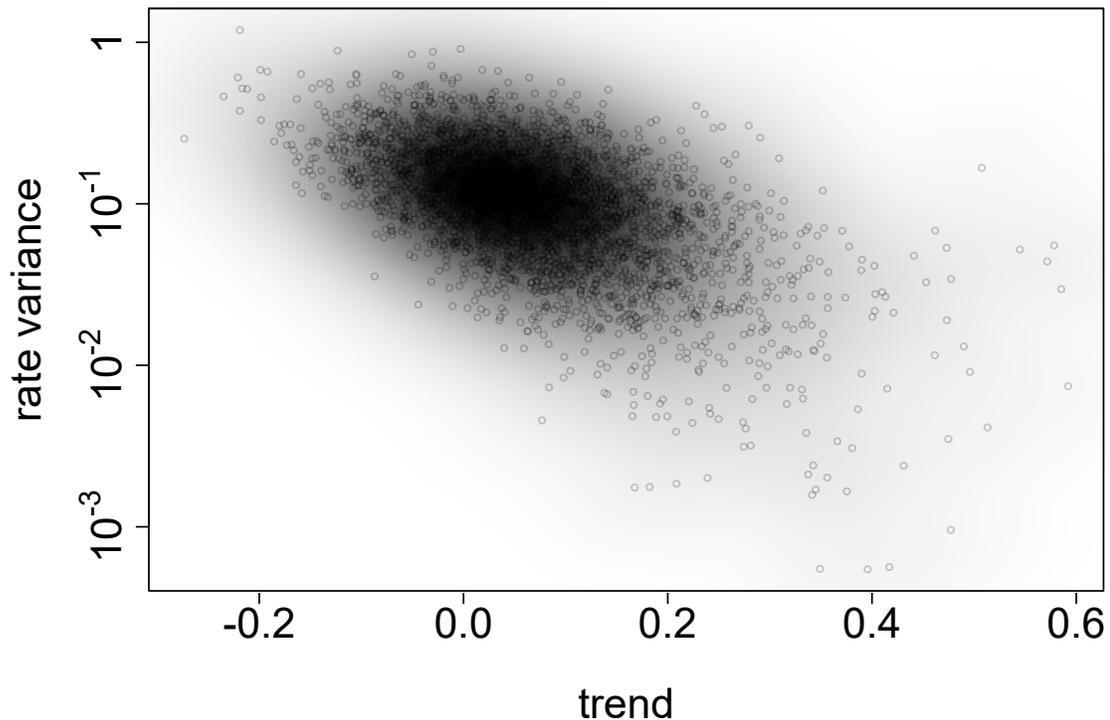


Figure 2C.1 Posterior samples of the trend and rate variance parameters inferred under the full, evolving rates model fit to the Monocalypt height data. Note the negative correlation between posterior samples of these parameters, indicating that Monocalypt height evolution is largely consistent with either rapid accumulation of random evolutionary rate variation (i.e., high rate variance, trend around 0) or accelerating evolutionary rates over time (i.e., low rate variance, positive trend).

Then, to render downstreams analyses more manageable, we thinned the resulting set of 3,484 contsimmaps to every tenth contsimmap, which was the maximum thinning rate whereby effective sample sizes (Geyer, 2011; Stan Development Team, 2019) of height values at each time point remained above 100. This procedure yielded the final set of 349 contsimmaps used in our empirical case study of Monocalypts. Conveniently, our procedure for generating contsimmaps under models fitted via the *evorates* package requires contsimmapping both the focal trait/factor (height in this case) as well as the rates at which the focal trait/factor evolved. Thus, to generate dummy factor histories, we simply simulated the evolution of a trait with starting trait values and contsimmapped rates identical to those in our final sample of 349 contsimmaps.

We now outline the details of our approach to generating contsimmaps based on posterior

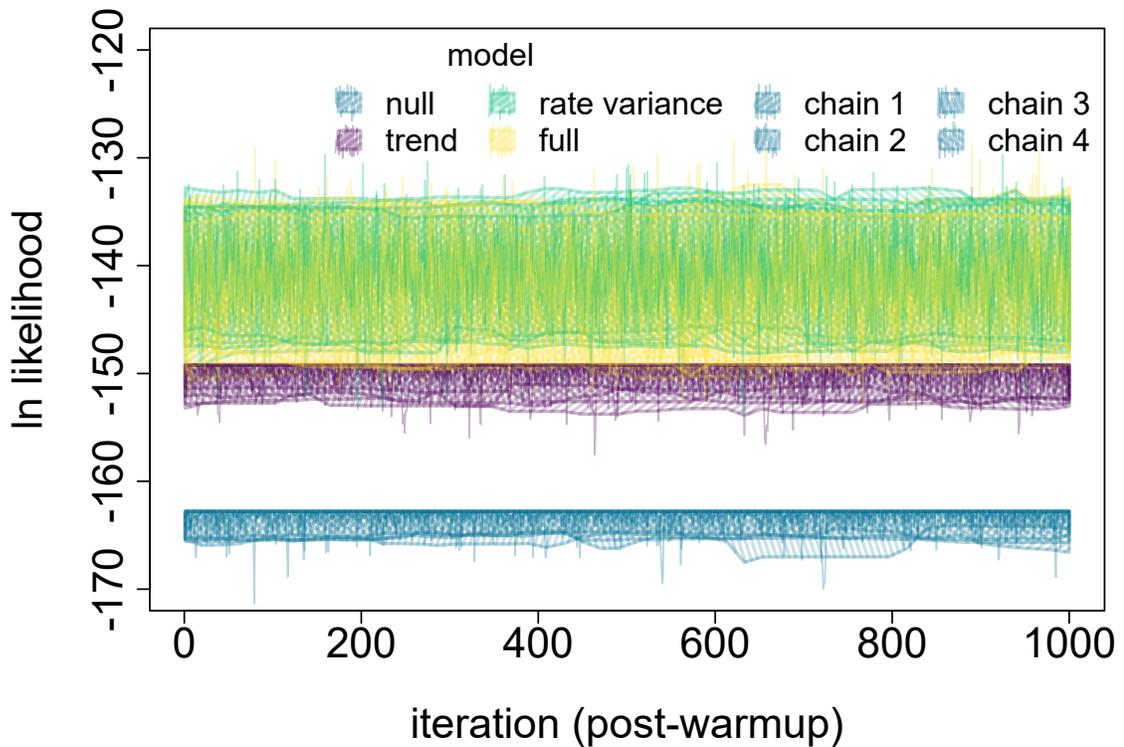


Figure 2C.2 Posterior traces of likelihoods of the Monocalypt height data under the null, constant-rate Brownian motion model; the trended, “early/late” burst model; the rate variance model whereby rates gradually diverge among lineages over time; and the full model including both an overall trend in rates over time and rate variance. The partially-shaded “ribbons” represent rolling 95% credible intervals of posterior likelihood samples. Different colors correspond to the different models, while the angle of the shading in each ribbon correspond to independent Hamiltonian Monte Carlo chains (note that the lines representing the actual traces of posterior likelihoods for independent chains are not distinguishable). Note that the ribbon corresponding to the full model (in light yellow) overlaps with those for the rate variance and (to a lesser extent) trend models (in a medium shade of green and dark purple, respectively).

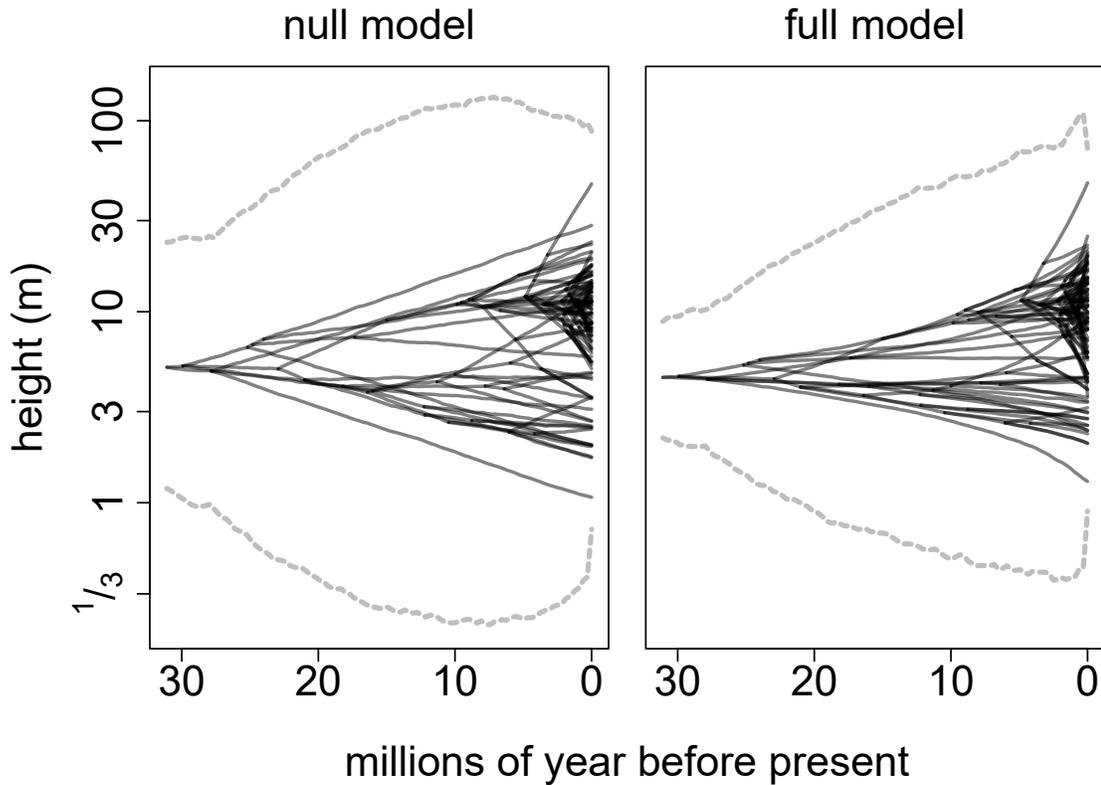


Figure 2C.3 Phenograms depicting the overall distribution of all 4,000 contsimmaps of *Monocalypt* height (i.e., prior to any filtering/thinning of samples) under both the null, constant-rate Brownian motion model and the full, evolving rates model including both an overall trend in rates over time and rate variance. The thick, solid lines depict the “central tendency” of the contsimmaps, corresponding to mean height value across all contsimmaps for each lineage in the *Monocalypt* phylogeny. To depict the overall range of the contsimmapped height values—represented by the lighter, dashed lines—we first calculated 95% confidence intervals of sampled height values for each lineage, then took the minimum and maximum interval bounds at 100 equally-spaced time points spanning the height of the phylogeny. Note that ancestral reconstructions were generally more precise under the full model.

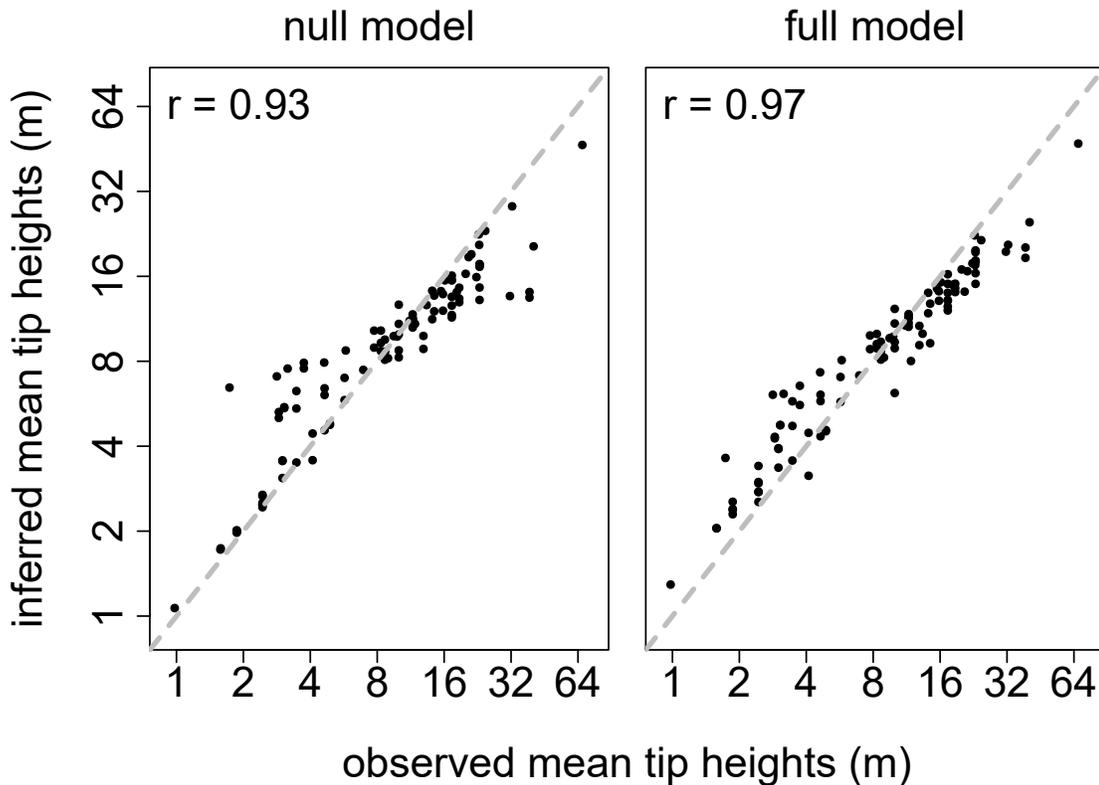


Figure 2C.4 Inferred versus observed mean species heights at the tips of the *Monocalypt* phylogeny under both the null, constant-rate Brownian motion model and the full, evolving rates model including both an overall trend in rates over time and rate variance. Correlation coefficients, denoted  $r$ , for the relationship between inferred and observed heights are provided in the top left corner of each plot. Note that inferred mean species heights under the full model better align with the dashed line, which depicts the position of observed mean species heights along the vertical axis.

samples of parameters under arbitrary models fitted via the *evorates* package. Briefly, models implemented in the *evorates* package generally work by assuming the rate parameter of a Brownian motion model of trait evolution itself “evolves” according to a constant-rate geometric Brownian motion process (i.e., the natural log of the evolutionary rates evolve according to a constant-rate Brownian motion process). Thus, the model infers a “rate evolution process” controlled by a trend ( $\mu_{\sigma^2}$ ) parameter determining whether rates tend to decrease or increase over time, as well as a rate variance ( $\sigma_{\sigma^2}^2$ ) parameter controlling how quickly random variation in rates accumulates over time. Additionally, the model estimates both the rate at the root of the phylogeny ( $\sigma_0^2$ ) and average rates of evolution along each branch of the phylogeny, which we term “branchwise rates” or  $\overline{\sigma_e^2}$ , where

$e$  denotes the index of a particular edge. At a broad level, to generate contsimmaps under such models, we implemented a two-step procedure whereby the inferred rate evolution process and branchwise rates are first used to generate contsimmaps of evolutionary rates of a user-specified resolution, which are in turn used to generate contsimmaps of trait (or factor) values of the same resolution by treating the rate contsimmaps as high-resolution regime maps (i.e., a regime for each time point spanning the preceding time interval). We summarize this process graphically in Fig. 2C.5.

Unfortunately, because the time-average of a geometric Brownian motion process has no closed-form probability distribution (Lepage et al., 2007), there is no simple and/or exact method for sampling rate values at arbitrary time points across a phylogeny (i.e., to generate rate contsimmaps) conditional on the inferred rate evolution process and branchwise rates. That being said, we can take advantage of the crude but effective approximation of geometric Brownian motion time-averages used by the *evorates* package itself (see APPENDIX 1B; see also Dufresne, 2004; Welch and Waxman, 2008; Martin et al., 2023). Namely, we can assume each branchwise rate is the sum of a trend and noise component, with the noise component assumed to follow the distribution of geometric, rather than arithmetic, time-averages of an untrended (i.e.,  $\mu_{\sigma^2} = 0$ ) geometric Brownian motion process. Under this assumption, we can derive straight-forward normal distributions describing how the natural log of rates at the nodes of a phylogeny are distributed. To start out, let  $\ln \sigma_{\tau_1}^2$  and  $\ln \sigma_{\tau_2}^2$  represent the natural log of the starting and ending values of a geometric Brownian motion process over an interval of length  $t$  with trend  $\mu_{\sigma^2}$ , rate variance  $\sigma_{\sigma^2}^2$ , and time-average  $\overline{\sigma^2}$ . Then the distributions of  $\ln \sigma_{\tau_1}^2$  and  $\ln \sigma_{\tau_2}^2$  under the aforementioned approximation are given by:

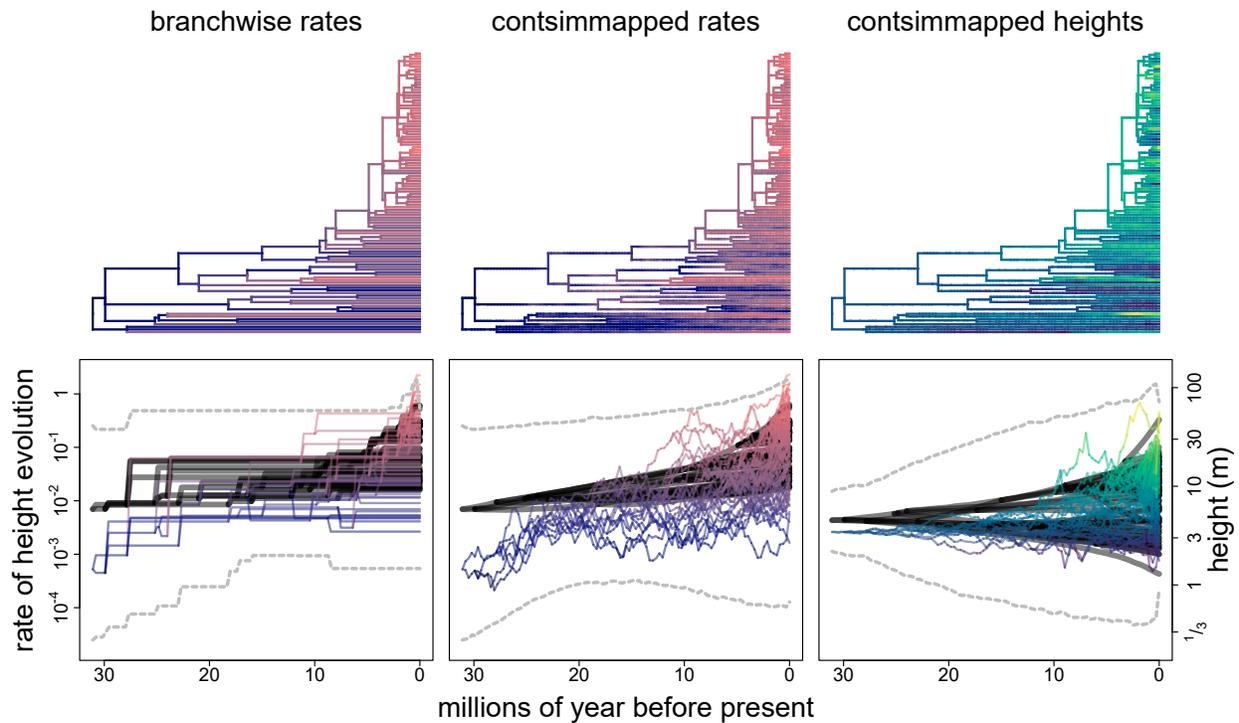


Figure 2C.5 A brief graphical summary of our procedure for generating contsimmaps under evolving rates models, using the full, evolving rates model fit to the Monoclypt height data as an example. The phenograms on bottom depict the distributions of: 1) inferred average rates of trait evolution along each branch (i.e., branchwise rates), which form the main input for our procedure (left); 2) the contsimmapped rate values sampled conditional on the inferred rate evolution process and branchwise rates, which are generated by the first step of our procedure (middle); and 3) contsimmapped height values sampled conditional on the observed height data at the tips and contsimmapped rates, which are generated by the second step of our procedure (right). Each phenogram consists of the overall mean rates/heights for each lineage in solid black lines, as well as a single example posterior sample/contsimmap in thinner lines, which are colored according to their position along the vertical axis (different color gradients are used to visually distinguish rates from heights). The phylograms on top provide an alternate view of the example posterior samples/contsimmaps, with rate/height values represented solely by these color gradients rather than positions along vertical axes. The lighter, dashed lines in each phenogram on bottom depict the overall range of the rate/heights, and were derived from calculating 95% confidence intervals along each lineage, then taking the minimum and maximum interval bounds at 100 equally-spaced time points spanning the height of the phylogeny.

$$\ln \sigma_{\tau_1}^2 \sim N \left( \ln \bar{\sigma}^2 - \beta, \frac{\sigma_{\sigma^2}^2}{3} \right) \quad (1)$$

$$\ln \sigma_{\tau_2}^2 \sim N \left( \ln \bar{\sigma}^2 - \beta + \mu_{\sigma^2} t, \frac{\sigma_{\sigma^2}^2 t}{2} \right) \quad (2)$$

$$\beta = \begin{cases} 0 & \text{if } \mu_{\sigma^2} = 0 \\ \ln |e^{\mu_{\sigma^2} t} - 1| - \ln |\mu_{\sigma^2}| - \ln t & \text{if } \mu_{\sigma^2} \neq 0 \end{cases} \quad (3)$$

Where  $N(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . We can generalize these results to define normal distributions for the natural log of rates at each node in a phylogeny given branchwise rates along each edge based on basic algebra of normal random variables. Specifically, Let  $\sigma_e^2$  represent the rate at the node immediately descending from edge  $e$  (not to be confused with  $\bar{\sigma}_e^2$ , the branchwise rate along edge  $e$ ), then:

$$\ln \sigma_e^2 \sim N \left( \alpha_e^{-1} \left( \frac{2(\ln \bar{\sigma}_e^2 - \beta_e)}{t_e} + \sum_{d \in \text{des}(e)} \frac{3(\ln \bar{\sigma}_d^2 - \beta_d)}{t_d} \right) + \mu_{\sigma^2} \tau_e, \alpha_e^{-1} \sigma_{\sigma^2}^2 \right) \quad (4)$$

$$\alpha_e = \frac{2}{t_e} + \sum_{d \in \text{des}(e)} \frac{3}{t_d} \quad (5)$$

$$\beta_e = \begin{cases} 0 & \text{if } \mu_{\sigma^2} = 0 \\ \ln |e^{\mu_{\sigma^2} \tau_e} - e^{\mu_{\sigma^2} \tau_{\text{anc}(e)}}| - \ln |\mu_{\sigma^2}| - \ln t_e & \text{if } \mu_{\sigma^2} \neq 0 \end{cases} \quad (6)$$

Where  $t_e$  now denotes the length of edge  $e$ ,  $\tau_e$  the height (i.e., distance from root) of the node immediately descending from edge  $e$ ,  $\text{des}(e)$  the indices for all edges immediately descending from edge  $e$ , and  $\text{anc}(e)$  the index for the edge immediately ancestral to edge  $e$ . Based on these formulae, we implemented a root-to-tips or preorder traversal algorithm that generates complete rate contsimmaps by jointly sampling rate values at the time points along each edge conditional on each edge's branchwise rate and sampled rate value at the edge's immediately ancestral node. Under our approximation, the joint distribution of the natural log of rate values along a given edge follows a straight-forward multivariate normal distribution conditioned to have a particular

weighted sum given by the corresponding branchwise rate. Let  $\vec{\sigma}_e^2$  and  $\vec{\tau}_e$  denote  $n_e$ -length vectors of the rate values and corresponding time points along edge  $e$  (i.e., the  $n_e$ th entries correspond to the node immediately descending from edge  $e$ ;  $\sigma_e^2 = \sigma_{e,n_e}^2$  and  $\tau_e = \tau_{e,n_e}$ ). Ignoring the condition that  $\vec{\sigma}_e^2$  must have a particular weighted sum, the distribution of  $\vec{\sigma}_e^2$  is given by:

$$\ln \vec{\sigma}_e^2 \sim \text{MVN} \left( \frac{(\mathbb{E}[\ln \sigma_e^2] - \ln \sigma_{\text{anc}(e)}^2)(\vec{\tau}_e - \tau_{\text{anc}(e)})}{t_e} + \ln \sigma_{\text{anc}(e)}^2, \Sigma_e \right) \quad (7)$$

$$\Sigma_{e,i,j} = \frac{\sigma_{\sigma^2}^2 (\tau_e - \max\{\vec{\tau}_{e,i}, \vec{\tau}_{e,j}\})(\min\{\vec{\tau}_{e,i}, \vec{\tau}_{e,j}\} - \tau_{\text{anc}(e)}) + \text{Var}(\ln \sigma_e^2) \sqrt{\vec{\tau}_{e,i} \vec{\tau}_{e,j}}}{t_e} \quad (8)$$

Where  $\text{MVN}(\vec{\mu}, \Sigma)$  denotes a multivariate normal distribution with mean vector  $\vec{\mu}$  and variance-covariance matrix  $\Sigma$ ,  $\mathbb{E}[\ln \sigma_e^2]$  and  $\text{Var}(\ln \sigma_e^2)$  are given by Eq. (4), and  $\Sigma_{e,i,j}$  is the entry in the  $i$ th row and  $j$ th column of matrix  $\Sigma_e$ . Notably, for edges descending from the root node,  $\ln \sigma_{\text{anc}(e)}^2$  corresponds to the inferred root rate parameter,  $\ln \sigma_0^2$ , and  $\tau_{\text{anc}(e)}$  to 0. We can calculate an approximate branchwise rate for a given sample of  $\vec{\sigma}_e^2$  as the weighted sum:

$$\hat{\sigma}_e^2 = \frac{1}{t_e} \sum_{i=1}^n \vec{\sigma}_e^2 (\vec{\tau}_{e,i} - \tau_{e,i-1}) \quad (9)$$

Where  $\vec{\tau}_{e,0}$  is replaced with  $\tau_{\text{anc}(e)}$  in the above expression (or with 0 if  $e$  is an edge descending from the root node). Here, we assume each sampled rate at a given time point is constant along its preceding time interval, which greatly simplifies calculations and negligibly differs from more accurate interpolations of rates along an edge given a sufficiently dense set of time points (notably, we use the same approximation to calculate likelihoods of continuous factor-dependent Brownian motion models).

We use Markov chain Monte Carlo to sample from Eq. (7) under a highly informative prior conditioning the sampled branchwise rate,  $\hat{\sigma}_e^2$ , to approximately equal the inferred branchwise rate,  $\overline{\sigma}_e^2$ . Specifically, we place a normal prior on  $\ln \hat{\sigma}_e^2$  with mean  $\ln \overline{\sigma}_e^2$  and a user-specified variance or “tolerance”, which defaults to 0.001. Lower tolerance values will produce more accurate rate samples at the cost of increased computation time (and vice versa for higher values). To simplify this procedure, we use an “uncentered” parameterization and do not sample  $\vec{\sigma}_e^2$  directly. Instead,

we sample  $n_e$  independent standard normal random variables,  $\vec{z}$ , which are transformed to follow the distribution given by Eq. (7) via  $\Sigma_e^{\frac{1}{2}}\vec{z} + \mathbb{E}[\ln \vec{\sigma}_e^2]$ , where  $\Sigma_e^{\frac{1}{2}}$  denotes the lower triangular Cholesky decomposition of  $\Sigma_e$  (Betancourt and Girolami, 2019). We propose new samples of  $\vec{z}$  by simply adding  $n_e$  normal random variables with mean 0 and variance 0.01 to the previously sampled values of  $\vec{z}$ . For each edge, Markov chain Monte Carlo sampling of rate values is run for a user-specified maximum number of iterations (which defaults to 100,000), but terminates early if  $(\ln \widehat{\sigma}_e^2 - \ln \overline{\sigma}_e^2)^2$  is less than the user-specified tolerance value (i.e., the sampled branchwise rate is within a standard deviation of the normal prior's mean). So far, we have found that this procedure essentially never reaches the maximum number of iterations under our default settings, ensuring a close match between inferred branchwise rates and the sampled branchwise rates in the resulting rate contsimmaps.

Finally, after generating rate contsimmaps, we can again assume that the sampled rates at each time point are constant along the preceding interval, forming a high-resolution map of regimes corresponding to different rates for each time interval (i.e., all time points are critical time points). Then, we can simply use the contsimmapping algorithm described in subsection 2.2.1 to sample trait/factor values at all time points conditional on the contsimmapped rates and observed trait/factor data.

## CHAPTER 3

### A NEW APPROACH FOR INFERRING STATE-DEPENDENT VARIATION IN CONTINUOUS TRAIT EVOLUTION DYNAMICS

#### 3.1 Introduction

Both paleontological and, more recently, phylogenetic comparative studies demonstrate that variation in the tempo and mode of macroevolutionary processes like phenotypic evolution and lineage diversification is pervasive across the tree of life (Simpson, 1944; Gingerich, 2009; Jablonski, 2017; Sauquet and Magallón, 2018; Harmon et al., 2021), resulting in the uneven distribution of biodiversity across space and taxa observed today. Elucidating the mechanisms driving this apparent “evolutionary heterogeneity”—for example, novel ecological opportunities or variation in life history traits—is critical for understanding many patterns that have long puzzled biologists, like the anomalous hyperdiversity of some taxa compared to their closest relatives (e.g., flowering plants, bats, beetles; Davies et al., 2004; Crepet and Niklas, 2009; Brock Fenton and Simmons, 2015; Stork et al., 2015) or the elevated species richness/phenotypic diversity of tropical ecosystems versus temperate ones (Hillebrand, 2004; Stevens et al., 2006; Schumm et al., 2019; Diamond and Roy, 2023; Saupe, 2023). Accordingly, recent decades have seen the development of many robust and powerful methods for investigating potential drivers of heterogeneity in both lineage diversification (Maddison et al., 2007; FitzJohn, 2010; Goldberg et al., 2011; FitzJohn, 2012; Beaulieu and O’Meara, 2016; Rabosky and Goldberg, 2017; Vasconcelos et al., 2022) and discrete trait evolution dynamics (Beaulieu et al., 2013; Goldberg and Foo, 2020; Boyko and Beaulieu, 2021, 2023). However, despite the plethora of available methods for estimating evolutionary correlations between continuously-measured traits and other variables (Felsenstein, 1985; Martins and Hansen, 1997; Hansen et al., 2008; Bartoszek et al., 2012; Felsenstein, 2012; Tolkoﬀ et al., 2018; Hassler et al., 2022b), approaches for inferring heterogeneity in continuous trait evolution dynamics (e.g., rate shifts, changes in the frequency/magnitude of evolutionary “pulses” *sensu* Landis and Schraiber, 2017) continue to lag behind those for lineage diversification and discrete traits in several key ways.

Most available methods for estimating how some variable or “factor” (e.g., habitat, reproductive strategy) is associated with heterogeneity in continuous trait evolution dynamics rely on “sequential” approximations consisting of two main steps: 1) inferring the evolutionary history of the factor via ancestral state reconstruction, and 2) fitting factor-dependent continuous trait evolution models based on these reconstructions. Such approaches are cumbersome and exhibit notable biases under certain conditions (Revell, 2013; May and Moore, 2020; Boyko et al., 2023b). In particular, by completely ignoring continuous trait data in the first step, these methods tend to rely on rather improbable reconstructions of factor histories (Caetano et al., 2018; Boyko et al., 2023b). This is because the continuous trait data reflect past evolutionary processes which, assuming a statistically well-supported relationship between the factor and continuous trait evolution dynamics, provide substantial information regarding the factor’s evolutionary history beyond the factor data itself (which is nearly always limited to “phylogenetically sparse” measurements among extant/fossilized taxa). Furthermore, methods based on sequential approximations tend to simplistically assume all heterogeneity in continuous trait evolution is caused by only one to a few explicitly-measured factors. In reality, however, continuous trait evolution dynamics are presumably affected by a tangled web of countless interconnected and context-dependent factors, inevitably generating “residual” or “background” heterogeneity in evolutionary dynamics on top of any heterogeneity associated with the particular factors a given study happens to focus on (Cooper and Purvis, 2009; May and Moore, 2020; Boyko et al., 2023b; Tribble et al., 2023; see also Donoghue and Sanderson, 2015). In turn, these methods often misattribute residual heterogeneity caused by unconsidered and/or unobserved factors to measured factors instead, thereby yielding spurious support for researchers’ *a priori* hypotheses (May and Moore, 2020; Boyko et al., 2023a,b).

Critically, inferring evolutionary heterogeneity driven by unobserved factors is not as impractical as it may first seem. In fact, phylogenetic comparative methods for analyzing heterogeneity in evolutionary dynamics are often described as either “hypothesis-driven” approaches designed to test whether heterogeneity is associated with particular factors of interest, or “data-driven” ap-

proaches meant to detect and quantify heterogeneity based on the provided comparative data alone, independently of any *a priori* hypothesis. In other words, data-driven approaches are precisely designed to infer general evolutionary heterogeneity driven by unobserved factors. Accordingly, methods that integrate both hypothesis- and data-driven approaches have become increasingly popular in recent years, as they are able to account for evolutionary heterogeneity driven by both observed and unobserved factors (Beaulieu and O’Meara, 2016; Uyeda et al., 2018; May and Moore, 2020; Boyko and Beaulieu, 2023; Boyko et al., 2023b). Hidden Markov model-based (HMM) frameworks are particularly effective in this regard, allowing researchers to directly infer unobserved discrete variables or “hidden states” (presumably representing simplified summaries of various interconnected factors; see discussion in Boyko et al., 2023b) based on their apparent impact on lineage diversification and/or discrete trait evolution dynamics (Beaulieu et al., 2013; Beaulieu and O’Meara, 2016; Caetano et al., 2018; Goldberg and Foo, 2020; Vasconcelos et al., 2022; Boyko and Beaulieu, 2023). Ancestral state reconstruction can even be used to map the occurrence of hidden states across a phylogeny, enabling powerful explorations of how evolutionary dynamics have changed throughout the history of a clade (e.g., Beaulieu and O’Meara, 2016; Vasconcelos et al., 2022). Unfortunately, existing approaches for inferring heterogeneity in continuous trait evolution dynamics are not directly compatible with HMM frameworks, instead employing sequential approximations to infer general/residual heterogeneity by iteratively sampling possible distributions of heterogeneity (implicitly driven by unobserved factors) across a phylogeny (Eastman et al., 2011; Thomas and Freckleton, 2012; Rabosky et al., 2014; May and Moore, 2020; Pagel et al., 2022; Boyko et al., 2023b; Martin et al., 2023; Tribble et al., 2023). While such methods are certainly powerful tools for exploring variation in continuous trait evolution dynamics, they are typically computationally-expensive and often incompatible with hypothesis-driven approaches (but see May and Moore, 2020; Boyko et al., 2023b; Tribble et al., 2023).

Ultimately, to make accurate and robust conclusions regarding what mechanisms drive heterogeneity in continuous trait evolution dynamics across the tree of life, researchers need methods capable of “jointly” inferring heterogeneous continuous trait evolutionary processes along with the

history of factors—both observed and unobserved—potentially driving such heterogeneity. Notably, currently available state-dependent speciation and extinction (SSE) models meet all these criteria in regards to inferring links between discretely-measured factors or “states” (but see FitzJohn, 2010) and lineage diversification dynamics (Maddison et al., 2007; Beaulieu and O’Meara, 2016). Accordingly, the SSE modeling framework has become extremely popular among empirical researchers and method developers alike (Helmstetter et al., 2023). The publication of the original SSE model has been cited over 1,000 times (Maddison et al., 2007) and since been elaborated into a diverse array of more sophisticated models themselves often associated with highly-cited publications (FitzJohn et al., 2009; FitzJohn, 2010; Goldberg et al., 2011; FitzJohn, 2012; Goldberg and Igić, 2012; Magnuson-Ford and Otto, 2012; Beaulieu and O’Meara, 2016; Caetano et al., 2018; Freyman and Höhna, 2018; Herrera-Alsina et al., 2019; Nakov et al., 2019; Verboom et al., 2020; Vasconcelos et al., 2022). Thus, an analogous framework for State-dependent Continuous trait Evolution or “SCE” models would likely constitute a tremendously useful tool for empirical macroevolutionary research, not to mention continued development of more realistic continuous trait evolution models.

Here, we begin to address this methodological gap by developing a new phylogenetic comparative method for inferring variation in continuous trait evolution dynamics driven by observed and/or unobserved states (i.e., discretely-varying factors), implemented in an R package named *sce* (pronounced “ski”). Critically, our approach employs a novel “pruning algorithm” (Felsenstein, 1973) that efficiently accounts for all possible state histories given the observed state and continuous trait data, thereby avoiding explicit reconstruction of state histories as in similar methods based on sequential approximations. Through an extensive simulation study, we verify that this SCE modeling framework can be used to reliably detect and estimate variation in rates of continuous trait evolution from phylogenetic comparative data. Furthermore, we use SCE models to show that tropical environments are associated with higher rates of flower size evolution in sages (Lamiaceae: *Salvia* L.), providing insights into the possible processes underlying latitudinal gradients of increasing phenotypic diversity (and perhaps species richness; see Schemske, 2001) towards the

equator (Stevens et al., 2006; Schumm et al., 2019; Diamond and Roy, 2023). Overall, SCE models provide a powerful and flexible new framework for studying heterogeneity in continuous trait evolution dynamics, ultimately allowing researchers to more confidently and accurately elucidate how various factors drive shifts in phenotypic evolutionary processes across the tree of life.

## 3.2 Materials and Methods

The *sce* package provides tools for inferring state-dependent variation in the evolutionary dynamics of a single (i.e., univariate) continuous trait given a phylogeny and comparative state/continuous trait data associated with its tips. Currently, our implementation only allows for *up to* one state and continuous trait measurement per tip (note that missing state/trait measurements are allowed), though we plan to extend the method to handle multiple measurements per tip in the future. Using the *sce* package, researchers can build, fit, and compare various SCE models corresponding to particular hypotheses—for example, one might compare models assuming rates of flower size evolution either are constant, differ among habitats, vary according to some unobserved factor/hidden state, or are affected by *both* habitat and hidden states. Notably, the package additionally allows researchers to map probable states, traits, and rates onto a phylogeny under a fitted model via marginal ancestral state reconstruction (Yang, 2007; Hiscott et al., 2016).

### 3.2.1 Model and Implementation

Our new framework for inferring SCE models largely depends on a novel pruning algorithm for calculating the joint likelihood of continuous and discrete phylogenetic comparative data evolving under “Markov-modulated Lévy processes” (MMLPs). Briefly, MMLPs model the evolution of a continuous trait as a continuous-time random walk (of which Brownian motion or BM is a special case) whose parameters depend on discrete states, which themselves evolve via a continuous-time Markov chain (CTMC). To achieve this, we effectively treat states and continuous traits as a unified compound trait, and use a modified version of Felsenstein’s pruning algorithm for discrete traits (Felsenstein, 1973) to recursively calculate the likelihood of observed data conditional on the state and continuous trait value at the root of a phylogeny. To effectively clarify the details of our new pruning algorithm, we first outline some key notations and concepts by providing a general

introduction to pruning algorithms below.

Assume we have a rooted phylogeny with  $m$  branches, including  $n$  terminal “tip branches” with associated phenotypic data. Let  $\tau$  be an  $m$ -length vector of branch lengths (typically measured in units of time) and  $\phi_e(x_l)$  denote the likelihood that the node immediately descending from branch  $e$  exhibits phenotype  $x_l$  given the observed phenotypes of all its descendants. Hereafter, we represent these so-called “partial likelihoods” (*sensu* Hassler et al., 2022a) of all possible phenotypes for a given branch  $e$  as  $\phi_e(x)$ . Notably, the partial likelihoods for each tip branch are directly derived from observed phenotypic data. Generally speaking, pruning algorithms aim to recursively calculate partial likelihoods for all remaining (i.e., non-tip) branches in the phylogeny to compute the full likelihood of all observed phenotypic data under a given model of phenotypic evolution. Specifically, for non-tip branch  $e$ , assuming phenotypic evolution is conditionally independent among sister branches (an assumption common to nearly all trait evolution models),  $\phi_e(x_l)$  is given by:

$$\phi_e(x_l) = \prod_{d \in \text{des}(e)} \int \psi(y; x_l, \tau_d, \theta) \phi_d(y) dy \quad (1)$$

Where  $\text{des}(e)$  is a function that returns the indices of all branches immediately descending from branch  $e$  and  $\psi$  denotes the “transition probability function” under the given model of phenotypic evolution. More precisely,  $\psi(y; x_l, \tau_d, \theta)$  represents the probability of observing phenotype  $y$  after  $\tau_d$  units of time given a starting phenotype of  $x_l$  and evolutionary model parameters  $\theta$ . The expression  $\int \psi(y; x_l, \tau_d, \theta) \phi_d(y) dy$  represents the likelihood that the node immediately descending from branch  $e$  exhibits phenotype  $x_l$  given the observed phenotypic data associated with all descendants of branch  $d$  only, and is sometimes called the “branch-inflated” partial likelihood for branch  $d$  (Hassler et al., 2022a), here denoted  $\phi_d^*(x_l)$ . Because phenotypic evolution along sister branches is assumed to be conditionally independent, taking the product of  $\phi_d^*(x_l)$  across all descendants of branch  $e$  thus yields the partial likelihood of phenotype  $x_l$  for branch  $e$ . To calculate the overall likelihood given all observed data, the phylogeny is traversed in “postorder” (i.e., from tips to root), calculating  $\phi(x)$  for all branches according to Eq. 1, including an implicit “root branch”

indexed  $m + 1$ . The final likelihood  $\mathcal{L}$  is computed by integrating the partial likelihoods for the root branch multiplied by some root prior,  $\pi(x)$ . For this work, we use ‘‘FitzJohn’s root prior’’, marginalizing over all possible root phenotypes according to the relative probability they gave rise to the observed data (FitzJohn et al., 2009):

$$\mathcal{L} = \int \phi_{m+1}(x)\pi(x)dx \quad (2)$$

$$= \int \phi_{m+1}(x) \frac{\phi_{m+1}(x)}{\int \phi_{m+1}(z)dz} dx \quad (3)$$

$$= \frac{\int \phi_{m+1}(x)^2 dx}{\int \phi_{m+1}(x) dx} \quad (4)$$

Note that the use of integrals in imply that the phenotype  $x$  in this example varies continuously. In the case of discretely-valued phenotypes, these integrals are instead replaced by sums over all  $s$  possible states.

Unfortunately, outside of special cases, MMLPs do not admit simple closed-form expressions for partial likelihood or transition probability functions over continuous trait values. Thus, like other pruning algorithms for calculating the likelihood of complex continuous trait evolution models from phylogenetic comparative data, our implementation discretizes a specified range of continuous trait values into an equally-spaced grid of  $c$  points and tracks the partial likelihoods at each point to flexibly approximate arbitrary functions over continuous trait values (FitzJohn, 2010; Boucher and Démery, 2016; Hiscott et al., 2016; Boucher et al., 2018). We specifically represent the partial likelihoods for each branch  $e$  as an  $s \times c$  matrix  $\Phi_e$ . More concretely, if we let  $x$  represent an  $c$ -length vector of trait values corresponding to each grid point, the partial likelihood of exhibiting a continuous trait value of  $x_l$  and state  $j$  is thus given by the entry in the  $j$ th row and  $l$ th column of  $\Phi_e$  denoted  $\Phi_{e,j,l}$ .

Unfortunately, even with this flexible representation of partial likelihoods over both discrete states and continuous trait values, calculating transition probabilities under MMLPs (and branch-inflated partial likelihoods by extension) remains challenging. To simplify these calculations, we represent transition probability functions under Lévy processes,  $\psi(x; 0, t, \theta)$ , via their Fourier

transform or characteristic function,  $\hat{\psi}(\xi; 0, t, \theta)$ . This representation is particularly convenient because, as continuous-time random walks, Lévy process transition probability functions are “infinitely divisible probability distributions” (i.e., distributions describing the sum of an arbitrary number of independent and identically distributed random variables; Sato, 2013). According to the convolution theorem, the characteristic function describing the sum of random variables is given by the element-wise product of each variable’s individual characteristic function, and Lévy process characteristic functions thus always correspond to “infinitely divisible products” of the form:

$$\hat{\psi}(\xi; 0, t, \theta) = \exp[\zeta(\xi, \theta)t] \quad (5)$$

Where  $t$  denotes elapsed time and  $\zeta(\xi, \theta)$  is the so-called “characteristic exponent” of a given Lévy process with parameters  $\theta$  evaluated at  $\xi$  (note that the domain of the characteristic function,  $\xi$ , is distinct from that of its corresponding probability distribution function,  $x$ ). Ultimately, all Lévy processes admit a characteristic function representation of their transition probability function that consists of (relatively) simple exponential functions with respect to time for any particular value of  $\xi$  (stated more formally by the Lévy-Khintchine formula for characteristic functions of Lévy processes; Sato, 2013). The convolution theorem also greatly simplifies calculation of branch-inflated partial likelihoods for a branch  $d$ ,  $\phi_d^*(x)$ , which is equivalent to the distribution of the sum of two random variables—one distributed according to the transition probability function given a starting state of  $x = 0$  and another to the partial likelihoods for branch  $d$ . Thus, we can Fourier transform the partial likelihoods for branch  $d$  to yield its characteristic function  $\hat{\phi}_d(\xi)$ , multiply  $\hat{\phi}_d(\xi)$  by  $\hat{\psi}(\xi; 0, \tau_d, \theta)$ , and take the inverse Fourier transform of the result to finally yield  $\phi_d^*(x)$ .

Representing Lévy process transition probability functions as exponential functions with respect to time allows us to form an important bridge between Lévy process and CTMC transition probability functions, ultimately allowing us to efficiently compute transition probabilities under MMLPs. Briefly, under a CTMC, lineages switch between  $s$  states according to an  $s \times s$  matrix,

$Q$ , where  $q_{j,k}$  denotes the instantaneous rate of transition from state  $j$  to state  $k$ . Exponentiating  $tQ$  yields a new matrix,  $P$ , where  $p_{j,k}$  now denotes the probability of a lineage starting in state  $j$  ending up in state  $k$  after  $t$  time units (Pagel, 1994). Importantly, branch-inflated partial likelihoods for branch  $d$  under a CTMC process are given by the matrix-vector product  $\exp[\tau_d Q] \phi_d(x)$ , where  $\phi_d(x)$  now represents an  $s$ -length vector of partial likelihoods for each state (Felsenstein, 1973). In any case, transition probability functions under CTMCs, similarly to characteristic functions representations of Lévy process transition probability functions, are ultimately defined by a kind of “characteristic exponent”—in this case, the matrix  $Q$ . Importantly, because characteristic functions are linear transformations of their corresponding probability distribution functions (i.e., the Fourier transform of  $A\psi_1(x) + B\psi_2(x)$  is equal to  $A\hat{\psi}_1(\xi) + B\hat{\psi}_2(\xi)$ ; Pinsky, 2009), characteristic function representations of MMLP transition probability functions for any  $\xi$  value can be calculated by exponentiating the  $Q$  matrix describing the CTMC evolution of states with modified diagonal entries. Under normal CTMCs, the diagonal entries of  $Q$  are given by:

$$q_{j,j} = - \sum_{\substack{k=1 \\ k \neq j}}^s q_{j,k} \quad (6)$$

Which ensures that probabilities are conserved under the CTMC process by balancing the overall transition rate into state  $j$  with the transition rate out of state  $j$  (Pagel, 1994). In the case of MMLPs, we must modify these diagonals for a given  $\xi$  value by adding the corresponding characteristic exponent for the Lévy process governing continuous trait evolution in state  $j$ , denoted  $\zeta_j(\xi, \theta)$  (note that the exponents are added because  $\exp[\zeta_j(\xi, \theta)t] \exp[q_{j,j}t] = \exp[(\zeta_j(\xi, \theta) + q_{j,j})t]$ ).

Because we represent the partial likelihoods for each branch with  $s \times c$  matrices  $\Phi$ , we do not treat  $\xi$  as a continuous variable but, like  $x$ , as a  $c$ -length vector of grid points. More specifically, the  $l$ th entry  $\xi$  is given by:

$$\xi_l = -2\pi \frac{l-1-c \lfloor \frac{l+c/2-2}{c} \rfloor}{x_c - x_1} \quad (7)$$

Where  $\lfloor a \rfloor$  denotes the greatest integer less than or equal to  $a$ . Now, let  $R$  represent a “rate array” of  $c$   $s \times s$  matrices, with the  $l$ th matrix encoding the characteristic function representation of the MMLP transition probability function at  $\xi_l$ . The off-diagonal elements of all matrices in  $R$  are equal and given by the  $Q$  matrix describing the CTMC evolution of discrete states. On the other hand, the diagonal entries for the  $l$ th matrix are equal to:

$$r_{j,j,l} = \zeta_j(\xi_l, \theta) - \sum_{\substack{k=1 \\ k \neq j}}^s q_{j,k} \quad (8)$$

To calculate branch-inflated partial likelihoods for a given branch  $d$  under MMLPs, we exponentiate each matrix in  $R$  multiplied by the branch length  $\tau_d$ , resulting in a new array denoted  $\hat{\Psi}_d$ . Next, we use the highly efficient fast Fourier transform algorithm (FFT) to compute the discrete Fourier transform of each row in the partial likelihood matrix  $\Phi_d$  to yield the characteristic function representation of the partial likelihood matrix denoted  $\hat{\Phi}_d$ . Then the  $l$ th matrix of  $\hat{\Psi}_d$  is multiplied by the  $l$ th column of  $\hat{\Phi}_d$  to compute the characteristic function representation of the branch-inflated partial likelihood matrix  $\hat{\Phi}_d^*$ . Next, we use the inverse FFT to convert each row of  $\hat{\Phi}_d^*$  back to its normal representation, yielding the branch-inflated partial likelihood matrix  $\Phi_d^*$ . Lastly, the partial likelihood matrix for a given branch  $e$  is simply the element-wise product of  $\Phi_d^*$  for all branches  $d$  immediately descending from  $e$ . This finally allows us to compute the necessary partial likelihoods to carry out a pruning algorithm.

Intriguingly, representing partial likelihoods in matrix form theoretically allows for incredibly flexible initialization of our pruning algorithm—that is, partial likelihood matrices for tip branches may be arbitrarily specified in accordance with observed phenotypic data. For example, provided sufficient within-tip sampling, one could even initialize the partial likelihood matrices for each tip with kernel density estimates of continuous trait distributions for any given discrete state. Nonetheless, for simplicity, we assume here that continuous traits for each tip  $e$  are normally distributed with mean  $\eta_e$  and variance  $\varepsilon_e^2$  across all discrete states, while the probability that a given tip exhibits discrete state  $j$  is given by the  $j$ th entry of an  $s$ -length vector  $\rho_e$ , denoted  $\rho_{e,j}$ . Under these assumptions, the characteristic function representation for the partial likelihood matrix for

tip branch  $e$ ,  $\hat{\Phi}_e$ , is given by (DasGupta, 2011):

$$\hat{\Phi}_{e,j,l} = \rho_{e,j} \exp \left[ i \left( \eta_e - \frac{x_c + x_1}{2} \right) \xi_l - \frac{\epsilon_e^2 \xi_l^2}{2} \right] \quad (9)$$

Where, critically,  $i$  is *not* a parameter or variable and actually denotes the imaginary unit (i.e.,  $\sqrt{-1}$ ).

Notably, our method for efficiently calculating partial likelihoods, branch-inflated partial likelihoods, and transition probabilities all assume continuous trait evolution exhibits “periodic boundary conditions”—a rather technical yet important caveat (Bowman and Roberts, 2011). Practically speaking, this means that continuous trait evolution “wraps around” on itself when it hits the boundaries of the specified grid (i.e.,  $x_1$  to the left and  $x_c$  to the right), such that low trait values are unrealistically “teleported” over to high trait values and vice versa. Fortunately, this problem is quite easily and effectively managed by simply extending the grid of trait values well beyond the range of observed data. In our current implementation, given an  $n$ -length vector of mean trait values for each tip  $\eta$ , we derive a “primary grid” of  $c/2$  points spanning from  $\min \eta - \omega(\max \eta - \min \eta)$  to  $\max \eta + \omega(\max \eta - \min \eta)$ . By default, we set  $\omega$  to 0.5 to expand the bounds of observed continuous trait data by 50% of the overall range of trait values on either side. The remaining  $c/2$  grid points come from symmetrically “zero-padding” partial likelihoods—that is, appending  $c/4$ -length vectors of 0s to either side of the partial likelihood vectors (Bowman and Roberts, 2011). By forcing these extreme trait values to be associated with partial likelihoods of 0 prior to each Fourier transformation, the risk of periodic boundary conditions influencing model inference is rendered effectively negligible, though we recommend against setting  $\omega$  to small values below  $\sim 0.05$ - $0.1$ .

In the current work, we focus on the simplest MMLP—namely, Markov-Modulated Brownian motion (MMBM). The characteristic exponent of a BM process is given by (DasGupta, 2011):

$$\zeta(\xi_l, \sigma^2, \mu) = i\mu \xi_l - \sigma^2 \xi_l^2 / 2 \quad (10)$$

Where  $\sigma^2$  and  $\mu$  correspond to the evolutionary rate and trend, respectively, of the BM process (notably, because the BM process must be propagated “backwards” in time for the pruning

algorithm, the drift parameter is multiplied by -1 for pruning algorithm calculations). While evolutionary trends under homogeneous BM models of trait evolution are generally unidentifiable from ultrametric phylogenetic comparative data (i.e., phylogenies with tips corresponding to contemporaneous taxa only, which the current work focuses on), preliminary tests of our approach with simulated data suggest that state-dependent *differences* among evolutionary trends are identifiable under MMBM models. Nonetheless, for simplicity, we assume  $\mu = 0$  throughout the simulation and empirical case studies presented here, leaving investigation of state-dependent evolutionary trend inference for future work.

We implemented our pruning algorithm using the C/C++ libraries *FFTW3* (Frigo and Johnson, 2005) and *Armadillo* (Sanderson and Curtin, 2016, 2019), interfaced via the R packages *Rcpp* (Eddelbuettel and Francois, 2011) and *RcppArmadillo* (Eddelbuettel and Sanderson, 2014). See APPENDIX 3B for further details on how we practically manage the efficiency and numerical stability of our pruning algorithm.

### 3.2.2 Hidden States and Hypothesis Testing

By calculating the likelihood of *both* state and continuous trait data under a joint evolutionary process, our new SCE modeling framework enables straight-forward inference of unobserved or hidden states associated with heterogeneity in continuous trait evolution dynamics. Ultimately, this is due to the fact that our approach allows continuous trait data to directly influence the likelihood of state histories, enabling the inference of evolutionary histories of unobserved states based solely on their apparent impact on continuous trait evolution dynamics. More concretely, inferring hidden states generally requires “splitting” the observed state data into multiple observed-hidden state combinations and allowing continuous trait evolution parameters like rates or trends to vary based on these additional states. Importantly, each tip is initialized such that partial likelihoods are identical across hidden states. As an example, if one was studying whether flower size evolutionary dynamics differ between tropical and temperate lineages, inferring a model with two hidden states would entail splitting each state into two new ones, conventionally labeled  $\text{tropical}_A$ ,  $\text{tropical}_B$ ,  $\text{temperate}_A$ , and  $\text{temperate}_B$ . Then, for any given tip, the partial likelihoods for states

tropical<sub>A</sub>/tropical<sub>B</sub> and temperate<sub>A</sub>/temperate<sub>B</sub> are given by the partial likelihoods for the original tropical and temperate states, respectively.

Ultimately, incorporating hidden states into hypothesis-testing frameworks allows for more comprehensive and realistic tests of evolutionary hypotheses. To test whether the evolution of a continuous trait is largely homogeneous or varies according to an observed and/or unobserved discrete variable, one may compare the fit of several candidate SCE models with differing constraints on which parameters of the continuous trait evolution process (e.g., evolutionary rates, trends) are allowed to vary across observed and/or hidden states. We demonstrate an example of this hypothesis-testing approach through the simulation study and empirical examples outlined below. Critically, despite the fact that hidden states alter how partial likelihoods are initialized in pruning algorithms, the likelihoods/information criteria associated with models fit to the same phylogenetic comparative data are directly comparable in this context whether or not they include hidden states.

### 3.2.3 Simulation Study

To assess the performance of our new approach for inferring SCE models, we tested whether our method could reliably detect and quantify observed and/or hidden state-dependent rates of continuous trait evolution from phylogenetic comparative data simulated under MMBM models. Hereafter, for both brevity and consistency with existing terminology in the field (e.g., Beaulieu and O’Meara, 2016; Boyko et al., 2023b), we refer to observed state-dependent rate variation as simply “state-dependent rate variation” and to unobserved/hidden state-dependent rate variation as “state-independent rate variation”. In general, we used the R package *phytools* (Revell, 2012) to first simulate both pure-birth phylogenies (all scaled to have heights of 1) and associated discrete state data, then used our new *sce* package to simulate associated continuous trait data under several different patterns of state-dependent rate variation.

For each simulation, we simulated CTMC evolution of four discrete states labeled “0A”, “0B”, “1A”, and “1B”, with the numeric prefixes denoting observed states 0 and 1 and letter suffixes denoting unobserved/hidden binary states A and B. All simulated CTMCs assumed that: 1) transitions between observed states do not depend on unobserved states and vice versa, 2) transitions in

both observed and unobserved states cannot occur in the same instant (i.e.,  $q_{0A,1B} = 0$ ,  $q_{1A,0B} = 0$ , etc.), and 3) transitions among unobserved states always occur with rate 1 and are “symmetric”—that is, transitions from A to B occur at the same rate as transitions from B to A. With the exception of the third assumption (which we made to simplify the design of our simulation study), these are all common simplifying assumptions made to render hidden state inference more statistically tractable (e.g., Beaulieu and O’Meara, 2016). Transitions among observed states 0 and 1 followed one of three alternative evolutionary dynamics: 1) “slow” symmetric transitions of rate 1, 2) “fast” symmetric transitions of rate 4, and 3) “asymmetric” transitions whereby transitions from 0 to 1 occur with rate 2 while transitions from 1 to 0 occur with rate 0.4. For each simulation, we sampled root states from the stationary distribution associated with a simulation’s given transition rate matrix. To prevent simulations from exhibiting too few state transitions to reliably infer state-(in)dependent rate variation, we repeated each CTMC simulation until it met the following criteria: 1) each state is represented by at least 10% of tips in the phylogeny and 2) each state is transitioned into along 5 or more distinct branches (allowing for no more than 1 exception; i.e., at least 3 states must exhibit this property).

We simulated continuous trait evolution according to one of five distinct patterns of state-dependent and/or state-independent rate variation: 1) “constant” whereby rates of continuous trait evolution are equal across all states observed or unobserved, 2) “state-independent” whereby rates only depend on unobserved states A and B, 3) “completely state-dependent” whereby rates only depend on observed states 0 and 1, 4) “strongly state-dependent” whereby rates primarily depend on observed states while also being influenced by unobserved states, and 5) “weakly state-dependent” whereby rates primarily depend on unobserved states while also being influenced by observed states. Across all simulation conditions, states 1 and B were associated with higher rates than states 0 and A, respectively, with one notable exception. To investigate how asymmetric state transitions affect inference of state-dependent rates, we swapped the rates associated with state 0 and 1 for some simulations with both asymmetric transition rates and unequal rates of continuous trait evolution across states 0 and 1. The lowest rate in all simulations was always set to 1, while the

higher rate for state-independent and completely state-dependent simulations was set to 8. For strongly state-dependent simulations, we set rates to 1, 4, 8, and 16 in states 0A, 0B, 1A, and 1B, respectively, but set these rates to 1, 8, 4, and 16 instead for weakly state-dependent simulations.

Ultimately, we defined 18 simulation conditions—15 corresponding to every possible combination of the 3 state transition dynamics and 5 state-dependent/independent rate variation patterns, plus an additional 3 conditions with asymmetric transition rates and rates of continuous trait evolution for states 0 and 1 swapped as described above. To also investigate how our approach was influenced by increasing sample size, we simulated phylogenies with either 50, 100, or 200 tips. We simulated 20 replicates for each condition and phylogeny size, ultimately yielding 1,080 simulated phylogenies and associated comparative datasets. Note that hidden state information was discarded prior to model fitting and analysis, such that states 0A/0B were converted to 0 and states 1A/1B to 1. To analyze each simulation, we fit 8 different models via maximum likelihood inference using our *sce* package—four models assuming symmetric transition rates among observed states plus four otherwise identical models instead assuming asymmetric transition rates. For simplicity, all models assumed transition rates among hidden states are symmetric. The four models in each category (symmetric versus asymmetric transition rates) consisted of: 1) a null model assuming constant rates across observed states without hidden states, 2) a state-independent model assuming constant rates across observed states with two hidden states, 3) a state-dependent model allowing rates to vary across observed states without hidden states, and 4) a full model allowing rates to vary across observed states while also including two hidden states. Like the simulation settings, all models assumed transition rates among hidden states did not depend on observed states and vice versa, and that transitions in both observed and hidden states cannot occur simultaneously. To reduce the time it took to analyze all these simulated datasets, we discretized continuous trait data with a relatively coarse grid of 512 points. While 512 grid points seems to suffice to preliminary analyses, we generally recommend using a grid 2,048 points for final parameter estimates and likelihood calculations (the FFT algorithm is most efficient with grid resolutions which are a power of two; Frigo and Johnson, 2005). For the purposes of model fitting, we assumed the con-

tinuous trait at each tip followed a normal distribution centered at the simulated continuous trait value with a small standard deviation equal to  $1/100^h$  the range of all continuous trait values for a given simulation.

We used the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LBFGS) algorithm implemented in the C++ library NLOPT (interfaced through the R package *nloptr*; Johnson, 2021; Ypma et al., 2022) to fit all models. Notably, the LBFGS algorithm requires information about the gradient of the likelihood surface, which we calculated via simple finite difference approximations (see APPENDIX 2B for an example of this approach to gradient approximation). While other numerical optimization algorithms do not require gradient calculations, we found these algorithms to be slower and less reliable in preliminary tests of our approach. Because transition and evolutionary rate parameters must be positive, we estimated all parameters on a natural log scale. For each model fit, we ran the LBFGS algorithm at least 5 times from initial parameter estimates uniformly sampled between -2.5 and 2.5 ( $\sim 0.1$ -10), taking whichever run achieved the highest maximum likelihood estimate. Likelihood surfaces for models including hidden states occasionally appeared to be multimodal and/or exhibit flat “ridges”, rendering numerical optimization more challenging—thus, we ran the algorithm 10 rather than 5 times for models with hidden states. We also bounded parameter estimates between -30 ( $\sim 1 \times 10^{-13}$ ) and 7 ( $\sim 1,000$ ) to prevent the algorithm from both getting “stuck” in likelihood ridges that may occur when parameter estimates are effectively equal to 0 and/or running into overflow issues if parameter estimates grew too large.

We analyzed the results of our simulation study through both “model selection” and “parameter inference”-based approaches. For the model selection-based approach, we calculated sample size corrected Akaike Information Criteria (AICc) for each model fit. Because simulated data often failed to exhibit overwhelming support for one particular model based on differences in AICc ( $\Delta$ AICc), we instead focused on analyzing variation in the AICc weights associated with different models across simulation conditions. In particular, we used “relative AICc weights” (hereafter RAWs), which we define here as the average AICc weight for all models supporting some hypoth-

esis (e.g., state-dependent rate variation) divided by itself plus the average AICc weight for all models supporting an alternative hypothesis (e.g., constant rates). RAWs thus provide normalized measures between 0 and 1 of the support for some hypothesis (more technically, RAWs correspond to the inverse logit transform of log “evidence ratios” *sensu* Burnham and Anderson, 2002). We specifically calculated RAWs comparing: 1) state-independent versus null models, 2) state-dependent/full versus null models, 3) state-dependent/full versus state-independent models, and 4) asymmetric versus symmetric models. Note that we *averaged* rather than summed AICc weights because some of these RAWs were “unbalanced” in comparing four models to two (Burnham and Anderson, 2002; Kittle et al., 2008). For the parameter inference-based approach, we simply model-averaged parameter estimates according to (non-relative) AICc weights and compared inferred parameter values to simulated ones. For consistency among fitted models and simulated parameters, we assumed the inferred hidden state associated with the highest average rate corresponded to the simulated unobserved state B and the other inferred hidden state to A. Note that the transition rate among hidden states could only be model-averaged across four of the eight models which explicitly inferred hidden states.

### 3.2.4 Empirical Example

To demonstrate potential empirical applications of our new method for inferring state-dependent rates of continuous trait evolution, we tested whether tropical lineages of sage (Lamiaceae: *Salvia* L.) exhibit higher rates of flower size evolution in accordance with the Biotic Interactions Hypothesis (BIH). The BIH predicts that the comparatively stable climatic conditions of tropical ecosystems causes biotic factors to drive more variation in evolutionary fitness relative to abiotic factors (Dobzhansky, 1950; Schemske, 2001; Schemske et al., 2009). Thus, assuming biotic selective pressures are typically more heterogeneous than abiotic ones, the characteristic spatiotemporal scale of evolutionary processes should shrink towards the tropics, resulting in more frequent and rapid bouts of evolutionary divergence among populations (i.e., more intense/pronounced “co-evolutionary mosaics”; see Thompson, 2005). As such, the BIH suggests that traits particularly important in mediating biotic interactions—like flower morphology—should exhibit elevated rates of

evolution among tropical lineages.

For this analysis, we combined the recently-generated, time-calibrated maximum clade credibility phylogeny of sages and associated corolla length measurements from Moein et al. (2023) with geographic occurrence data from Kriebel et al. (2019). To more thoroughly contextualize the results of our analyses comparing corolla length evolution among tropical and temperate sage lineages, we used additional discrete trait data from Moein et al. (2023) to also examine whether other key sage traits—specifically, active staminal lever presence/absence and woodiness/herbaceousness—affect corolla length evolution. We dropped 5 tips from the phylogeny corresponding to species outside the genus *Salvia*, as well as an additional 37 tips lacking any associated data, yielding a final phylogeny consisting of 334 tips with corresponding corolla (328 tips), lever (314 tips), and/or woodiness (302 tips) data. We coded missing discrete trait measurements by assigning identical partial likelihoods to all possible discrete states, similarly to hidden states, and coded missing continuous trait measurements by assigning equal partial likelihoods to all grid points corresponding to different corolla length measurements. The geographic occurrences from Kriebel et al. (2019) provided data for 289 of these tips, but we supplemented these data with coordinates downloaded from the Global Biodiversity Information Facility (GBIF) for any species lacking coordinate data. We used the R package *CoordinateCleaner* (Zizka et al., 2019) to remove likely erroneous and/or duplicate records from the combined data under the package’s default settings, with the exception of keeping records potentially corresponding to country centroids, which were the only records available for a substantial number of tips in the phylogeny. Ultimately, this yielded nearly 71,000 occurrence records for 304 tips, with a median of 10 occurrences per tip (ranging from 1 to 25,788; 57 or ~20% of tips were only associated with a single record).

There are, unfortunately, numerous partially conflicting strategies for classifying geographic locations as either tropical or temperate (Feeley and Stroud, 2018). Because the BIH suggests climatic stability ultimately drives increased evolutionary rates towards the tropics, we chose to use ratio of annual to daily temperature ranges (hereafter “isothermality”) to delineate geographic occurrences as either tropical or temperate. First, we calculated isothermality based on the definition

of Feeley and Stroud (2018) for each occurrence record using 2.5 minute resolution WorldClim rasters (Fick and Hijmans, 2017) via the R package *geodata* (Hijmans et al., 2023). In this case, isothermality is a strictly positive variable with values below 1 corresponding to tropical conditions whereby daily temperature fluctuations are comparable to or exceed seasonal variation in temperature. After log-transforming isothermality such that it ranged from  $-\infty$  to  $\infty$ , we computed isothermality ranges for each tip as the 5 and 95% quantiles of each tip’s empirical isothermality distribution, classifying each tip as either tropical if the midpoint of the range fell below  $\ln 1 = 0$  and temperate otherwise (Fig. 3A.1). To assess how sensitive our results were to this particular classification scheme, we also devised a more conservative coding strategy that took the extremes of each tip’s isothermality range into account. Notably, however, the widths of isothermality ranges tended to increase with the number of occurrence records among tips with roughly 20 or fewer records (Fig. 3A.2). Thus, we calculated the mean range width across all tips with more than 20 records and symmetrically expanded the ranges for all tips with 20 or fewer records to at least this width. We then classified tips as tropical only if both the upper extreme and midpoint of their range fell below  $\ln 1.25 \approx 0.22$  and  $\ln 0.75 \approx -0.29$ , respectively. Conversely, tips were only classified as temperate if the lower extreme and midpoint of their range exceeded  $\ln 0.75 \approx -0.29$  and  $\ln 1.25 \approx 0.22$ , respectively. All other tips (50 out of the 304 with occurrence records) were coded as missing because they exhibited no strong climatic preference, either occurring under both tropical and temperate conditions or exclusively occurring right around the isothermality cut-off point of  $\ln 1 = 0$  (Fig. 3A.3).

For each discrete variable (“strict” and “conservative” codings of tropical/temperate states, presence/absence of staminal levers, and woodiness), we inferred joint models of state and natural log-transformed corolla length evolution following procedures largely identical to those used for the simulation study, with a few key exceptions. First, in addition to the 8 models described in the simulation study, we assessed evidence for asymmetric transition rates among hidden states via 4 additional models: state-independent and full models assuming symmetric transitions among observed states but asymmetric transitions among hidden states, plus two otherwise identical models

assuming asymmetric transitions among both observed and hidden states. Second, because we lack data on corolla length measurement errors/within-tip variation, we allowed all models to simultaneously estimate a “tip error” parameter (i.e.,  $\epsilon$  in Eq. 9; see Landis and Schraiber, 2017) rather than fixing this parameter *a priori* based on the overall range of continuous trait data as in the simulation study. Lastly, to better ensure the accuracy of all inferences, we increased the resolution of the continuous trait value grid from 512 to 2,048 and repeatedly fit each model 20 times rather than 5 or 10.

We analyzed our empirical results following the procedures used for the simulation study as well, calculating AICc weights and corresponding model-averaged estimates of transition rates among states (again defining hidden state B as whichever inferred hidden state is associated with the highest average rate of continuous trait evolution), rates of corolla length evolution, and tip error in corolla length measurements. We also calculated RAWs measuring the evidence for state-independent/state-dependent rate variation versus constant rates, state-dependent versus state-independent rate variation, and asymmetric versus symmetric transition rates among observed/hidden states. Additionally, we computed ancestral state, rate, and corolla length estimates under each model using marginal ancestral state reconstruction, summarizing the results for each of the four discrete variables by model-averaging estimated state probabilities, rates, and corolla lengths at each node and tip in the phylogeny.

### **3.3 Results**

#### **3.3.1 Simulation Study**

Overall, our simulation study demonstrates that our new SCE modeling framework can accurately detect and quantify variation in rates of continuous trait evolution from phylogenetic comparative data—whether such variation is associated with observed or unobserved discrete variables. Interestingly, sample size corrected Akaike Information Criteria (AICc) suggested data simulated under relatively “simple” conditions (e.g., constant rates, state-independent rate variation only) generally did not exhibit strong evidence against models implying more complex patterns of rate variation, instead tending to yield equivocal support for a range of simple to more complex models.

Accordingly, selecting best-fitting models via common “rules of thumb” (e.g., lowest AICc,  $\Delta\text{AICc} < 2$ ) resulted in incorrectly rejecting simpler models about 10-15% of the time. Fortunately, using relative AICc weights (RAWs) to measure the evidence in favor of particular *hypotheses* rather than models resulted in much better error rates (Figure 3.1). More specifically, using a RAW threshold of  $>85\%$  in favor of “more complex” hypotheses (e.g., variation in rates, state-dependent rates, asymmetric transitions among states) yielded error rates of about 5% or lower. Critically, many datasets simulated under state-independent rates yielded strong support for state-dependent over constant rate models, but not state-dependent over state-independent models, directly demonstrating the importance of accounting for the possible influence of unobserved discrete variables in testing for associations between discrete variables and rates of continuous trait evolution. Beyond error rates, our method performed well in terms of its statistical power to detect rate variation. While our approach sometimes struggled to detect rate variation—particularly state-independent rate variation—from 50-tip datasets, its power generally grew rapidly with increasing sample size. As might be expected, increasing ratios of state-independent to state-dependent rate variation (i.e., weakly state-dependent simulations compared to completely state-dependent ones) lead to more equivocal support for state-dependent over state-independent models. Intriguingly, faster transition rates seemed to slightly decrease power rates for detecting state-dependent rate variation, consistent with previous work based on a more approximate method for inferring state-dependent rates of continuous trait evolution (Revell, 2013). Our method’s power to detect asymmetric transition rates (i.e., transitions from one state to another occurring more frequently than the reverse) among observed states was noticeably weaker than its power to detect rate variation, aligning with similar findings in the context of state-dependent SSE models (Beaulieu and O’Meara, 2016). Nonetheless, our approach could detect asymmetric transition rates more sensitively from data simulated under state-dependent rates, likely because apparent variation in evolutionary rates across the phylogeny provided additional information on the evolutionary history of observed states (see Boyko et al., 2023b).

Overall, our method yielded rather unbiased and accurate parameter estimates which grew

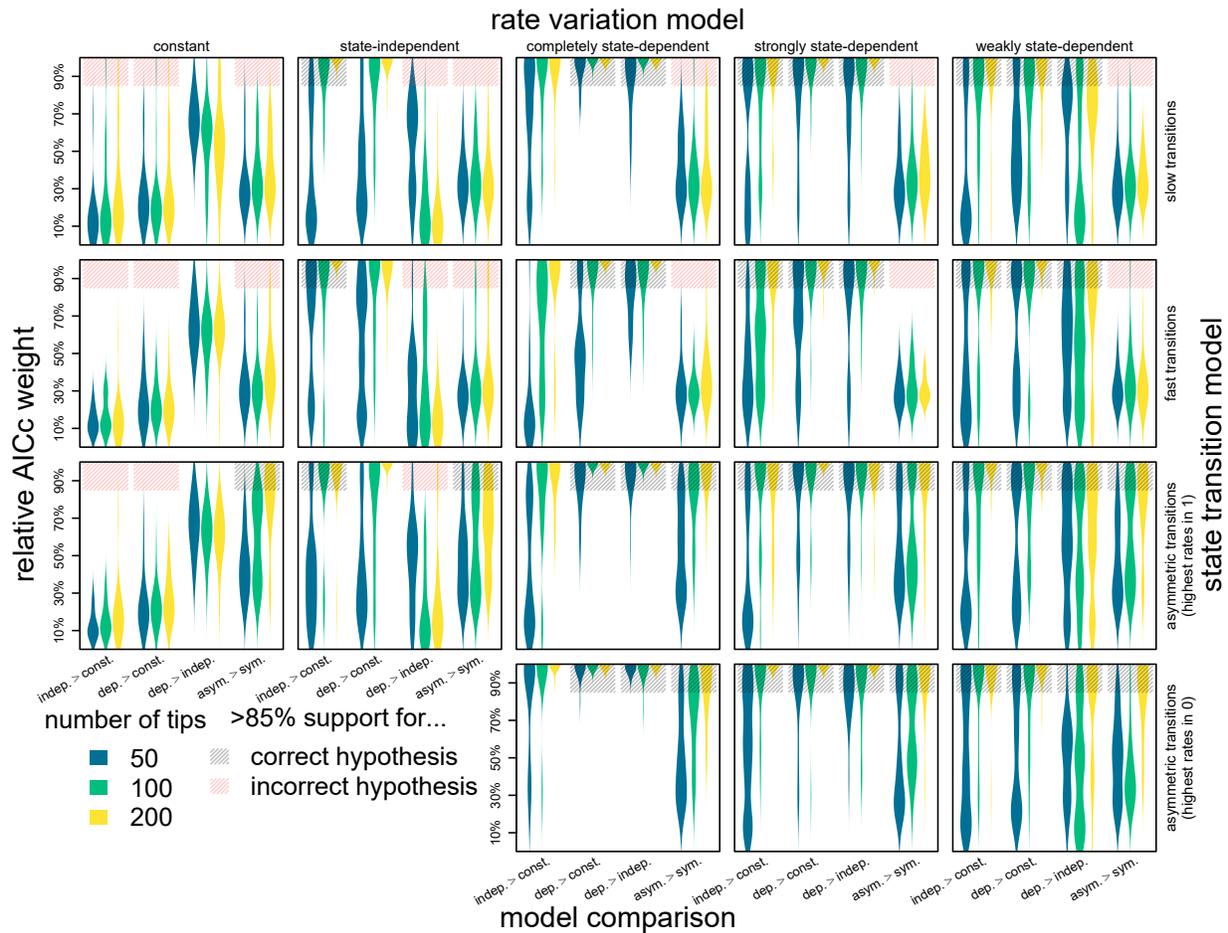


Figure 3.1 Distributions of relative sample size corrected Akaike Information Criterion (AICc) weights (i.e., average AICc weight associated with models supporting one hypothesis versus another, normalized to vary between 0 and 1; hereafter RAWs) measuring the support for several key evolutionary hypotheses across different simulation conditions based on our novel approach to inferring state-(in)dependent continuous trait evolution models. We specifically quantified the evidence that: 1) rates of continuous trait evolution vary according to an unobserved discrete variable (indep. > const.), 2) rates vary according to the observed discrete variable (dep. > const.), 3) rates vary according to the observed discrete variable rather than or in addition to an unobserved variable (dep. > indep.), and 4) transition rates from observed state 0 to 1 differ from those for 1 to 0 (i.e., transition rates are asymmetric; asym. > sym.). Different plots correspond to distinct simulation conditions, with different sample sizes (i.e., number of tips; distributions in lighter, warmer colors correspond to larger sample sizes) and hypotheses arrayed along each plot's x-axis. Hatched boxes correspond to RAW values > 85%, which indicate substantial support for either correct (in darker gray) or incorrect (in lighter red) hypotheses depending on the given simulation conditions. Hatched boxes were omitted in cases where a particular hypothesis is irrelevant given the simulation conditions.

more precise with increasing sample size (Figs. 3.2-3.3). Interestingly, in some cases, the method inferred hidden states which were very unlikely to occur anywhere across a phylogeny, resulting in “outlier” rate estimates (represented by the long, skinny tails of some distributions and  $\times$  symbols in Fig. 3.2) because the rate parameters associated with these states had virtually no effect on model likelihoods. Fortunately, such situations are not especially problematic—ancestral state reconstruction may be used to quickly verify whether states associated with anomalously low or high rate estimates are actually likely to have occurred anywhere on a given phylogeny. Unsurprisingly, inferred state-dependent rate differences tended to be more accurate and precise than state-independent rate differences. In fact, state-independent rate estimates often exhibited a specific pattern of bias for smaller 50-tip datasets, with differences among hidden states within a given observed state “collapsing” to 0 (e.g., rate estimates for states 0A and 0B being biased towards the overall average rate for state 0). Fortunately, this bias largely disappears for larger sample sizes as state-independent rate variation could be inferred with greater confidence (see in particular rate estimates for simulations with state-independent rates and slow transition rates in Fig. 3.2). A similar pattern occurred for rate differences between observed states under some simulation conditions with fast transition rates among observed states, aligning with results from previous work (Revell, 2013). Analogously, transition rate estimates for smaller datasets simulated under asymmetric transition rates tended to exhibit some bias towards inferring more symmetric rates, consistent with the general difficulty of confidently detecting asymmetric transition rates (Fig. 3.1). Notably, inferred parameters for transition rates among hidden states varied from fairly precise to exceptionally imprecise depending on strength of state-independent rate variation, as inference of hidden states is directly tied to the apparency of their effect on rates of continuous trait evolution—for example, hidden transition rate estimates for data simulated under completely state-dependent rates varied over six orders of magnitude regardless of sample size. Thus, unsurprisingly, simulations generally must have sufficiently strong state-independent relative to state-dependent rate variation for accurate inference of hidden state transition rates.



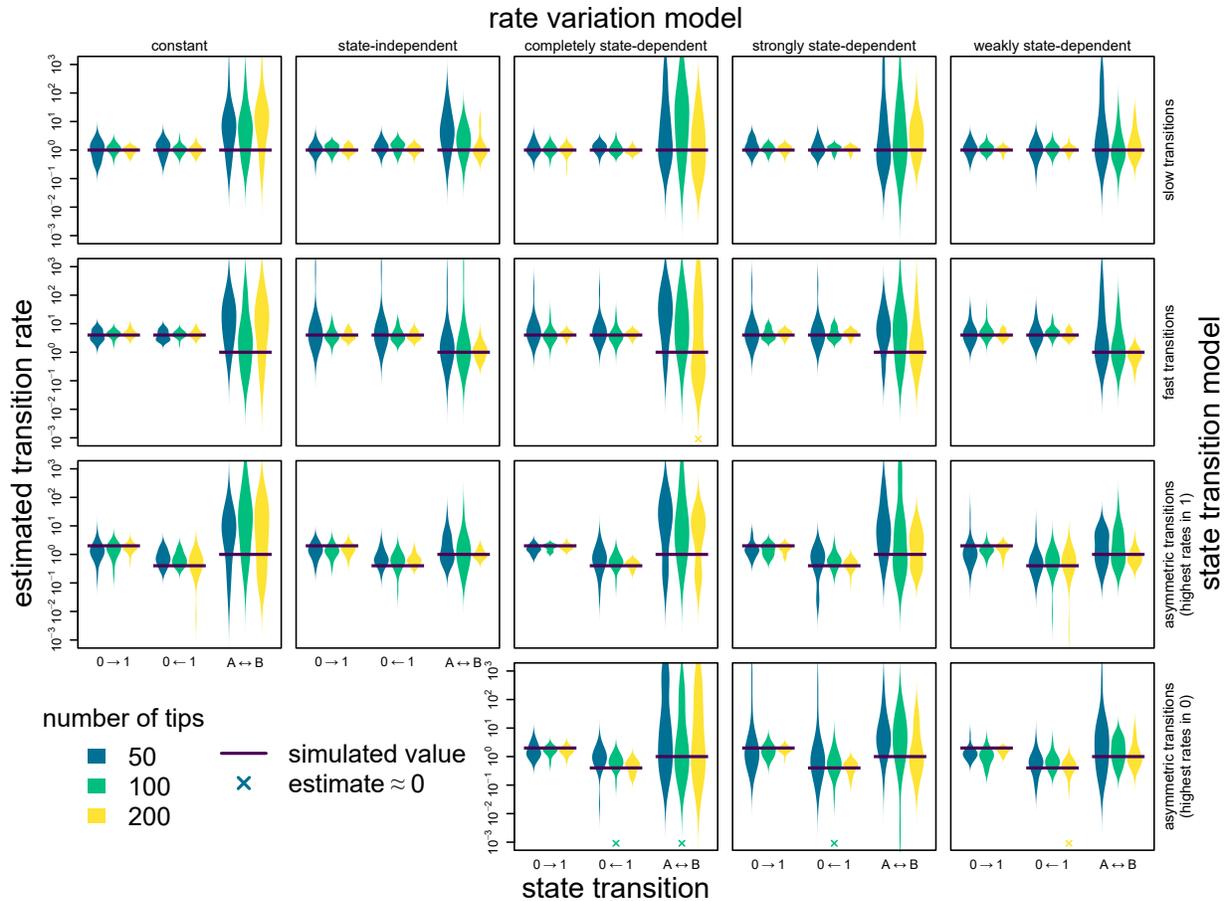


Figure 3.3 Distributions of model-averaged transition rates among states (based on sample size corrected Akaike Information Criterion—or AICc—weights) across all simulation conditions estimated under our novel approach to inferring state-dependent continuous trait evolution models. Different plots correspond to distinct simulation conditions, with different sample sizes (i.e., number of tips; distributions in lighter, warmer colors correspond to larger sample sizes) and state transitions arrayed along each plot’s x-axis. Horizontal lines depict the values of simulated transition rate parameters under different simulation conditions. We used  $\times$  symbols at the bottom of plots to indicate five cases where transition rate estimates fell below  $10^{-3}$  (ranging from about  $3 \times 10^{-5}$  to  $4 \times 10^{-4}$ ). Outside of these cases, estimated transition rates never exceeded the plot boundaries, though some came quite close (note that we set an upper bound of  $\exp[7] \approx 10^3$  on parameter estimates), causing some distributions to appear “cut off” at their extremes.

### 3.3.2 Empirical Example

We applied our new approach for inferring state-dependent variation in rates of continuous trait evolution to assess whether tropical lineages of sage exhibit higher rates of flower size (measured via natural log-transformed corolla length) evolution than temperate lineages as predicted by the BIH. Based on RAWs, we found substantial support for elevated rates of corolla length evolution among tropical sages (i.e., relevant RAWs around  $>85\%$ ; see Table 3.1; see also Table 3A.1 for associated AICc weights), with estimated rates among tropical lineages around double to triple those for temperate lineages (Table 3.2). This was true regardless of whether we used a strict (Fig. 3.4) or conservative (Fig. 3A.4) scheme for classifying extant sages as tropical versus temperate, and stands in stark contrast to the results for other discrete traits (i.e., staminal levers and woodiness; Figs. 3A.5 and 3A.6), which do not exhibit any apparent association with rates of corolla length evolution. On the other hand, evidence for state-independent rate variation was decidedly equivocal with a RAW of around  $56\%$ . Reassuringly, such results were consistent across models of different discrete traits, as would be expected because discrete and continuous trait evolutionary processes are completely independent under models completely state-independent rate variation. Indeed, excepting transition rates among observed states, parameter estimates under these models were extremely consistent across different discrete traits (compare rows 3, 7, 9, and 11 within/across Tables 3A.2–3A.5).

Beyond variation rates of corolla length evolution, the only discrete trait to yield notable evidence for asymmetric transition rates was woodiness, with transitions to herbaceousness occurring at nearly twice the rate of transitions to woodiness among sages, agreeing with previous results (Moein et al., 2023). With regards to hidden states, while our raw parameter estimates suggest transitions to the “fast” hidden state B (i.e., higher rates of corolla length evolution) occur at a rate some three times higher than that for transition into the “slow” hidden state A, evidence for asymmetric transition rates among hidden states based on RAWs never exceeded  $50\%$ , presumably due to the general lack of evidence for state-independent rate variation in the first place. Accounting for the association between tropicality and rates of corolla length evolution eroded any support

for asymmetry in hidden state transition rates even further. Inferred tip error in log corolla length measurements (i.e., variation in log corolla length due to measurement error/within-tip variation) remained rather consistent across all models at  $\sim 0.3$ , roughly corresponding to an error of about  $\pm 30\%$  around raw corolla length measurements. However, intriguingly, inferred tip errors were slightly yet consistently lower under models accounting for associations between rates and tropicality, suggesting that other models misattributed some signals of this rate heterogeneity to increased corolla length measurement error instead (see tip error/ $\epsilon$  estimates in Tables 3A.2–3A.5).

Table 3.1 Relative sample size corrected Akaike Information Criterion (AICc) weights (i.e., average AICc weight associated with models supporting one hypothesis versus another, normalized to vary between 0 and 1; hereafter RAWs) measuring the support for evolutionary hypotheses based on joint models of corolla length and discrete trait evolution among sages (Lamiaceae: *Salvia* L.). Each column corresponds to one of the four discrete traits we analyzed: 1-2) alternative “strict” and “conservative” codings (refer to subsection 3.2.4 for details) of tropicality versus temperateness (strict/conserv. trop.), 3) the presence/absence of staminal levers (lever pres.), and 4) woodiness versus herbaceousness (woodiness). Each row corresponds to a particular hypothesis we measured support for: 1) rates of corolla length evolution vary according to an unobserved discrete variable (indep. > const.), 2) rates vary according to the observed discrete trait (dep. > const.), 3) rates vary according to the observed discrete trait rather than or in addition to an unobserved variable (dep. > indep.), 4) transition rates from observed state 0 (i.e., temperateness, staminal lever absence, herbaceousness) to 1 (i.e., tropicality, lever presence, woodiness) differ from those for 1 to 0 (i.e., transition rates are asymmetric; obs. asym. > sym.), and 5) transition rates from the “slow” (i.e., lower rate of corolla length evolution) hidden state A to the “fast” state B are asymmetric (hid. asym. > sym.). A RAW of  $\sim 85\text{-}90\%$  or more indicates strong evidence in favor of a given hypothesis.

hypothesis	discrete trait			
	strict trop.	conserv. trop.	lever pres.	woodiness
indep. > const.	56%	56%	56%	55%
dep. > const.	91%	96%	23%	31%
dep. > indep.	89%	95%	19%	26%
obs. asym. > sym.	35%	31%	26%	88%
hid. asym. > sym.	37%	35%	49%	48%

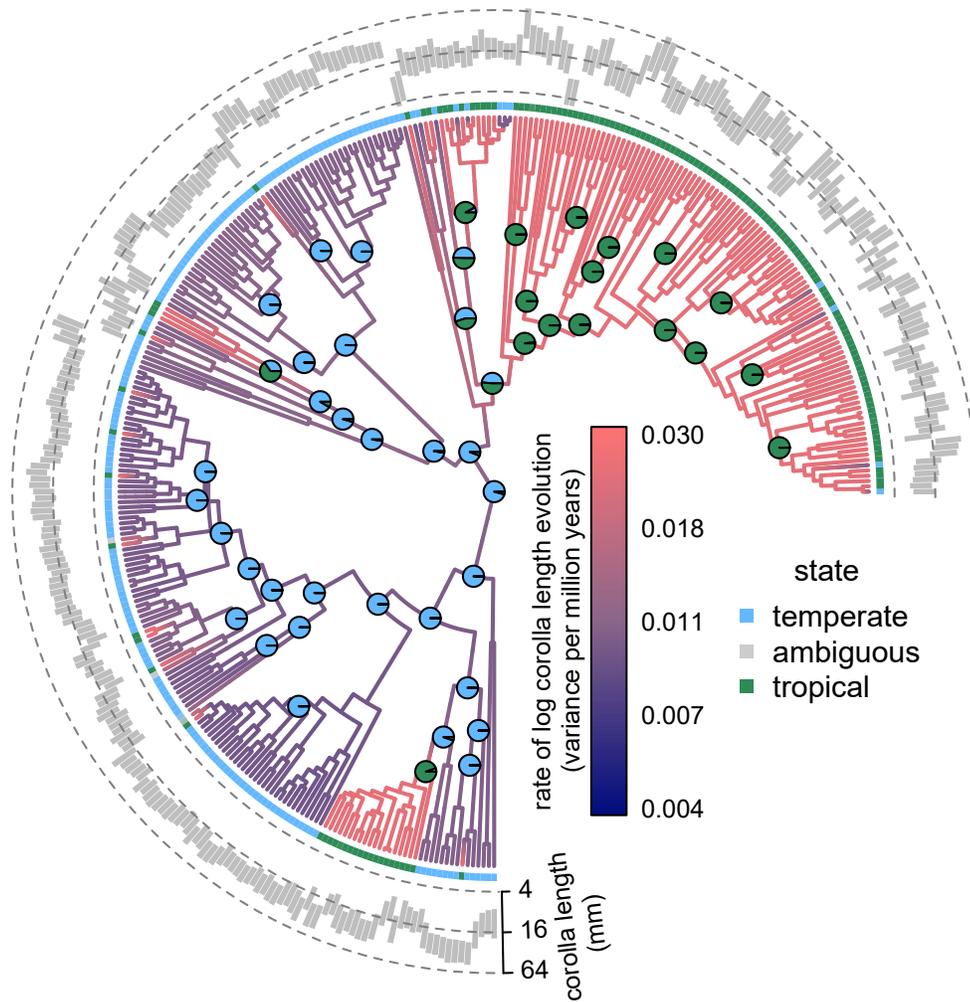


Figure 3.4 Phylogram depicting model-averaged marginal ancestral rate and state estimates (based on sample size corrected Akaike Information Criterion—or AICc—weights) based on our joint analysis of corolla length evolution and temperate-tropical transitions (strictly-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The color of branches correspond to inferred rates of corolla length evolution, with darker, cooler and lighter, warmer colors denoting relatively slow and fast rates, respectively. Pie charts at select nodes depict the probability that a given node tended to occur in either temperate (light blue) or tropical (dark green) environments. Because we lacked data for some tips and accounted for uncertainty in tropicality and corolla lengths, we also depict inferred tropicality probabilities (via colored boxes; light gray indicates even chances of being tropical or temperate) and 95% confidence intervals on corolla lengths (via gray bars) arrayed along the tips.

Table 3.2 Model-averaged parameter estimates (based on sample size corrected Akaike Information Criterion—or AICc—weights) based on joint models of corolla length and discrete trait evolution among sages (Lamiaceae: *Salvia* L.). Each column corresponds to one of the four discrete traits we analyzed: 1-2) alternative “strict” and “conservative” codings (refer to subsection 3.2.4 for details) of tropicality versus temperateness (strict/conserv. trop.), 3) the presence/absence of staminal levers (lever pres.), and 4) woodiness versus herbaceousness (woodiness). Each row corresponds to a particular parameter. Parameters denoted  $q_{x,y}$  represent transition rates from state  $x$  to state  $y$  per million years. States 0 and 1 refer to observed states (i.e., 0 = temperateness/staminal lever absence/herbaceousness, 1 = tropicality/lever presence/woodiness) while states A and B refer to “slow” (i.e., lower rate of corolla length evolution) and “fast” hidden states, respectively. Parameters denoted  $\sigma_{xw}^2$  refer to rates of log-transformed corolla length evolution (i.e., increase in variance per million years) in observed state  $x$  and hidden state  $w$ . Lastly,  $\varepsilon$  denotes “tip error”—the inferred standard deviation of (presumably normally-distributed) log-transformed corolla length measurements across all tips due to measurement error and/or within-tip phenotypic variation.

parameter	discrete trait			
	strict trop.	conserv. trop.	lever pres.	woodiness
$q_{0,1}$	0.018	0.008	0.024	0.040
$q_{1,0}$	0.016	0.007	0.024	0.076
$q_{A,B}$	0.018	0.014	0.034	0.032
$q_{B,A}$	0.013	0.015	0.011	0.010
$\sigma_{0A}^2$	0.009	0.009	0.006	0.007
$\sigma_{0B}^2$	0.011	0.011	0.019	0.020
$\sigma_{1A}^2$	0.022	0.026	0.007	0.006
$\sigma_{1B}^2$	0.025	0.027	0.020	0.021
$\varepsilon$	0.292	0.287	0.303	0.303

### 3.4 Discussion

Here, we outlined and demonstrated the capabilities of a new approach for inferring how discrete variables affect continuous trait evolution dynamics based on a novel pruning algorithm for directly calculating the likelihood of phylogenetic comparative data under a joint evolutionary process of both discrete state and continuous trait evolution. Unlike other methods for inferring these SCE models, our approach avoids relying on explicit reconstructions of state histories, rendering the method relatively efficient and convenient to use. Overall, our simulation study verifies that our new framework for fitting SCE models not only yields largely accurate parameter estimates, but also exhibits both acceptable error rates and relatively high power for detecting variation in rates of continuous trait evolution. A particular strength of our framework is its ability to account for residual heterogeneity in continuous trait evolution dynamics caused by unobserved discrete variables

or hidden states, enabling more accurate and robust parameter inference and evolutionary hypothesis testing (May and Moore, 2020; Boyko et al., 2023b; Tribble et al., 2023). Our simulation study concretely demonstrate this fact, showing that residual heterogeneity in rates of continuous trait evolution are frequently mistaken for state-dependent heterogeneity when only constant-rate and completely state-dependent models are considered (Fig. 2.2).

### **3.4.1 Increased rates of flower size evolution among tropical sages**

Using our new SCE modeling framework, we found that tropical sage lineages exhibit higher rates of flower size evolution than temperate lineages, consistent with predictions of the BIH. In fact, the two main subclades predominantly consisting of tropical sage taxa in our analyses (Fig. 3.4) are rather unusual among sage lineages for exhibiting multiple evolutionary shifts from bee to bird pollination (Kriebel et al., 2019), and bird pollination has been explicitly linked to larger flower sizes in sages (Wester et al., 2020; Moein et al., 2023). While more research is needed to confidently determine what mechanisms underlie elevated rates of flower size evolution among tropical sages, this association between tropical environments and elevated rates of both pollinator interaction and flower size evolution among sages is rather striking and certainly seems to agree with some of the key predictions of the BIH. Alternatively, such elevated rates of flower size evolution among tropical sage lineages may result from higher lineage diversification rates in the tropics driving increased incomplete lineage sorting and/or hybridization among sages. A single species tree is cannot fully describe the complex interrelationships of clades exhibiting high rates of introgression and hybridization, distorting expected patterns of phenotypic similarity across species and inflating evolutionary rate estimates (Mendes et al., 2018; Hibbins and Hahn, 2021; Hibbins et al., 2023). While our new method cannot explicitly account for these “reticulate” evolutionary processes, one can approximately model them by integrating comparative analyses over a sample of “gene trees” (Hibbins et al., 2023). To this end, our new approach to inferring SCE models theoretically makes integrating comparative analyses over multiple possible tree topologies easier by avoiding repeated sampling of discrete state histories over different topologies. This should make sampling parameters that fit the observed data under a variety of different topologies much easier

and computationally feasible, as the most likely parameters of a continuous trait evolution model can vary widely depending on the assumed state history, an issue presumably only exaggerated by different overall tree topologies (e.g., Caetano and Harmon, 2017; Boyko et al., 2023b). In any case, future ecological and microevolutionary work comparing pollination interactions, selection on floral traits, and/or population genetic patterns between tropical and temperate sage taxa may help more precisely elucidate the mechanisms driving increased rates of flower size evolution among tropical sages.

While much macroevolutionary research has investigated whether speciation and/or extinction rates differ among temperate and tropical lineages, comparatively little research has examined whether rates of phenotypic evolution are elevated among tropical lineages. Those that have investigated this question have yielded mixed results, from those broadly in agreement with the BIH (Schumm et al., 2019; Chartier et al., 2021), to those finding no consistent differences in rates among temperate/tropical lineages (Drury et al., 2021), to even those finding opposing patterns of increased rates of trait evolution in temperate ecosystems (Hipsley et al., 2014). Our results here notably agree with a previous study demonstrating that heathers and allies (order Ericales) exhibit greater floral morphological diversity towards the tropics (Chartier et al., 2021). However, Chartier et al. (2021) did not conduct any phylogenetic comparative analyses, and it remains unclear what evolutionary processes generated these apparent patterns. To our knowledge, this study is the first to explicitly compare *rates* of floral morphology evolution across tropical and temperate environments. Notably, while sages and heathers diverged from one another quite some time ago (molecular and fossil evidence roughly suggest the mid to early Cretaceous, some 90-120 million years ago; Zhang et al., 2020), they are both members of the larger Asterid clade. To determine whether these results reflect more general patterns, future studies should investigate whether rates of floral trait evolution and/or floral morphological diversity increase towards the tropics not only in other Asterid clades but also more broadly across the Angiosperm phylogeny (e.g., Rosids, Monocots, Magnoliids).

### 3.4.2 Relationship to previous methods and possible extensions

This is not the first approach to joint inference of state histories and continuous trait evolution processes from phylogenetic comparative data. Theoretically, sequential approximations converge to a truly joint modeling approach if the likelihood of a continuous trait evolution model is averaged over a sufficiently representative sample of state histories. Generating such a sample, however, is not trivial, as most sampling methods (e.g., simmapping; Nielsen, 2002; Bollback, 2006; Revell, 2013) generate state histories associated with extremely low likelihoods that barely contribute to the overall average (Boyko et al., 2023b). Nonetheless, previous researchers have developed both effective Bayesian approaches (Caetano et al., 2018; May and Moore, 2020; Quintero and Landis, 2020) and clever greedy algorithms (Boyko et al., 2023b) for achieving just this. While our new SCE framework generally offers greater computational efficiency and, in some respects, flexibility compared to previous approaches, these alternative methods still offer some important strengths worth considering.

While our approach could be extended to multivariate continuous traits quite easily via the use of multidimensional FFTs, the computational complexity of our pruning algorithm would unfortunately scale poorly. In particular, each additional trait increases the number of grid points to consider by a factor equal to the specified grid resolution—for example, two traits each coarsely discretized into just 256 grid points would together still require  $256 \times 256 = 65,536$  grid points. We believe sparse/adaptive grids (Brumm and Scheidegger, 2017) offer a possible workaround to this issue, but is far beyond the scope of the current paper. However, methods like *ratematrix* (Caetano et al., 2018) and *MuSSCRat* (May and Moore, 2020) are already capable of accommodating several or more continuous traits. Additionally, while our SCE framework can model a wide array of possible evolutionary dynamics via Lévy processes, it is not easily extendable to more “adaptive” models of trait evolution like Ornstein Uhlenbeck (OU; Butler and King, 2004) and Fokker-Planck-Komologrov processes (Boucher et al., 2018). Unlike Lévy processes, these processes do not admit a characteristic exponent representation and are thus not directly compatible with our current approach. Fortunately, the recently-developed *hOUwie* is capable of inferring

state-dependent OU processes via a clever stochastic algorithm for approximating the likelihood of the model rather efficiently despite being based on a sequential approximation.

In general, we believe our method offers the most straight-forward and computationally-efficient modeling approach in the case of univariate continuous traits evolving under more “drift-like” processes (e.g., Brownian motion, “pulsed” evolution; see Landis et al., 2013; Landis and Schraiber, 2017). More broadly, however, we believe our novel algorithm establishes a helpful mathematical link between popular discrete and continuous trait evolution models. Future research could build off this framework to develop additional SCE models tailored to a variety of applications—for example, joint inference of geographic history and its effect on trait evolution (e.g., Goldberg et al., 2011; Caetano et al., 2018) or even jointly modeling the influence of a continuous variable on continuous trait evolution (e.g., FitzJohn, 2010). Another interesting—though perhaps more challenging—avenue for development would be uniting continuous, discrete, and lineage diversification models under a single framework, which would be possible by combining our algorithm with those used for SSE models—particularly the quantitative SSE model (i.e., quasse; FitzJohn, 2010). One potentially useful application of such a framework would be modeling more dynamic interactions between speciation, extinction, and continuous trait evolution using insights from cladogenetic SSE models (e.g., Goldberg and Igić, 2012; see also Bokma, 2008).

### **3.4.3 Conclusion**

Macroevolutionary researchers still remain limited in their ability to rigorously detect and quantify variation in continuous trait evolution dynamics, despite the ubiquity of evolutionary heterogeneity across the tree of life. Our new SCE modeling framework allows for joint inference of discrete state histories and their influence of the evolution of a continuous trait. Such states could represent different habitats, reproductive strategies, or even entirely unobserved variables/hidden states used to model generic background and/or residual heterogeneity in continuous trait evolution dynamics. By using this method to fit and compare several candidate models, researchers can easily and robustly test a potentially wide variety of hypotheses regarding what factors are associated with shifts in the tempo or mode of continuous trait evolution. Furthermore, the mathematical basis

of our approach presents numerous opportunities for further elaboration, extension, and connection with other methods going forward. Ultimately, we believe our method fills an important gap among phylogenetic comparative methods and will benefit a broad array of both method developers and empirical macroevolutionary researchers alike.

## BIBLIOGRAPHY

- Bartoszek K., Pienaar J., Mostad P., Andersson S., and Hansen T.F. 2012. A phylogenetic comparative method for studying multivariate adaptation. *J Theor Biol* 314:204–215.
- Beaulieu J.M. and O’Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst Biol* 65:583–601.
- Beaulieu J.M., O’Meara B.C., and Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. *Syst Biol* 62:725–737.
- Bokma F. 2008. Detection of “punctuated equilibrium” by Bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. *Evolution* 62:2718–2726.
- Bollback J.P. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* 7:88.
- Boucher F.C. and Démery V. 2016. Inferring bounded evolution in phenotypic characters from phylogenetic comparative data. *Syst Biol* 65:651–661.
- Boucher F.C., Démery V., Conti E., Harmon L.J., and Uyeda J. 2018. A general model for estimating macroevolutionary landscapes. *Syst Biol* 67:304–319.
- Bowman J.C. and Roberts M. 2011. Efficient dealiased convolutions without padding. *SIAM J. Sci. Comput.* 33:386–406.
- Boyko J.D. and Beaulieu J.M. 2021. Generalized hidden Markov models for phylogenetic comparative datasets. *Methods Ecol Evol* 12:468–478.
- Boyko J.D. and Beaulieu J.M. 2023. Reducing the biases in false correlations between discrete characters. *Syst Biol* 72:476–488.
- Boyko J.D., Hagen E.R., Beaulieu J.M., and Vasconcelos T. 2023a. The evolutionary responses of life-history strategies to climatic variability in flowering plants. *New Phytol* 240:1587–1600.
- Boyko J.D., O’Meara B.C., and Beaulieu J.M. 2023b. A novel method for jointly modeling the evolution of discrete and continuous traits. *Evolution* 77:836–851.
- Brock Fenton M. and Simmons N.B. 2015. *Bats*. University of Chicago Press, Chicago, IL.
- Brumm J. and Scheidegger S. 2017. Using adaptive sparse grids to solve high-dimensional dynamic models. *Econometrica* 85:1575–1612.

- Burnham K.P. and Anderson D.R. 2002. Information and likelihood theory: A basis for model selection and inference. Pages 49–97 *in* Model Selection and Multimodel Inference: A Practical information-theoretic Approach. Springer New York, New York, NY.
- Butler M.A. and King A.A. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am Nat* 164:683–695.
- Caetano D.S. and Harmon L.J. 2017. ratematrix: An R package for studying evolutionary integration among several traits on phylogenetic trees. *Methods Ecol Evol* 8:1920–1927.
- Caetano D.S., O’Meara B.C., and Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution* 72:2308–2324.
- Chartier M., von Balthazar M., Sontag S., Löfstrand S., Palme T., Jabbour F., Sauquet H., and Schönenberger J. 2021. Global patterns and a latitudinal gradient of flower disparity: Perspectives from the angiosperm order Ericales. *New Phytol* 230:821–831.
- Cooper N. and Purvis A. 2009. What factors shape rates of phenotypic evolution? A comparative study of cranial morphology of four mammalian clades. *J Evol Biol* 22:1024–1035.
- Crepet W.L. and Niklas K.J. 2009. Darwin’s second ‘abominable mystery’: Why are there so many angiosperm species? *Am J Bot* 96:366–381.
- DasGupta A. 2011. Characteristic functions and applications. Pages 293–322 *in* Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics. Springer New York, New York, NY.
- Davies T.J., Barraclough T.G., Chase M.W., Soltis P.S., Soltis D.E., and Savolainen V. 2004. Darwin’s abominable mystery: Insights from a supertree of the angiosperms. *Proc Natl Acad Sci USA* 101:1904–1909.
- Diamond J. and Roy D. 2023. Patterns of functional diversity along latitudinal gradients of species richness in eleven fish families. *Glob Ecol Biogeogr* 32:450–465.
- Dobzhansky T. 1950. Evolution in the tropics. *Am Sci* 38:209–221.
- Donoghue M.J. and Sanderson M.J. 2015. Confluence, synnovation, and depauperons in plant diversification. *New Phytol* 207:260–274.
- Drury J.P., Clavel J., Tobias J.A., Rolland J., Sheard C., and Morlon H. 2021. Tempo and mode of morphological evolution are decoupled from latitude in birds. *PLoS Biol* 19:e3001270.
- Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L., and Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65:3578–

3589.

- Eddelbuettel D. and Francois R. 2011. Rcpp: Seamless R and C++ integration. *J Stat Softw* 40:1–18.
- Eddelbuettel D. and Sanderson C. 2014. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput Stat Data Anal* 71:1054–1063.
- Feeley K.J. and Stroud J.T. 2018. Where on Earth are the “tropics”? *Front Biogeogr* 10:e38649.
- Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst Zool* 22:240–249.
- Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat* 125:1–15.
- Felsenstein J. 2012. A comparative method for both discrete and continuous characters using the threshold model. *Am Nat* 179:145–156.
- Fick S.E. and Hijmans R.J. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315.
- FitzJohn R.G. 2010. Quantitative traits and diversification. *Syst Biol* 59:619–633.
- FitzJohn R.G. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods Ecology Evol* 3:1084–1092.
- FitzJohn R.G., Maddison W.P., and Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst Biol* 58:595–611.
- Freyman W.A. and Höhna S. 2018. Cladogenetic and anagenetic models of chromosome number evolution: A Bayesian model averaging approach. *Syst Biol* 67:195–215.
- Frigo M. and Johnson S.G. 2005. The design and implementation of FFTW3. *Proc IEEE Inst Electr Electron Eng* 93:216–231.
- Gingerich P.D. 2009. Rates of evolution. *Annu Rev Ecol Evol Syst* 40:657–675.
- Goldberg E.E. and Foo J. 2020. Memory in trait macroevolution. *Am Nat* 195:300–314.
- Goldberg E.E. and Igić B. 2012. Tempo and mode in plant breeding system evolution. *Evolution* 66:3701–3709.
- Goldberg E.E., Lancaster L.T., and Ree R.H. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Syst Biol* 60:451–465.

- Hansen T.F., Pienaar J., and Orzack S.H. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62:1965–1977.
- Harmon L.J., Pennell M.W., Francisco Henao-Diaz L., Rolland J., Siple B.N., and Uyeda J.C. 2021. Causes and consequences of apparent timescaling across all estimated evolutionary rates. *Annu Rev Ecol Evol Syst* 52:587–609.
- Hassler G., Tolkoff M.R., Allen W.L., Ho L.S.T., Lemey P., and Suchard M.A. 2022a. Inferring phenotypic trait evolution on large trees with many incomplete measurements. *J Am Stat Assoc* 117:678–692.
- Hassler G.W., Gallone B., Aristide L., Allen W.L., Tolkoff M.R., Holbrook A.J., Baele G., Lemey P., and Suchard M.A. 2022b. Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods Ecol Evol* 13:2181–2197.
- Helmstetter A.J., Zenil-Ferguson R., Sauquet H., Otto S.P., Méndez M., Vallejo-Marin M., Schönenberger J., Burgarella C., Anderson B., de Boer H., Glémin S., and Käfer J. 2023. Trait-dependent diversification in angiosperms: Patterns, models and data. *Ecol Lett* 26:640–657.
- Herrera-Alsina L., van Els P., and Etienne R.S. 2019. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. *Syst Biol* 68:317–328.
- Hibbins M.S., Breithaupt L.C., and Hahn M.W. 2023. Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. *Proc Natl Acad Sci USA* 120:e2220389120.
- Hibbins M.S. and Hahn M.W. 2021. The effects of introgression across thousands of quantitative traits revealed by gene expression in wild tomatoes. *PLoS Genet* 17:e1009892.
- Hijmans R.J., Barbosa M., Ghosh A., and Mandel A. 2023. geodata: download geographic data. R package version 0.5-9.
- Hillebrand H. 2004. On the generality of the latitudinal diversity gradient. *Am Nat* 163:192–211.
- Hipsley C.A., Miles D.B., and Müller J. 2014. Morphological disparity opposes latitudinal diversity gradient in lacertid lizards. *Biol Lett* 10:20140101.
- Hiscott G., Fox C., Parry M., and Bryant D. 2016. Efficient recycled algorithms for quantitative trait models on phylogenies. *Genome Biol. Evol.* 8:1338–1350.
- Jablonski D. 2017. Approaches to macroevolution: 2. Sorting of variation, some overarching issues, and general conclusions. *Evol Biol* 44:451–475.
- Johnson S.G. 2021. The NLOpt nonlinear-optimization package. Version 2.7.1.

- Kittle A.M., Fryxell J.M., Desy G.E., and Hamr J. 2008. The scale-dependent impact of wolf predation risk on resource selection by three sympatric ungulates. *Oecologia* 157:163–175.
- Kriebel R., Drew B.T., Drummond C.P., González-Gallegos J.G., Celep F., Mahdjoub M.M., Rose J.P., Xiang C.L., Hu G.X., Walker J.B., Lemmon E.M., Lemmon A.R., and Sytsma K.J. 2019. Tracking temporal shifts in area, biomes, and pollinators in the radiation of *Salvia* (sages) across continents: Leveraging anchored hybrid enrichment and targeted sequence data. *Am J Bot* 106:573–597.
- Landis M.J., Matzke N.J., Moore B.R., and Huelsenbeck J.P. 2013. Bayesian analysis of biogeography when the number of areas is large. *Syst Biol* 62:789–804.
- Landis M.J. and Schraiber J.G. 2017. Pulsed evolution shaped modern vertebrate body sizes. *Proc Natl Acad Sci USA* 114:13224–13229.
- Maddison W.P., Midford P.E., and Otto S.P. 2007. Estimating a binary character's effect on speciation and extinction. *Syst Biol* 56:701–710.
- Magnuson-Ford K. and Otto S.P. 2012. Linking the investigations of character evolution and species diversification. *Am Nat* 180:225–245.
- Martin B.S., Bradburd G.S., Harmon L.J., and Weber M.G. 2023. Modeling the evolution of rates of continuous trait evolution. *Syst Biol* 72:590–605.
- Martins E.P. and Hansen T.F. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat* 149:646–667.
- May M.R. and Moore B.R. 2020. A Bayesian approach for inferring the impact of a discrete character on rates of continuous-character evolution in the presence of background-rate variation. *Syst Biol* 69:530–544.
- Mendes F.K., Fuentes-González J.A., Schraiber J.G., and Hahn M.W. 2018. A multispecies coalescent model for quantitative traits. *Elife* 7.
- Moein F., Jamzad Z., Rahiminejad M., Landis J.B., Mirtadzadini M., Soltis D.E., and Soltis P.S. 2023. Towards a global perspective for *Salvia* L.: Phylogeny, diversification and floral evolution. *J Evol Biol* 36:589–604.
- Moler C. and Van Loan C. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev Soc Ind Appl Math* 45:3–49.
- Nakov T., Beaulieu J.M., and Alverson A.J. 2019. Diatoms diversify and turn over faster in freshwater than marine environments. *Evolution* 73:2497–2511.

- Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol* 51:729–739.
- Pagel M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc R Soc B* 255:37–45.
- Pagel M., O’Donovan C., and Meade A. 2022. General statistical model shows that macroevolutionary patterns and processes are consistent with Darwinian gradualism. *Nat Commun* 13:1113.
- Pinsky M.A. 2009. Fourier transforms on the line and space. Pages 89–167 *in* Introduction to Fourier Analysis and Wavelets (D. Cox, S. G. Krantz, R. Mazzeo, and M. Scharlemann, eds.) vol. 102 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.
- Quintero I. and Landis M.J. 2020. Interdependent phenotypic and biogeographic evolution driven by biotic interactions. *Syst Biol* 69:739–755.
- Rabosky D.L., Donnellan S.C., Grundler M., and Lovette I.J. 2014. Analysis and visualization of complex macroevolutionary dynamics: An example from Australian scincid lizards. *Syst Biol* 63:610–627.
- Rabosky D.L. and Goldberg E.E. 2017. FiSSE: A simple nonparametric test for the effects of a binary character on lineage diversification rates. *Evolution* 71:1432–1442.
- Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223.
- Revell L.J. 2013. A comment on the use of stochastic character maps to estimate evolutionary rate variation in a continuously valued trait. *Syst Biol* 62:339–345.
- Sanderson C. and Curtin R. 2016. Armadillo: A template-based C++ library for linear algebra. *J Open Source Softw* 1:26.
- Sanderson C. and Curtin R. 2019. Practical sparse matrices in C++ with hybrid storage and template-based expression optimisation. *Math Comput Appl* 24:70.
- Sato K.I. 2013. Characterization and existence of Lévy and additive processes. Pages 31–68 *in* Lévy Processes and Infinitely Divisible Distributions (B. Bollobás, W. Fulton, A. Katok, F. Kirwan, P. Sarnak, B. Simon, and B. Totaro, eds.) vol. 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, UK.
- Saupe E.E. 2023. Explanations for latitudinal diversity gradients must invoke rate variation. *Proc Natl Acad Sci USA* 120:e2306220120.
- Sauquet H. and Magallón S. 2018. Key questions and challenges in angiosperm macroevolution. *New Phytol* 219:1170–1187.

- Schemske D.W. 2001. Biotic interactions and speciation in the tropics. Pages 219–239 in *Speciation and Patterns of Diversity* (R. Butlin, J. Bridle, and D. Schluter, eds.). Cambridge University Press, Cambridge, UK.
- Schemske D.W., Mittelbach G.G., Cornell H.V., Sobel J.M., and Roy K. 2009. Is there a latitudinal gradient in the importance of biotic interactions? *Annu Rev Ecol Evol Syst* 40:245–269.
- Schumm M., Edie S.M., Collins K.S., Gómez-Bahamón V., Supriya K., White A.E., Price T.D., and Jablonski D. 2019. Common latitudinal gradients in functional richness and functional evenness across marine and terrestrial systems. *Proc R Soc B* 286:20190745.
- Simpson G.G. 1944. *Tempo and Mode in Evolution*. Columbia University Press, New York, NY.
- Stan Development Team . 2019. *Stan Modeling Language Users Guide and Reference Manual*. Version 2.21.0.
- Stevens R.D., Willig M.R., and Strauss R.E. 2006. Latitudinal gradients in the phenetic diversity of New World bat communities. *Oikos* 112:41–50.
- Stork N.E., McBroom J., Gely C., and Hamilton A.J. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci USA* 112:7519–7523.
- Thomas G.H. and Freckleton R.P. 2012. MOTMOT: Models of trait macroevolution on trees. *Methods Ecol Evol* 3:145–151.
- Thompson J.N. 2005. *The Geographic Mosaic of Coevolution*. University of Chicago Press, Chicago, IL.
- Tolkoff M.R., Alfaro M.E., Baele G., Lemey P., and Suchard M.A. 2018. Phylogenetic factor analysis. *Syst Biol* 67:384–399.
- Tribble C.M., May M.R., Jackson-Gain A., Zenil-Ferguson R., Specht C.D., and Rothfels C.J. 2023. Unearthing modes of climatic adaptation in underground storage organs across Liliales. *Syst Biol* 72:198–212.
- Uyeda J.C., Zenil-Ferguson R., and Pennell M.W. 2018. Rethinking phylogenetic comparative methods. *Syst Biol* 67:1091–1109.
- Vasconcelos T., O’Meara B.C., and Beaulieu J.M. 2022. A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. *Evolution* 76:1420–1433.
- Verboom G.A., Boucher F.C., Ackerly D.D., Wootton L.M., and Freyman W.A. 2020. Species selection regime and phylogenetic tree shape. *Syst Biol* 69:774–794.

- Wester P., Cairampoma L., Haag S., Schramme J., Neumeyer C., and Claßen-Bockhoff R. 2020. Bee exclusion in bird-pollinated *Salvia* flowers: The role of flower color versus flower construction. *Int J Plant Sci* 181:770–786.
- Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
- Ypma J., Johnson S.G., Stamm A., Borchers H.W., Eddelbuettel D., Ripley B., Hornik K., Chiquet J., Adler A., Dai X., and Ooms J. 2022. nlotpr: R Interface to NLOPT. R package version 2.0.3.
- Zhang C., Zhang T., Luebert F., Xiang Y., Huang C.H., Hu Y., Rees M., Frohlich M.W., Qi J., Weigend M., and Ma H. 2020. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol Biol Evol* 37:3188–3210.
- Zizka A., Silvestro D., Andermann T., Azevedo J., Duarte Ritter C., Edler D., Farooq H., Herdean A., Ariza M., Scharn R., Svanteson S., Wengstrom N., Zizka V., and Antonelli A. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol Evol* 10:744–751.

## APPENDIX 3A

### SUPPLEMENTAL TABLES AND FIGURES

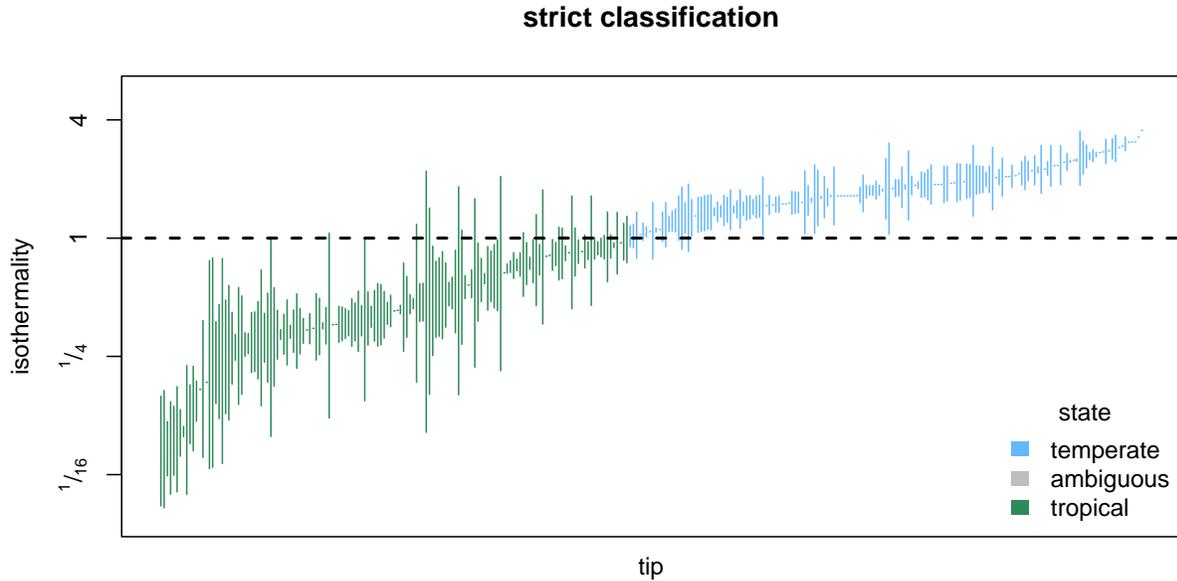


Figure 3A.1 The calculated isothermality (i.e., the ratio of seasonal variation in temperature to daily temperature fluctuations; see Feeley and Stroud, 2018) ranges used to classify each tip/taxon in the sage (*Lamiaceae: Salvia* L.) phylogeny for which we had geographic occurrence data as either tropical (vertical line segments colored darker green) or temperate (lines colored lighter blue) under our “strict” coding scheme (refer to subsection 3.2.4 for details). The lower and upper bounds of each vertical line segment represent the empirical 5 and 95% quantiles, respectively, of the isothermality values associated with occurrence records for a single taxon. The dashed horizontal line depicts the position where seasonal and daily temperature variation are equal—here, taxa with range midpoints below and above this line were considered tropical and temperate, respectively.

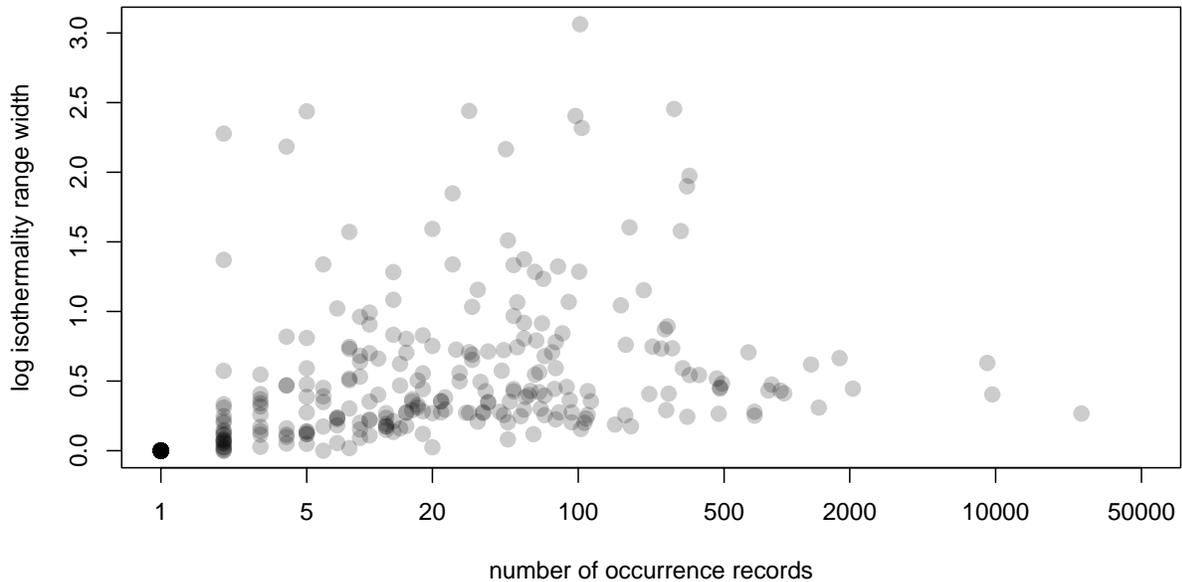


Figure 3A.2 The widths of calculated isothermality (i.e., the ratio of seasonal variation in temperature to daily temperature fluctuations; see Feeley and Stroud, 2018) ranges for each tip/taxon in the sage (Lamiaceae: *Salvia* L.) phylogeny for which we had geographic occurrence data, plotted against the number of occurrence records associated with each tip on the x-axis. Widths are given by the difference between the natural log-transformed empirical 5 and 95% quantiles of the isothermality values associated with occurrence records for a single taxon. Note that, while the range widths generally increase among taxa with more occurrence records (i.e., because the geographic distribution of such taxa are better-sampled), this trend largely “levels” out for taxa associated with more than 20 records. For our “conservative” coding of tropicality versus temperateness (refer to subsection 3.2.4 for details), range widths for taxa with 20 or fewer records were symmetrically expanded (on the log scale) to the mean width among taxa with more than 20 records, which came out to  $\sim 0.7$  (corresponding to 1 to 2-fold range of isothermality values).

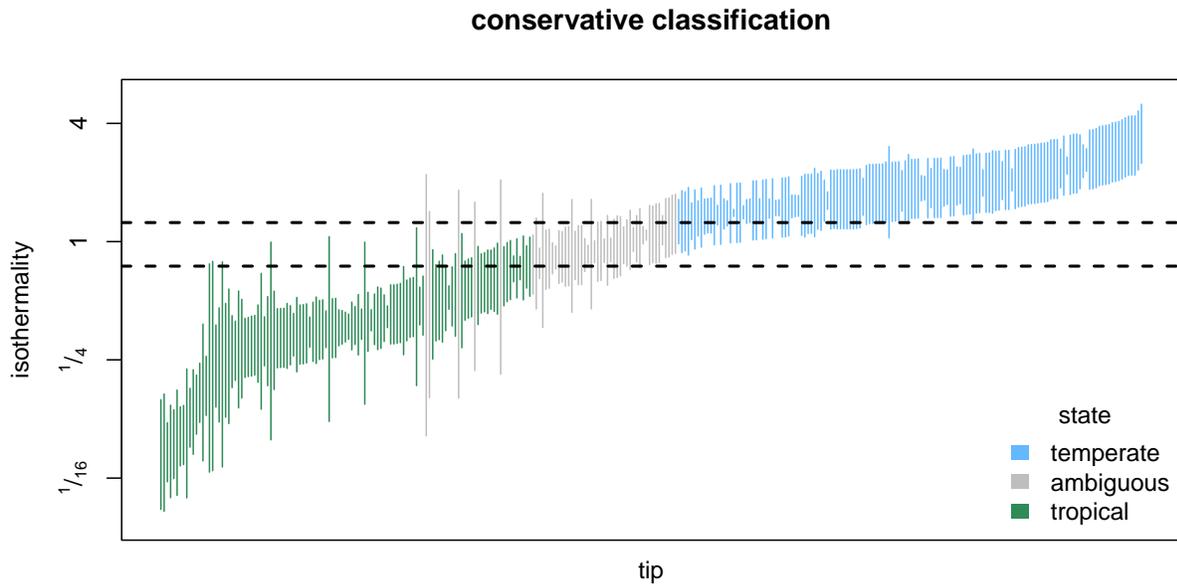


Figure 3A.3 The calculated isothermality (i.e., the ratio of seasonal variation in temperature to daily temperature fluctuations; see Feeley and Stroud, 2018) ranges used to classify each tip/taxon in the sage (Lamiaceae: *Salvia* L.) phylogeny for which we had geographic occurrence data as either tropical (vertical line segments colored darker green), temperate (lines colored lighter blue), or ambiguous (lines colored gray) under our “conservative” coding scheme (refer to subsection 3.2.4 for details). Generally, the lower and upper bounds of each vertical line segment represent the empirical 5 and 95% quantiles, respectively, of the isothermality values associated with occurrence records for a single taxon. However, the widths of isothermality ranges for any undersampled taxa (i.e., 20 or fewer occurrence records) were expanded to at least the mean range width of well-sampled taxa (i.e., more than 20 records; see Fig. 3A.2). The dashed horizontal lines depict the positions where annual temperature variation is 25% lower and higher than daily temperature variation. Here, taxa were only considered tropical if their isothermality range’s midpoint and upper bound both fell below the lower and higher dashed lines, respectively. Similarly, taxa were only considered temperate if their range’s midpoint and lower bound both exceeded the higher and lower dashed lines, respectively. All other taxa failed to exhibit a strong preference for either tropical or temperate environments and were thus considered ambiguous.

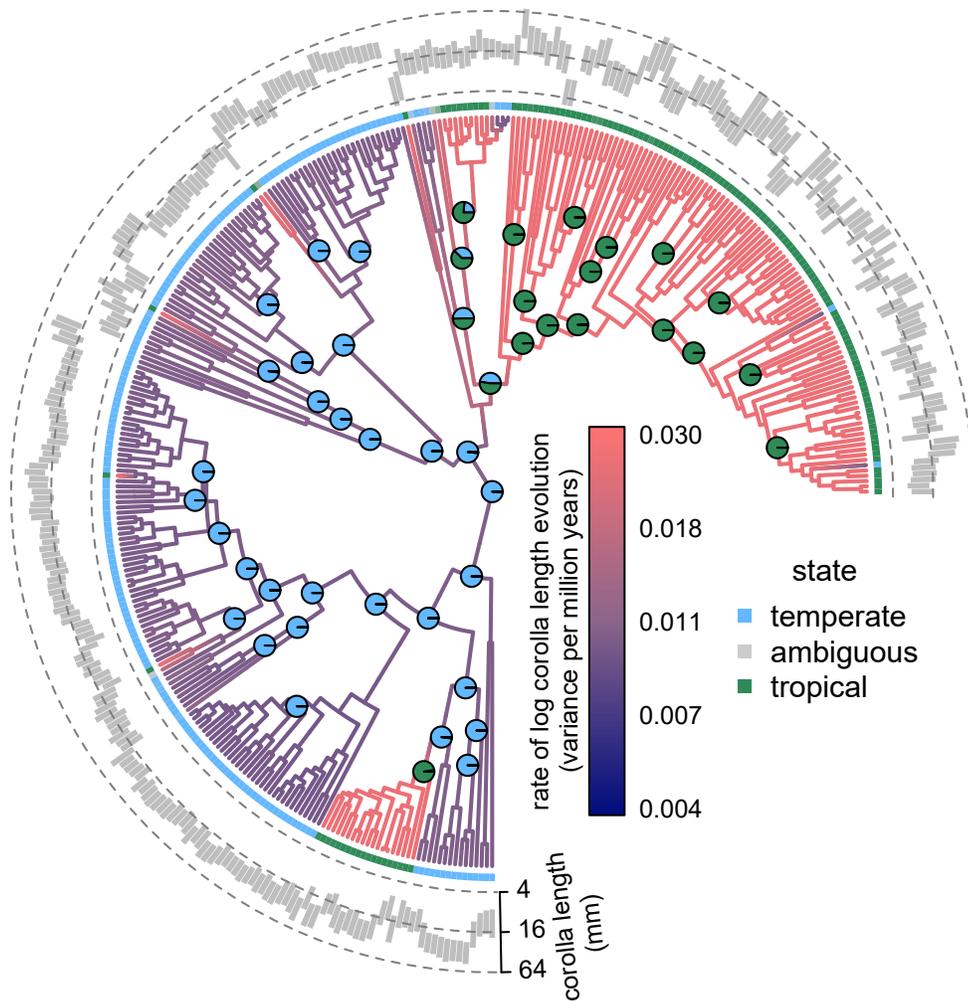


Figure 3A.4 Phylogram depicting model-averaged marginal ancestral rate and state estimates (based on sample size corrected Akaike Information Criterion—or AICc—weights) based on our joint analysis of corolla length evolution and temperate-tropical transitions (conservatively-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The color of branches correspond to inferred rates of corolla length evolution, with darker, cooler and lighter, warmer colors denoting relatively slow and fast rates, respectively. Pie charts at select nodes depict the probability that a given node tended to occur in either temperate (light blue) or tropical (dark green) environments. Because we lacked data for some tips and accounted for uncertainty in tropicality and corolla lengths, we also depict inferred tropicality probabilities (via colored boxes; light gray indicates even chances of being tropical or temperate) and 95% confidence intervals on corolla lengths (via gray bars) arrayed along the tips.

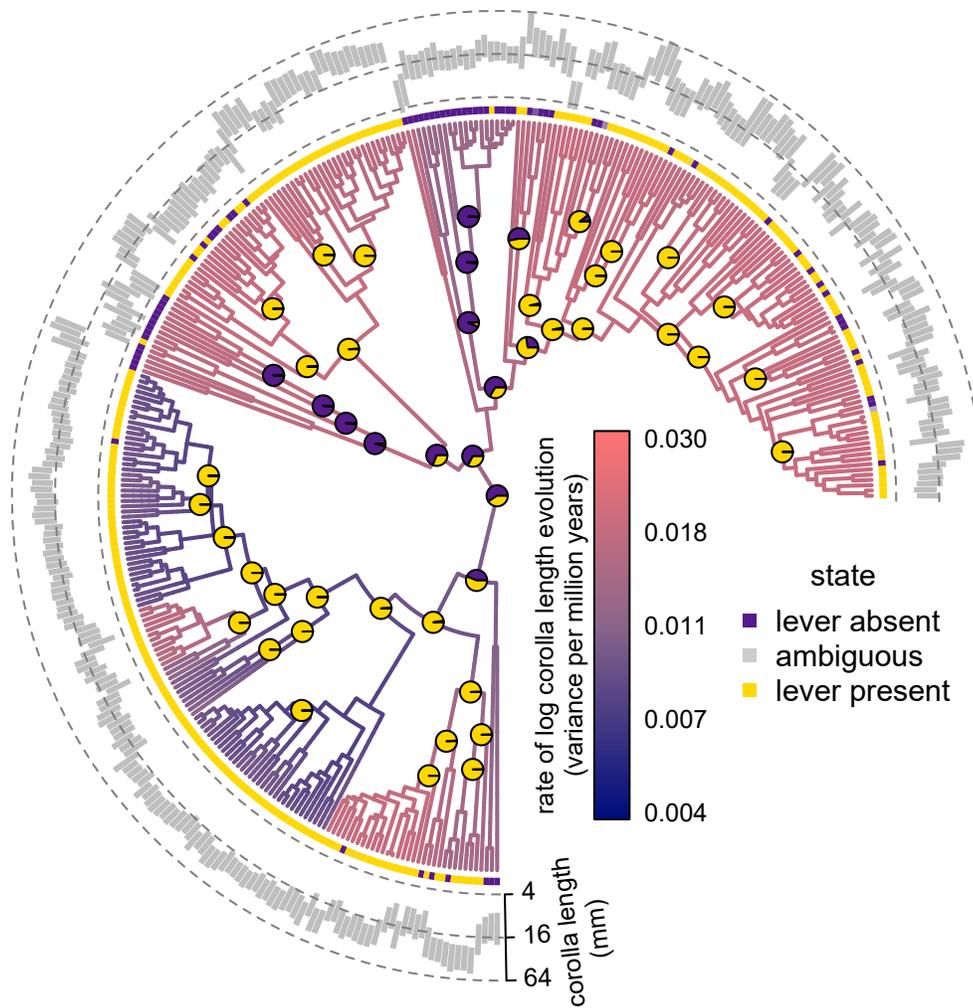


Figure 3A.5 Phylogram depicting model-averaged marginal ancestral rate and state estimates (based on sample size corrected Akaike Information Criterion—or AIC<sub>c</sub>—weights) based on our joint analysis of corolla length and staminal lever evolution among sages (Lamiaceae: *Salvia* L.). The color of branches correspond to inferred rates of corolla length evolution, with darker, cooler and lighter, warmer colors denoting relatively slow and fast rates, respectively. Pie charts at select nodes depict the probability that a given node either possessed (light yellow) or lacked (dark purple) staminal levers. Because we lacked data for some tips and accounted for uncertainty in tropicality and corolla lengths, we also depict inferred staminal lever presence probabilities (via colored boxes; light gray indicates even chances of having or lacking staminal levers) and 95% confidence intervals on corolla lengths (via gray bars) arrayed along the tips.

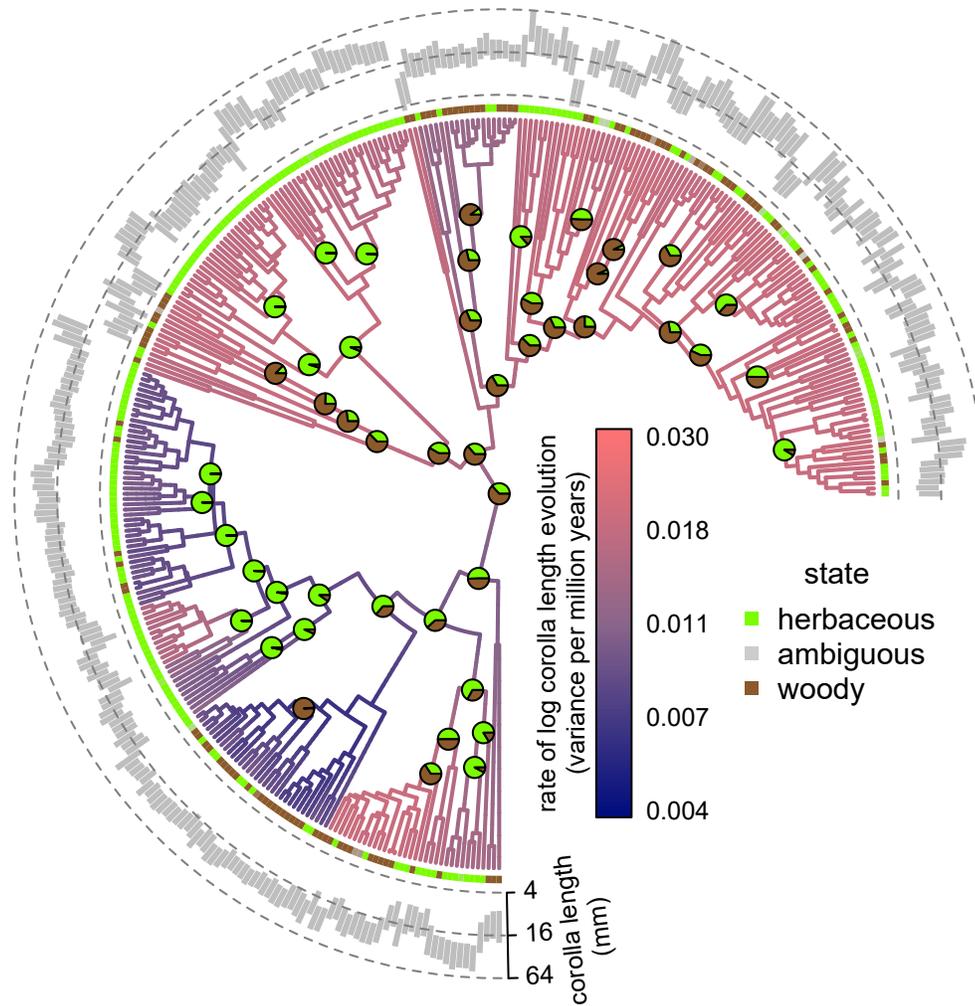


Figure 3A.6 Phylogram depicting model-averaged marginal ancestral rate and state estimates (based on sample size corrected Akaike Information Criterion—or AICc—weights) based on our joint analysis of corolla length and woodiness evolution among sages (Lamiaceae: *Salvia* L.). The color of branches correspond to inferred rates of corolla length evolution, with darker, cooler and lighter, warmer colors denoting relatively slow and fast rates, respectively. Pie charts at select nodes depict the probability that a given node was either woody (dark brown) or herbaceous (light green). Because we lacked data for some tips and accounted for uncertainty in woodiness and corolla lengths, we also depict inferred tropicality probabilities (via colored boxes; light gray indicates even chances of being woody or herbaceous) and 95% confidence intervals on corolla lengths (via gray bars) arrayed along the tips.

Table 3A.1 Sample size corrected Akaike Information Criterion (AICc) weights for joint models of corolla length and discrete trait evolution among sages (Lamiaceae: *Salvia* L.), demonstrating the support for alternative models for any given discrete trait. The rows correspond to alternative models that differed in four assumptions corresponding to the four leftmost columns: 1) whether rates of corolla length evolution vary independently of the observed discrete trait (i.e., whether the model included “hidden states”; labeled indep?), 2) whether rates vary according to the observed discrete trait (dep?), 3) whether transitions among states of the observed discrete trait occur at unequal, “asymmetric” rates (obs. asym?), and 4) whether transition rates among hidden states are asymmetric (hid. asym?). ✓ and × symbols indicate which models made and did not make these assumptions, respectively. The remaining four rightmost columns correspond to the different discrete traits we analyzed: 1-2) alternative “strict” and “conservative” codings (refer to subsection 3.2.4 for details) of tropicality versus temperateness (strict/conserv. trop.), 3) the presence/absence of staminal levers (lever pres.), and 4) woodiness versus herbaceousness (woodiness). Note that the weights only describe the support for a given model for a given discrete trait, and are not directly comparable across discrete traits.

model assumptions				discrete trait			
indep?	dep?	obs. asym?	hid. asym?	strict trop.	conserv. trop.	lever pres.	woodiness
×	×	×	—	1.9%	1.0%	16.8%	2.4%
×	✓	×	—	52.2%	59.4%	6.2%	0.9%
✓	×	×	×	2.4%	1.2%	21.7%	3.1%
✓	✓	×	×	4.5%	4.4%	4.1%	1.8%
×	×	✓	—	0.9%	0.4%	6.0%	18.4%
×	✓	✓	—	28.5%	27.1%	2.2%	6.6%
✓	×	✓	×	1.2%	0.5%	7.7%	23.4%
✓	✓	✓	×	2.3%	1.9%	1.4%	9.0%
✓	×	×	✓	2.3%	1.2%	20.6%	2.9%
✓	✓	×	✓	1.8%	1.8%	4.5%	1.2%
✓	×	✓	✓	1.1%	0.5%	7.3%	22.1%
✓	✓	✓	✓	0.9%	0.8%	1.6%	8.3%

Table 3A.2 Parameter estimates based on our joint analysis of corolla length evolution and temperate-tropical transitions (strictly-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The rows correspond to alternative models that differed in four assumptions corresponding to the four leftmost columns: 1) whether rates of corolla length evolution vary independently of tropicality/temperateness (i.e., whether the model included “hidden states”; labeled indep?), 2) whether rates vary according to tropicality/temperateness (dep?), 3) whether transitions into versus out of the tropics occur at unequal, “asymmetric” rates (obs. asym?), and 4) whether transition rates among hidden states are asymmetric (hid. asym?). ✓ and × symbols indicate which models made and did not make these assumptions, respectively. The remaining 9 rightmost columns correspond to different parameters. Parameters denoted  $q_{x,y}$  represent transition rates from state x to state y per million years. States 0 and 1 refer to temperate and tropical states, respectively, while states A and B refer to “slow” (i.e., lower rate of corolla length evolution) and “fast” hidden states, respectively. Parameters denoted  $\sigma_{xw}^2$  refer to rates of log-transformed corolla length evolution (i.e., increase in variance per million years) in observed state x and hidden state w. Lastly,  $\epsilon$  denotes “tip error”—the inferred standard deviation of (presumably normally-distributed) log-transformed corolla length measurements across all tips due to measurement error and/or within-tip phenotypic variation.

model assumptions				parameters								
indep?	dep?	obs. asym?	hid. asym?	$q_{0,1}$	$q_{1,0}$	$q_{A,B}$	$q_{B,A}$	$\sigma_{0A}^2$	$\sigma_{0B}^2$	$\sigma_{1A}^2$	$\sigma_{1B}^2$	$\epsilon$
×	×	×	—	0.017	0.017	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	×	—	0.017	0.017	—	—	0.010	0.010	0.026	0.026	0.290
✓	×	×	×	0.017	0.017	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	×	×	0.017	0.017	0.007	0.007	0.000	0.012	0.000	0.027	0.293
×	×	✓	—	0.019	0.014	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	✓	—	0.020	0.013	—	—	0.010	0.010	0.026	0.026	0.290
✓	×	✓	×	0.019	0.014	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	✓	×	0.020	0.014	0.007	0.007	0.000	0.012	0.000	0.028	0.293
✓	×	×	✓	0.017	0.017	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	×	✓	0.017	0.017	0.000	0.035	0.006	0.017	0.028	0.021	0.296
✓	×	✓	✓	0.019	0.014	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	✓	✓	0.020	0.013	0.000	0.035	0.006	0.017	0.028	0.022	0.296

Table 3A.3 Parameter estimates based on our joint analysis of corolla length evolution and temperate-tropical transitions (conservatively-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The rows correspond to alternative models that differed in four assumptions corresponding to the four leftmost columns: 1) whether rates of corolla length evolution vary independently of tropicality/temperateness (i.e., whether the model included “hidden states”; labeled indep?), 2) whether rates vary according to tropicality/temperateness (dep?), 3) whether transitions into versus out of the tropics occur at unequal, “asymmetric” rates (obs. asym?), and 4) whether transition rates among hidden states are asymmetric (hid. asym?). ✓ and × symbols indicate which models made and did not make these assumptions, respectively. The remaining 9 rightmost columns correspond to different parameters. Parameters denoted  $q_{x,y}$  represent transition rates from state x to state y per million years. States 0 and 1 refer to temperate and tropical states, respectively, while states A and B refer to “slow” (i.e., lower rate of corolla length evolution) and “fast” hidden states, respectively. Parameters denoted  $\sigma_{xw}^2$  refer to rates of log-transformed corolla length evolution (i.e., increase in variance per million years) in observed state x and hidden state w. Lastly,  $\epsilon$  denotes “tip error”—the inferred standard deviation of (presumably normally-distributed) log-transformed corolla length measurements across all tips due to measurement error and/or within-tip phenotypic variation.

model assumptions				parameters								
indep?	dep?	obs. asym?	hid. asym?	$q_{0,1}$	$q_{1,0}$	$q_{A,B}$	$q_{B,A}$	$\sigma_{0A}^2$	$\sigma_{0B}^2$	$\sigma_{1A}^2$	$\sigma_{1B}^2$	$\epsilon$
×	×	×	—	0.008	0.008	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	×	—	0.008	0.008	—	—	0.010	0.010	0.028	0.028	0.286
✓	×	×	×	0.008	0.008	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	×	×	0.008	0.008	0.007	0.007	0.000	0.011	0.000	0.029	0.290
×	×	✓	—	0.009	0.006	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	✓	—	0.009	0.006	—	—	0.010	0.010	0.028	0.028	0.286
✓	×	✓	×	0.009	0.006	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	✓	×	0.009	0.006	0.010	0.010	0.000	0.012	0.027	0.028	0.288
✓	×	×	✓	0.008	0.008	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	×	✓	0.008	0.008	0.000	0.036	0.006	0.017	0.030	0.023	0.293
✓	×	✓	✓	0.009	0.006	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	✓	✓	0.009	0.006	0.000	0.036	0.006	0.017	0.030	0.023	0.293

Table 3A.4 Parameter estimates based on our joint analysis of corolla length and staminal lever evolution (strictly-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The rows correspond to alternative models that differed in four assumptions corresponding to the four leftmost columns: 1) whether rates of corolla length evolution vary independently of lever presence/absence (i.e., whether the model included “hidden states”; labeled indep?), 2) whether rates vary according to lever presence/absence (dep?), 3) whether gains and losses of levers occur at unequal, “asymmetric” rates (obs. asym?), and 4) whether transition rates among hidden states are asymmetric (hid. asym?). ✓ and × symbols indicate which models made and did not make these assumptions, respectively. The remaining 9 rightmost columns correspond to different parameters. Parameters denoted  $q_{x,y}$  represent transition rates from state x to state y per million years. States 0 and 1 refer to the absence and presence of staminal levers, respectively, while states A and B refer to “slow” (i.e., lower rate of corolla length evolution) and “fast” hidden states, respectively. Parameters denoted  $\sigma_{xw}^2$  refer to rates of log-transformed corolla length evolution (i.e., increase in variance per million years) in observed state x and hidden state w. Lastly,  $\epsilon$  denotes “tip error”—the inferred standard deviation of (presumably normally-distributed) log-transformed corolla length measurements across all tips due to measurement error and/or within-tip phenotypic variation.

model assumptions				parameters								
indep?	dep?	obs. asym?	hid. asym?	$q_{0,1}$	$q_{1,0}$	$q_{A,B}$	$q_{B,A}$	$\sigma_{0A}^2$	$\sigma_{0B}^2$	$\sigma_{1A}^2$	$\sigma_{1B}^2$	$\epsilon$
×	×	×	—	0.024	0.024	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	×	—	0.024	0.024	—	—	0.013	0.013	0.015	0.015	0.315
✓	×	×	×	0.024	0.024	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	×	×	0.024	0.024	0.019	0.019	0.000	0.018	0.004	0.024	0.301
×	×	✓	—	0.024	0.024	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	✓	—	0.025	0.024	—	—	0.013	0.013	0.015	0.015	0.315
✓	×	✓	×	0.024	0.024	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	✓	×	0.025	0.024	0.019	0.019	0.000	0.018	0.004	0.024	0.301
✓	×	×	✓	0.024	0.024	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	×	✓	0.024	0.024	0.046	0.000	0.000	0.021	0.004	0.022	0.294
✓	×	✓	✓	0.024	0.024	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	✓	✓	0.026	0.023	0.046	0.000	0.000	0.021	0.004	0.022	0.294

Table 3A.5 Parameter estimates based on our joint analysis of corolla length and woodiness evolution (strictly-coded; refer to subsection 3.2.4 for details) among sages (Lamiaceae: *Salvia* L.). The rows correspond to alternative models that differed in four assumptions corresponding to the four leftmost columns: 1) whether rates of corolla length evolution vary independently of woodiness/herbaceousness (i.e., whether the model included “hidden states”; labeled indep?), 2) whether rates vary according to woodiness/herbaceousness (dep?), 3) whether transitions to woody and herbaceous habits occur at unequal, “asymmetric” rates (obs. asym?), and 4) whether transition rates among hidden states are asymmetric (hid. asym?). ✓ and × symbols indicate which models made and did not make these assumptions, respectively. The remaining 9 rightmost columns correspond to different parameters. Parameters denoted  $q_{x,y}$  represent transition rates from state x to state y per million years. States 0 and 1 refer to herbaceous and woody states, respectively, while states A and B refer to “slow” (i.e., lower rate of corolla length evolution) and “fast” hidden states, respectively. Parameters denoted  $\sigma_{xw}^2$  refer to rates of log-transformed corolla length evolution (i.e., increase in variance per million years) in observed state x and hidden state w. Lastly,  $\epsilon$  denotes “tip error”—the inferred standard deviation of (presumably normally-distributed) log-transformed corolla length measurements across all tips due to measurement error and/or within-tip phenotypic variation.

model assumptions				parameters								
indep?	dep?	obs. asym?	hid. asym?	$q_{0,1}$	$q_{1,0}$	$q_{A,B}$	$q_{B,A}$	$\sigma_{0A}^2$	$\sigma_{0B}^2$	$\sigma_{1A}^2$	$\sigma_{1B}^2$	$\epsilon$
×	×	×	—	0.051	0.051	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	×	—	0.051	0.051	—	—	0.014	0.014	0.013	0.013	0.316
✓	×	×	×	0.051	0.051	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	×	×	0.051	0.051	0.017	0.017	0.007	0.019	0.000	0.030	0.301
×	×	✓	—	0.038	0.080	—	—	0.014	0.014	0.014	0.014	0.316
×	✓	✓	—	0.038	0.079	—	—	0.014	0.014	0.014	0.014	0.316
✓	×	✓	×	0.038	0.080	0.021	0.021	0.004	0.023	0.004	0.023	0.300
✓	✓	✓	×	0.039	0.079	0.017	0.017	0.007	0.021	0.000	0.024	0.301
✓	×	×	✓	0.051	0.051	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	×	✓	0.051	0.051	0.037	0.000	0.007	0.020	0.000	0.029	0.293
✓	×	✓	✓	0.038	0.080	0.048	0.000	0.003	0.022	0.003	0.022	0.295
✓	✓	✓	✓	0.038	0.079	0.036	0.000	0.009	0.021	0.000	0.027	0.293

## APPENDIX 3B

### PRUNING ALGORITHM DETAILS

Here, we briefly outline some of the more technical details of implementing the pruning algorithm described in subsection 3.2.1, focusing on how we address three key issues. In particular, a naïve implementation of our pruning algorithm would: 1) be quite computationally expensive and slow because it requires exponentiating anywhere from several hundred to a few thousand matrices for each branch in a phylogeny. Additionally, due to the recursive nature of our algorithm, numerical errors in partial likelihood calculations may accumulate due to both 2) numerical artifacts associated with inverting Fourier transforms (e.g., “aliasing” and rounding errors) and 3) general under- and overflow (i.e., partial likelihoods becoming smaller and larger than what can be represented using floating point arithmetic). We describe how we effectively manage each of these problems below, starting with reducing the computational burden of matrix exponentiation.

Unfortunately, matrix exponentiation is infamous for being a slow and computationally difficult operation (Moler and Van Loan, 2003). Nonetheless, these challenges have at least driven the development of many different methods for computing matrix exponentials, each with their own strengths and weaknesses. We chose to use one of the more straight-forward approaches, diagonalization, whereby a matrix to be exponentiated,  $Q$ , is eigendecomposed into a matrix of its eigenvectors,  $V$ , and a vector of its eigenvalues,  $\lambda$ . Then (Moler and Van Loan, 2003):

$$\exp[Qt] = V \text{diag}(\exp[\lambda t]) V^{-1} \quad (1)$$

Where  $t$  is a scalar,  $\text{diag}(\exp[\lambda t])$  denotes a square matrix with the exponentiated eigenvalues multiplied by  $t$  along its diagonal, and  $V^{-1}$  is the inverse of the eigenvector matrix. Notably, several other phylogenetic comparative methods based on pruning algorithms use diagonalization to exponentiate matrices as well (Pagel, 1994; Boucher and Démary, 2016; Boucher et al., 2018), and for good reason—such pruning algorithms require repeated computation of  $\exp[Qt]$  for the same matrix  $Q$  but different values of  $t$ , which corresponds to a phylogeny’s branch lengths in this context. While calculating  $V$ ,  $\lambda$ , and  $V^{-1}$  is computationally expensive, they only depend on  $Q$  and

may be pre-computed once prior to carrying out the pruning algorithm. Then, calculating  $\exp [Qt]$  for each branch in the phylogeny only requires basic (i.e., non-matrix) exponentiation and matrix multiplication, which is generally far simpler and quicker than direct matrix exponentiation.

In the case of our pruning algorithm specifically, we store  $V$  and  $V^{-1}$  for each matrix making up the “rate array”  $R$  (see Eqs. 7 and 8) in identically-structured “eigenvector/inverse eigenvector arrays”. Similarly, the  $\lambda$  vectors for each matrix in  $R$  are stored as the columns of an “eigenvalue matrix”. This approach further improves the performance of our pruning algorithm by allowing us to vectorize the exponentiation, addition, and multiplication operations involved in calculating Eq. 1 across the hundreds to thousands of matrices making up  $R$ . One last benefit of the diagonalization approach is that, unlike some other available methods for computing matrix exponentials, it directly generalizes to matrices of complex numbers (i.e., numbers of the form  $a + bi$ , where  $i = \sqrt{-1}$ ), which is necessary for our purposes as characteristic functions are generally complex-valued.

Moving onto managing the numerical instabilities associated with inverting Fourier transforms, we sought to devise a procedure for “cleaning” apparent artifacts in branch-inflated partial likelihoods ( $\phi_d^*(x)$ ; see Eq. 1 and accompanying text) resulting from two key limitations of the FFT algorithm. First, because we evaluate characteristic functions on a grid of necessarily limited extent and resolution, inverse FFTs may yield branch-inflated partial likelihoods exhibiting rapid, spurious oscillations under certain conditions—a phenomenon known as “aliasing” (Bowman and Roberts, 2011). Generally speaking, this occurs when partial likelihoods vary too rapidly across grid points to be fully represented by their discretized characteristic function representations. Fortunately, given the fact that partial likelihoods across grid points tend to rapidly “smooth out” under most trait evolution models (e.g., FitzJohn, 2010), such artifacts occur rather infrequently. Second, because we evaluate branch-inflated partial likelihoods on relatively high-resolution grids meant to approximate continuous domains, many partial likelihood values (particularly towards the boundaries of the grid) will be closer to 0 than what can be represented using floating point arithmetic (i.e., partial likelihoods around machine epsilon or lower— $\sim 1e-16$  on a typical computer). Accordingly, rounding errors during Fourier transform inversion can generate effectively

random “noise” around these extremely low partial likelihood values, resulting in slightly negative or even complex-valued partial likelihoods.

Through trial and error, we developed a relatively simple cleaning procedure that effectively manages these problems while avoiding any costly computation. We refer to the input and output of this procedure as “raw” and “cleaned” branch-inflated partial likelihoods, respectively. The raw partial likelihoods are directly given by the inverse Fourier transform of their characteristic function representation ( $\hat{\phi}_d(\xi)$ ; see text following Eq. 5). To clean the raw partial likelihoods, we first tackle some of the rounding errors by discarding the imaginary components of any complex-valued partial likelihoods (i.e., number of the form  $a + bi$  are “rounded” to  $a$ ). Then, to eliminate any rapid, likely erroneous oscillations reflecting aliasing errors, the raw partial likelihoods are “smoothed out” by rounding partial likelihood values surrounded by negative values on both sides to  $1e-16$ . To ensure all partial likelihoods are positive, any remaining values below  $1e-16$  are also rounded to  $1e-16$ . Lastly, to yield the final output, the partial likelihoods are rescaled such that they add up to the sum of the original, raw partial likelihoods—which is actually given by the first value of the characteristic function representation of the raw partial likelihoods and calculated quite accurately under our approach (DasGupta, 2011). To be clear, rounding all sufficiently low partial likelihood values up to  $1e-16$  is an approximation, as it causes the branch-inflated partial likelihoods to have “fatter tails” than they really should. Nonetheless, this procedure improves the numerical accuracy of subsequent Fourier transforms while also ensuring products of branch-inflated partial likelihoods always have a positive-valued sum.

Lastly, as the pruning algorithm proceeds and branch-inflated partial likelihoods are repeatedly multiplied together, partial likelihoods tend to rapidly approach values either too low or high to be represented using floating point arithmetic (i.e., under- and overflow, respectively). To prevent this, after multiplying cleaned branch-inflated partial likelihoods together for each state to yield the full partial likelihood matrix for a new edge  $e$  ( $\Phi_e$ ; see paragraph following Eqs. 1-4), the resulting matrix is rescaled to have a maximum of 1 (notably, this also ensures all branch-inflated partial likelihoods exhibit the same overall scale, rendering it more appropriate to use  $1e-16$  as a universal

“lower bound” when cleaning branch-inflated partial likelihoods). The cumulative product of the rescaling factors is in turn tracked on a logarithmic scale, and the final partial likelihoods at the root are subsequently log-transformed and divided by the final rescaling factor on the log scale. Additionally, we use the log-sum-exp trick (Stan Development Team, 2019) to stably sum partial likelihoods at the root including extremely low or high values according to Eqs. 1-4. Intriguingly—though perhaps somewhat unsurprisingly—the rescaling factor, rather than partial likelihoods at the root themselves, seem to determine most of the variation in the overall likelihoods associated with different parameter estimates. This reflects the fact that the rescaling factor in some sense measures how well the peaks of different branch-inflated partial likelihood matrices “match up”.