Reliable and interpretable inference of evolutionary history using Bayesian phylogenetic approaches

Andrew F Magee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Vladimir Minin, Chair

Frederick Matsen IV, Chair

Joseph Felsenstein

Program Authorized to Offer Degree:
Biology

University of Washington

Abstract

Reliable and interpretable inference of evolutionary history using Bayesian phylogenetic approaches

Andrew F Magee

Co-Chairs of the Supervisory Committee:
Professor Vladimir Minin
Biology

Professor Frederick Matsen IV
Statistics

Phylogenetic trees are key objects for understanding evolutionary history, first used to describe relationships between groups of species. Phylogenies help us to fill out the tree of life and to describe the dynamics that have given rise to the diversity of life on Earth. As we have not witnessed the entire history of any group, phylogenies must be inferred from character data (often DNA sequence data) using statistical models. If we can specifically infer trees with a time component, such that we can measure the lengths of branches in real time, we can attempt to make inferences about the processes that gave rise to the phylogeny itself. In the case of species histories (a macroevolutionary process), we use birth-death models. Birth-death models, and time-calibrated phylogenies in general, are also useful in describing the course of infectious disease outbreaks, an application area known as infectious disease phylodynamics. In this thesis, I (and co-authors) develop new birth-death models applicable to both macroevolutionary and phylodynamic applications. First, we describe a parameter-rich time-varying birth-death model, which allows for birth, death, sampling, and death-upon-sampling. In macroevolutionary applications, birth is speciation, death is extinction, and sampling is fossilization (plus later recovery of the fossil). Death-upon-sampling is primarily useful in phylodynamic applications, where it models treatment or isolation after

a diagnosis, and where birth is infection, death is recovery (absent treatment), and sampling is sequencing of the infectious disease agent (such as a virus). Our model includes all these processes for individual lineages, plus the possibility that there are instantaneous events applicable to all lineages. It is the first model to include these all-lineage-event versions of all four processes. Using Bayesian inference, we demonstrate the usefulness of this model in application to a previously inferred phylogeny of Crocodylomorpha (crocodiles and their relatives). We investigate the impact of the K-Pg (end Cretaceous) mass extinction and find that there is a very strong, and very robust, imprint of the K-Pg mass extinction in the phylogeny of Crocodylomorpha. Next, we describe time-varying priors applicable to rates of birth, death, and sampling through time. Specifically, we investigate performance of the horseshoe Markov random field as a birth-death model prior, and contrast its performance with a Gaussian Markov random field. In simulations, the horseshoe model performs quite well and appears to be capable of balancing both the power to detect rate variation with the ability to distinguish true rate variation from noise in the birth-death process. In full Bayesian analyses of real datasets (inferring the tree and birth-death model from sequence data), we detect a clear signature of a speciation-rate decrease in a group of Australian geckos and estimate that the HIV epidemic among Russian and Ukrainian drug users peaked between roughly 1993 and 2000. Lastly, we turn our attention back to the matter of inferring phylogenies. As phylogenetic posterior distributions are difficult to work with, we must instead approximate them using samples from Markov chain Monte Carlo. In this chapter, we ask if it is possible to quantify the variability (also called Monte Carlo error) inherent in this procedure. Using a novel simulation approach, we find that the Monte Carlo error in important quantities (such as the summary tree) can in fact be reliably quantified. Application to benchmark datasets shows the danger inherent in the currently common approaches of either ignoring the sampling variability in the tree or using proxies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Number Page

GLOSSARY

BDP: Birth-death process

ESS: Effective sample size

GMRF: Gaussian Markov random field

HSMRF: Horseshoe Markov random field

Markov: Built on the Markov property (the future is independent of the past given the present)

MCMC: Markov chain Monte Carlo

phylogeny: A description of evolutionary history consisting of a topology (a set of nested relationships) and branch lengths describing evolutionary distances or time.

topology: The discrete component of a phylogeny

NOTATION

τ: A phylogenetic tree topology

T: A phylogenetic time tree including a topology and branch lengths in real time

$\lambda(t)$: A time-varying birth rate

$\mu(t)$: A time-varying death rate

$\phi(t)$: A time-varying sampling rate

$r(t)$: A time-varying rate of becoming non-infectious after being sampled

$\boldsymbol{\lambda}$: A vector of birth rates

$\boldsymbol{\mu}$: A vector of death rates

$\boldsymbol{\phi}$: A vector of sampling rates

$\boldsymbol{r}$: A vector of rates of becoming non-infectious after being sampled

$\boldsymbol{\Lambda}$: A vector of birth burst probabilities

$\boldsymbol{M}$: A vector of death burst (mass extinction) probabilities

$\boldsymbol{\Phi}$: A vector of mass sampling probabilities

$\boldsymbol{R}$: A vector of probabilities of becoming non-infectious after being sampled at a mass sampling

ACKNOWLEDGMENTS

A Ugandan proverb states, "a child does not grow up only in a single home." Carl Sagan once said that if you wish to make a pie from scratch, you must first invent the universe. And F. Scott Fitzgerald opens *The Great Gatsby* with the exhortation, "remember that all the people in this world haven't had the advantages that you've had." Any, and all, of the same, can be said of a PhD. I am grateful to my parents, Becky and Larry Magee, who fostered in me an appreciation for learning. And who put up with me along the way. I am thankful for the many wonderful teachers in elementary, middle, and high school that fostered that appreciation and gave me an incredible framework for learning. To name but a few, Nicole Della Santina was formative in my becoming a biologist, Rita Korsunsky showed me calculus isn't that bad (and can even be fun), and Fritz Torp was a source of support when I needed it most. And I am indebted too to the many professors in college who helped me to find my way in a path that led me to this PhD. I want to especially thank Bob Kimsey, who showed me that I can successfully do research (and have fun doing so), and to Brian Moore, who gave me an incredibly strong foundation in research and phylogenetic modeling. And of course, I am thankful for the many, many people who have helped me during my PhD itself.

My committee has been incredibly supportive. They have given me room to explore, and have always been accepting of the fact that plans change. Carl Bergstrom and Trevor Bedford both have made time out of their incredibly busy pandemic year(s) for me. Adam Leaché has been a great source of support and encouragement. Joe

Felsenstein has been an incredible trove of wisdom and advice, and I thank him too for taking the time to read this (rather long) thesis.

And, of course, I am immensely grateful to my co-advisors Vladimir Minin and Erick Matsen. Their guidance, patience, and support has been absolutely key to my success. Their willingness to indulge some of my stranger ideas and research rabbit holes helped me develop my own unique interests, but they never let me get so far out I felt truly lost.

I am grateful to the labs of my advisors, which are full of clever people happy to share their knowledge. A good chunk of this thesis would never have happened without Jim Faulkner, who pioneered the horseshoe Markov random field that I port to birth-death models in Chapter 3 and from whom I learned extensively about those models. Mike Karcher has been a great labmate twice over, and helped me get Chapter 4 off the ground before either of us knew it was going anywhere. Will DeWitt has perhaps made me question myself more than anyone else in my graduate career, but I firmly believe I, and my work, are better for this. Will, and our co-conspirator Sarah Hilton, deserve a special shout-out. Together we conceived of, designed, and carried out a research project on our own, with little outside guidance, and far from our own research areas. They are both great co-authors. I am grateful as well to Amrit Dhar, Jean Feng, Cheng Zhang, Chris Whidden, and Arman Bilge.

I have been lucky enough to have a great community here in the systematics wing of the Biology department. These fine biologists have both given me hope for our community and inspiration for why I work on these statistical phylogenetic models in the first place. I am grateful especially to the Leaché lab, and particularly to Leonard Jones and Itzue Caviedes-Solis, who are some of the wisest people I have met. I also wish to express gratitude to Laura Frost, Sima Bouzid, Rebecca Harris, Jared Grummer, Audrey Ragsac, Ana Maria Bedoya, Cooper French, and Dave Slager.

Further abroad, I want to thank Sebastian Höhna and his lab, who all welcomed me when I went to work with them for a month. Sebastian has been an invaluable source of advice on matters career and technical, and a very helpful co-author. Chapters 2 and 3 would not exist without his help.

Finally, I am thankful to those who helped keep me sane along the way. This includes many of the above friends and colleagues, and many others. I wish in particular to thank my cohort for all their support and commiseration. And everyone who joined me for boardgames. Elizabeth Perkins has been incredibly supportive throughout the course of my degree, and also gets most of the credit for us having a dog, Verna, who has herself been instrumental to my mental health.

Thank you all, and to the many people whose names I have inevitably forgotten to include on this list.

DEDICATION

To my family—by blood or by friendship, human or otherwise.

And to absent friends.

Chapter 1
INTRODUCTION

Phylogenetic trees are statistical models of evolution, for which inference is a computationally intensive endeavor, and as such phylogenetics is an interdisciplinary field. In this thesis, I will assume that readers have some basic background in statistics, evolutionary biology, and epidemiology. I will endeavor to explain myself plainly such that readers from a wide variety of backgrounds can understand the content, but I do hope that readers will be forgiving if occasionally I slip up. I also note that there are letters that will get recycled in meaning between these chapters. There are only so many letters in the English and Greek alphabets combined, and sometimes it is easier on the comprehension to redefine n than to seek new and as-yet unused symbols to represent the number of some key quantity.

1.1 Phylogenetics, the short, short version

Phylogenetic models are many and varied, but at the core of all of them is a phylogenetic tree. A phylogenetic tree is a model of the evolutionary history of a set of samples for which we have character data available. Admittedly, this description is a bit too general to be particularly helpful, so let us now narrow the scope to what matters for the kinds of phylogenetics in this thesis. I will be vague about what these samples are, and refer to them as *taxa* (or samples/sequences), rather than species, infectious disease agents, or infections. In this thesis, we will assume that the data we have is a sequence of discrete values, most commonly a sequence of nucleotides from the DNA of our study taxa. (Coincidentally, those are the two areas we will be applying trees to in this thesis.) This data must be arranged in a *multiple sequence alignment*, a matrix where the taxa are in rows, and every character in

a column is evolutionarily homologous. This is to say, every column (or site) represents the state that each taxon has for a single evolving character. Homology is important because it is the aggregate evolutionary history of these characters that we are describing with the phylogeny. For the purposes of this thesis, a phylogeny is assumed to be *completely resolved*, such that we do not have any nodes in the tree where more than 2 lineages simultaneously descend from any other. This tree can be a timetree, in which case the branch lengths are in units of time (days, months, millions of years, etc.) [81], as in the first two chapters where we deal in birth-death processes. In other words, time-calibrated phylogenies tell us who is most closely related to whom, and it may tell us when the common ancestor of any group of taxa lived. There are multiple kinds of trees that are not time trees, but in this thesis the only alternative we care about is an *unrooted* tree, where branch lengths are measured in evolutionary distances (expected numbers of substitutions per site in the alignment) [81]. And, lastly, I wish to note that in this thesis, all phylogenetics that will be discussed is Bayesian phylogenetics.

Broadly, there are two reasons that phylogenetic trees crop up in research. First, there are applications where the phylogenetic tree is itself an object of inference. This is where phylogenetic models began, as ways to assess the relationships between the many species of life on Earth that scientists study. I will call this use-case the case where trees are parameters of interest (to highlight the fact that they are inferred parameters in statistical models). Secondly, there are applications where the tree is in some sense a nuisance parameter. This includes cases where the object of inference is the process that gives rise to the tree, the spread of the lineages in the tree through space and time, or the evolution of specific traits. I will refer to this use-case as the case where trees are data, as we are using information from the tree to make inferences about the evolutionary processes that gave rise to it. In the literature, there has historically been a lot of focus on "inferring the tree of X" and then using that tree to answer these questions. I would argue that in the present day, this approach is largely misguided, given that neither does it allow researchers to account for uncertainty in the phylogeny, nor does it allow for the model of interest to shape the inferred

phylogeny. Except in cases where phylogenetic inference is particularly burdensome, as time progresses I hope to see the field moving as much as possible towards joint inference of the phylogeny and the actual question at hand. I must confess at this point that Chapter 2 rests upon previously inferred phylogenies, though I (and my co-author) have still endeavored to account for some uncertainty in the tree. In the rest of this introduction, I will attempt to sketch out a brief introduction to phylogenetic models, with increasing complexity, to help frame the rest of the thesis.

1.2 Trees as parameters of interest

1.2.1 An unrooted species tree

A researcher heads out into the field and collects a number of specimens from the family of frogs she studies. She sequences a number of genes from tissues in s species of frogs, and wants to infer the family tree of these frogs without worrying about when the species diverged, so she chooses to infer an unrooted phylogeny. Based on prior research and morphology, she adds one frog from a different family, so that she can root the tree after inference is done. The object of interest is the posterior distribution on the tree topology, τ, given the sequence data, $\boldsymbol{D}$, $\Pr(\tau \mid \boldsymbol{D})$. Note that here the herpetologist is assuming that there is a single tree topology that describes the data. However, as she knows, in reality, there is not a single topology for more than one site in the genome. Evolution is messy, and processes like recombination act to decouple evolutionary history across the genome. Yet all phylogenetic models currently in use assume *some* sharing of a topology. It is around this point when the aphorism, "all models are wrong, some are useful" usually makes an appearance. In other words, we know that it is wrong to assume a shared topology, however, this can be an incredibly useful abstraction to make inferences about the past. And in some cases, it even appears to be a surprisingly decent approximation to reality.

Let us return to our intrepid herpetologist, who is now confronted with the fact that a topology τ needs *branch lengths* to define evolutionary distances. She knows that different

genes evolve at different rates, and that these rates don't have to be consistent through time or across genes, so she chooses to assume that each gene in the dataset gets its own vector of branch lengths $\boldsymbol{v}$. As an example of this phenomenon, consider a phylogeny of the great apes, and specifically the branch of the tree leading to humans. Genes involved in our ability to speak have evolved quite rapidly compared to the genomic average. When measuring a tree's branches in substitutions, we would expect that the branch lengths for those genes would be longer than genes for, say, the construction of ribosomes or the cell membrane. Giving every gene its own branch lengths allows us to capture this sort of variation, in order to not let a few extreme genes wreak havoc on our inference procedure.

Our herpetologist also needs to describe how evolution proceeds along the branches of the tree in terms of a continuous time Markov chain (CTMC) model. Such a model characterizes how long any site in the alignment remains in a state, and how probable the various changes are. For DNA, we know that certain kinds of changes generally happen faster than others (purine changes to purine more often than pyramidine), and factors like this are ignored at one's peril. Knowing that sequence evolution is more complex than our models generally capture, she chooses the most parameter-rich model commonly used, a general time reversible (GTR) model [145]. To account for different rates of evolution across a gene, such as mutational hotspots or the fact that third codon positions evolve more quickly than first codon positions, she adds gamma-distributed among-site rate variation [165] (GTR+G). As different parts of the genome have their own mutation and substitution dynamics, she gives every gene its own GTR+G model.

We can now write out the posterior distribution that this researcher wants to know about. It is,

$$\Pr(\tau \mid \boldsymbol{D}) \propto \left(\prod_{i=1}^{G} \int_{\boldsymbol{\theta}_i, \boldsymbol{v}_i} \Big(\Pr(\boldsymbol{D}_i \mid \tau, \boldsymbol{v}_i, \boldsymbol{\theta}_i) \Pr(\boldsymbol{\theta}_i) \Pr(\boldsymbol{v}_i) \Big) \right) \Pr(\tau),$$

where there are G genes, $\boldsymbol{D}_i$ is the alignment for the ith gene (with $\boldsymbol{D}$ the alignment of all genes combined), $\boldsymbol{v}_i$ are the branch lengths for the ith gene, $\boldsymbol{\theta}_i$ are the substitution model parameter for the ith gene (here the GTR+G parameters), and the likelihood $\Pr(\boldsymbol{D}_i \mid \tau, \boldsymbol{v}_i, \boldsymbol{\theta}_i)$

of the ith gene can be efficiently computed using the Felsenstein pruning algorithm [45]. We make a number of assumptions of (conditional) independence here, mainly independence between genes given the tree, and our choice of a GTR+G model means we assume every site in a gene is independent given the model parameters. These assumptions may not be realistic, but they enable phylogenetics to work in a universe with a finite supply of time.

Back to our intrepid researcher. With her modeling decisions in hand, she can easily set up her model in programs like MrBayes [67] or RevBayes (in which the work in Chapters 2 and 3 is implemented) [63], which run Markov chain Monte Carlo (MCMC) sampling to approximate the posterior distribution (understanding the properties of this procedure is the focus of Chapter 4). Since the integration over branch lengths and other model parameters is difficult, these programs will actually sample from

$$\Pr(\tau, \boldsymbol{v}, \boldsymbol{\theta} \mid \boldsymbol{D}) \propto \left(\prod_{i=1}^{G} \Pr(\boldsymbol{D}_i \mid \tau, \boldsymbol{v}_i, \boldsymbol{\theta}_i) \Pr(\boldsymbol{\theta}_i) \Pr(\boldsymbol{v}_i) \right) \Pr(\tau),$$

and integrating out the other parameters is accomplished by simply ignoring them in the MCMC samples. Then our researcher is confronted with what is essentially a pile of samples of the tree topology, $\boldsymbol{\tau}$. She can summarize this set of trees with a single estimate, $\hat{\tau}$, common choices would be the majority rule consensus tree or the maximum *a posteriori* (MAP) tree [81]. She can also compute the posterior probabilities of certain splits (bipartitions of taxa that correspond to edges in an unrooted tree) in order to understand how confident she is that different groupings in the tree exist. With a summary tree and measures of certainty in hand, she is now free to investigate and revise the taxonomy of the family as needed. In Chapter 4, we will revisit this inference framework in an attempt to quantify the error inherent in using samples of trees to make statements about the posterior distribution of trees.

1.2.2 A time-calibrated species tree

For a later research project, our intrepid herpetologist now has questions about when certain clades of species originated in the past. (A *clade* is a group consisting of all descendants of a particular lineage in the tree and only those lineages). This topic, called *divergence time estimation*, has had much ink spilled on it elsewhere [81, is a useful starting point], so I will aim only to give a brief explanation. Conceptually, the biggest difference is that the object of interest is no longer a tree topology, τ, but a rooted, time-calibrated phylogeny T. This is a phylogeny where branch lengths correspond to real time, and the ages of nodes in the tree tell us about the ages of clades of species we care about.

One approach to estimating timetrees is to essentially co-opt the above model for unrooted trees. In this case, we assume that there is a shared timetree for all loci, but that each gene has a vector of *branch rates*, which specify how quickly evolution proceeds along each branch (the branch length in substitutions is the branch length in time multiplied by the branch rate). The posterior distribution on all model parameters is,

$$\Pr(T, \boldsymbol{r}, \boldsymbol{\theta} \mid \boldsymbol{D}, \boldsymbol{f}) \propto \left(\prod_{i=1}^{G} \Pr(\boldsymbol{D}_i \mid T, \boldsymbol{r}_i, \boldsymbol{\theta}_i) \Pr(\boldsymbol{\theta}_i) \Pr(\boldsymbol{r}_i) \right) \Pr(T),$$

and again we can obtain samples from $\Pr(T \mid \boldsymbol{D})$ by marginalizing out (ignoring) the other model parameters. Some information, $\boldsymbol{f}$, is required in order to disentangle the evolutionary rates from the times over which evolution occurred, this is usually fossil information about the age of certain groups of species. There are many important biological reasons that this is not always an appropriate model, but it is among the simplest (and thus relatively easy to understand) and it sets up the models we will use in later chapters. Models like this can easily be run in programs like BEAST [143] or RevBayes, and a graphical representation can be found in Figure 1.1.

One important feature of these models for time trees is the *tree prior*, $\Pr(T)$. Unlike in the unrooted case, these distributions can be parameterized in ways that are biologically

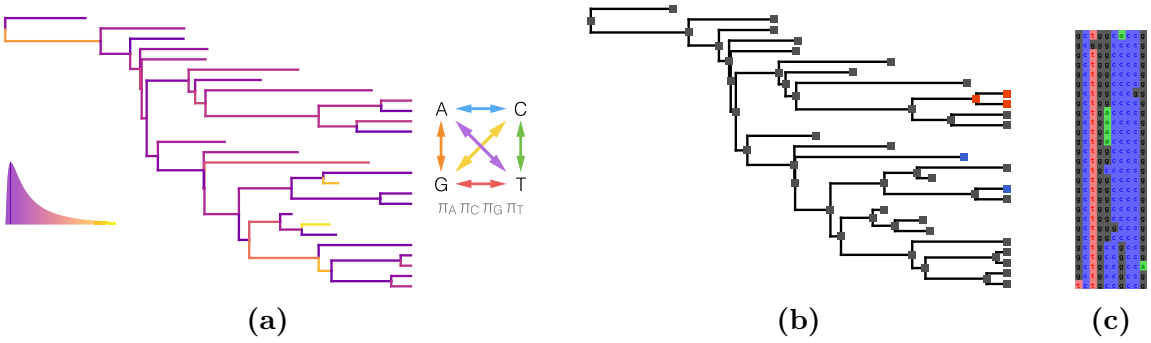


Figure 1.1: A graphical representation of time-calibrated Bayesian phylogenetic models. In panel (a), we see the tree, with each branch colored by the rate (drawn from the distribution shown to the left) at which evolution proceeds on it and a cartoon of a nucleotide substitution model which governs how changes between different states occur. Using these rates, character histories such as that in panel (b) may be sampled. The states at the tips of the tree (b) are collected as a column in a multiple sequence alignment (c).

meaningful. Our intrepid herpetologist, for example, is probably using a constant-rate birth-death model with species sampling [105, 166]. This model has three parameters,

1. The rate of speciation, λ
2. The rate of extinction, μ
3. The proportion of species that belong in this group which we have sequenced, Φ_0 (generally called ρ in the literature).

While we generally assume that we know Φ_0, λ and μ are estimated and biologically interesting. A group of species with a high speciation rate is adding new species quickly, but if $\mu \approx \lambda$, then the total diversity of the clade is relatively static. Comparing these parameters across clades can tell us about factors that drive speciation and extinction, and we can try to understand the processes that have shaped biodiversity on Earth. In the next sections, we will explore reasons we care about these, and related, model parameters in more detail.

1.3 Trees as data

As we will see in the next two chapters, there are cases where we are not very interested in the tree itself, but what we really care about is the tree prior, $\Pr(T)$. For example, our intrepid herpetologist may want to know whether the uplift of a mountain range was important for speciation in a particular clade of frogs. Geographic isolation can cause speciation, so isolating populations of many species on two different sides of a mountain could lead to many new species. In this case, the tree topology, and even the specific divergence times, are not terribly important. Instead, the question is whether there is an elevated speciation rate during the mountain uplift. Such questions make a Bayesian modeling framework natural, as we can sample the tree with MCMC, and then marginalize it out after the fact. Where possible, marginalizing is better than using a single tree as an estimate, as there can be considerable uncertainty in the tree and the divergence times, and this matters for the model parameters we will use to answer our questions!

In this thesis, our tree priors birth-death models, and we are primarily concerned with time-varying birth-death models, as rates of birth and death are not constant through time. Specifically, the models we will see are all birth-death-sampling models [*e.g.* 166, 135, 49], as we must have a sample (with some character data available) to put it in our tree. The sampling process is both a strength and a difficulty of birth-death models, as it adds a lot of information but is not always interpretable and its estimation does not come without difficulty.

In macroevolutionary applications, such as the one our herpetologist confronts, births are speciation events (represented by $\lambda(t)$, the time-varying birth rate), deaths are extinction events (represented by $\mu(t)$, the time-varying death rate), and sampling is often called the fossilization rate (represented by $\phi(t)$, the time-varying sampling rate) [49]. Where $\lambda(t)$ and $\mu(t)$ are interpretable as the instantaneous rates of speciation and extinction, the sampling parameter is not interpretable as the instantaneous rate of the lineage fossilizing. This is because not all fossils end up in our tree. To be in the tree, someone has to discover and

then describe a fossil's morphology, so other processes, like the intensity of looking for fossils in a given rock stratum, can effect this rate. Macroevolutionary applications also require a parameter to account for sampling species at the present day (Φ_0), as most phylogenies are missing at least some extant species. We are usually not interested in either sampling parameter directly, we simply need them to understand how the speciation and extinction rates have changed through time.

Phylogenetics can also be useful in epidemiology, where it is often referred to as phylodynamics (as one is using phylogenies to investigate infectious disease dynamics). Infectious diseases are caused by many things, but most causes have genomes which mutate quite rapidly. Thus, the process of infection spreading through a population leaves behind an evolutionary history that we can infer. Here, births occur when one individual infects another (making $\lambda(t)$ a time-varying infection rate) while deaths are recoveries (making $\mu(t)$ a time-varying recovery rate) [49]. As sampling is often accompanied by either treatment or individuals adopting behaviors that mitigate transmission, we can add a parameter $r(t)$ which models the probability that an individual who is sequenced becomes noninfectious [139, 49]. In chapters 2 and 3, we will refer to $r(t)$ as the "treatment parameter." In disease-modeling contexts, we are generally interested most in a compound parameter, the effective reproductive number, $R_e(t)$, which is the average number of people that a newly infected individual will go on to infect. It can be shown that $R_e(t) = \lambda(t)/(\mu(t) + \phi(t)r(t))$ [49], such that we can obtain $R_e(t)$ from our birth-death model parameters. Thanks to easy to use implementations in software, phylodynamic modeling has historically made a practice of integrating out uncertainty in the tree.

1.4 A brief guide to the rest of the thesis

Chapters 2 and 3 focus on birth-death process models, and are thus examples of phylogenetics where trees are data. The work for these two chapters overlaps in many ways. In Chapter 2 I use priors developed in Chapter 3. In Chapter 3, a number of analyses use the model developed in Chapter 2 (rather, a simpler sub-model using the same implementation). I

have ordered them such that we first explain how to compute likelihoods of time-varying birth death models, and then we visit the matter of priors for time-varying rates of birth and death. In Chapter 4 I pivot and examine the Bayesian inference machinery that the field uses to infer phylogenies, and try to quantify how well the inference procedure works. It is therefore focused on the case where the phylogenetic tree itself is a focal parameter. In Chapter 5, I contemplate future directions based on Chapters 2–4.

Chapter 2 is a version of Magee, AF and Höhna, S, 2021. Impact of K-Pg Mass Extinction Event on Crocodylomorpha Inferred from Phylogeny of Extinct and Extant Taxa. bioRxiv. Chapter 3 is a version of Magee AF, Höhna S, Vasylyeva TI, Leaché AD, and Minin VN (2020) Locally adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. PLOS Computational Biology 16(10): e1007999. Chapter 4 has not yet appeared elsewhere, but has been authored by Magee, AF, Karcher, MD, Matsen, FA, and Minin, VN.

Chapter 2

IMPACT OF K-PG MASS EXTINCTION EVENT ON CROCODYLOMORPHA INFERRED FROM PHYLOGENY OF EXTINCT AND EXTANT TAXA

2.1 Abstract

Crocodilians and their allies have survived several mass extinction events, including the K-Pg (end-Cretaceous) mass extinction event in which other clades of archosaurs, namely all non-avian dinosaurs and pterosaurs, went extinct completely. Surprisingly, the impact of the K-Pg mass extinction event on crocodylomorphs is debated and widely considered as minor or non-existent. In this chapter, we develop a new phylogenetic approach with extinct and extant species as tips of the time-calibrated phylogeny. To utilize this combined data and to infer the impact of the K-Pg mass extinction event, we derive the likelihood function for a time-varying (episodic) fossilized birth-death model that additionally incorporates mass extinctions and bursts of births. Contrary to previous research, we find strong evidence for a global signal of the K-Pg extinction event in crocodiles and their allies. This signal is robust to uncertainty in the phylogeny, the prior on the mass extinctions, and uncertainty in fossil ages.

2.2 Introduction

Crocodylomorpha (living crocodilians and their extinct relatives) is an ancient clade that first appeared in the late Triassic [94, 10]. Crocodylomorphs have survived several mass extinction events, including the Triassic-Jurassic extinction event (201.3 Ma), the Jurassic-Cretaceous extinction (145 Ma), the Cretaceous-Paleogene extinction event (K-Pg extinction, 66 Ma) and the Eocence-Oligocene extinction event (33.9 Ma). The fossil records shows

significantly higher historical diversity with a peak in Jurassic compared to the diversity of extant crocodilians [11]. Previous studies have found evidence for the Triassic-Jurassic extinction [94] and the Jurassic-Cretaceous extinction [146] although some debate of the true impact remains [41]. Most surprisingly, previous studies assessed the impact of the K-Pg mass extinction on the whole Crocodylomorpha clade as minor or non-existent [94, 150, 10, 11] although sister clades, such as non-avian dinosaurs and pterosaurs, with similar lifestyle (*e.g.* body size, habitat and food sources) and anatomy where severely impacted and went extinct 66 Ma. However, it has been noted that several lineages of crocodylomorphs went extinct at the K-Pg boundary [92], and turnover of crocodylomorph communities in Europe was elevated at the K-Pg boundary [24, 114]. These focused studies on sub-groups of all crocodylomorphs indicate evidence of the K-Pg mass extinction, but the impact on the whole clade remains elusive.

Previous studies used fossil occurrence data to address what impact the K-Pg mass extinction had on crocodylomorphs [94, 92]. Alternatively, time-calibrated phylogenies of extant taxa have been used successfully to infer the impact of mass extinction events [23, 95, 25], *e.g.* on fishes (Tetraodontiformes) [2] and geckos [9]. However, the signal of mass extinction erodes the further back the event was in the past [95]. The age of the crown group of living crocodylomorphs (Crocodylia) dates back to the Campanian with estimates ranging between 90 to 118 Ma [106, 162]. Even considering the older estimates of the age of the crown group, only three lineages that crossed the K-Pg boundary have living descendants (ancestral alligators, crocodiles, and gharials). Thus, a time-calibrated phylogeny of living crocodilians alone is also not sufficient to show the impact of the K-Pg mass extinction on crocodylomorphs.

Instead, we suggest that placing fossil crocodylomorphs as tips into a time-calibrated phylogeny with or without extant taxa and adopting the powerful methods of phylogenetic diversification studies can shed a new light into this question. The time-calibrated phylogeny of extant crocodilians provides insights into the pattern of recent biodiversity changes whereas fossilized crocodylomorphs provide insights about the more ancient past. The two data

sources combined into a single phylogeny with extinct and extant taxa provide a holistic view of the historical biodiversity.

Here, we develop a new statistical approach using phylogenies of both living and extinct species. Our approach extends previous fossilized-birth-death models [135, 57, 49, 132]. Specifically, our new model includes piecewise-constant (episodic) rates of speciation, extinction, and fossilization, plus instantaneous events of mass extinction with an estimated survival probability [136, 60, 95]. In our phylogenetic model, the rates of speciation and extinction are the rates of origination and termination of lineages in the tree (*e.g.* [104, 108, 99]). The fossilization rate represents the rate of fossils placed in the phylogeny, and thus is composed of (a) a lineage leaving a fossil behind, (b) the fossil being recovered, and (c) the fossil being placed into the phylogeny. Thus, the fossilization rate is influenced by several biases, such as fossil recovery and sampling, and is itself challenging to interpret at face value [158]. For mathematical completeness and availability for future research, we also complement the continuous events of speciation and fossilization with instantaneous events of speciation bursts (*e.g.* [107]) and mass fossilization. By leveraging information from both the ages of included fossils and from the inferred speciation times, our approach provides a holistic view of diversification. We also draw a link to infectious disease modeling and present our new fossilized birth-death process as a generalization of both macroevolutionary and phylodynamic models (*e.g.* [140, 74]).

We fit our new model to a recently published phylogeny of crocodylomorphs including extinct and extant species [162]. We use a reversible-jump Markov chain Monte Carlo algorithm [54] to infer whether a mass extinction event has impacted our study group or not. We use Horseshoe Markov random field (HSMRF) smoothing priors [90] on speciation, extinction, and fossilization rates to allow for abrupt changes while favoring constant rates over rate variation (Occam's razor principle). We infer an overwhelming support for a mass extinction event about 68 Ma. We show that this signal is robust to a number of factors, including the prior on the mass extinctions, the assumed phylogeny and fossil ages. Posterior predictive simulations and a post-hoc simulation study confirm that our approach is powerful, robust,

and has a low false positive rate. Thus, in addition to providing a new, comprehensive birth-death model, and the first empirical application of the fossilized birth-death process with mass extinctions, we demonstrate a rigorous Bayesian workflow for interrogating the phylogenetic evidence for a mass extinction.

2.3 Methods

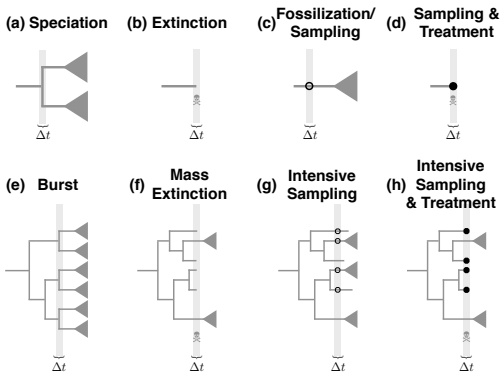


Figure 2.1: Schematic of possible events. The top row shows continuously occurring events affecting a single lineage and the bottom row shows the same types of events but instantaneously and affecting all lineages (tree wide). The types of events are speciation, extinction, fossilization/sampling, and sampling with treatment.

2.3.1 Notation

We begin with the notation of the generalized fossilized-birth-death process (or birth-death-sampling-treatment process as it might be called in the phylodynamics literature). At any point in time, lineages may speciate (infect another individual), go extinct (become noninfectious), or fossilize (be sampled). Additionally, to allow for phylodynamic applications, when a lineage is sampled it may be treated and then become noninfectious (go extinct). Each of these four types of events may occur continuously over time affecting a single lineage (Figure 2.1 a-d) or instantaneously at a pre-defined time affecting all lineage alive at the time (Figure 2.1 e-h). The continuously occurring events are governed by time-varying rates

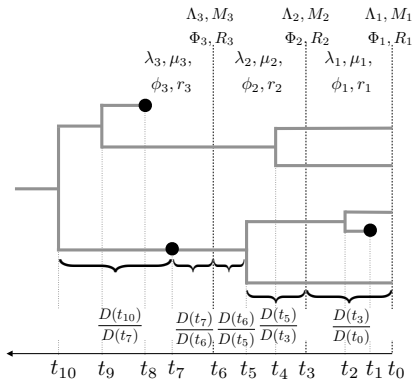


Figure 2.2: Schematic decomposition of probability density function for our generalized episodic fossilized birth-death process. In this example, we have three epochs and thus three rates (*e.g.* $\{\lambda_1, \lambda_2, \lambda_3\}$) for each continuous type event and three probabilities (*e.g.* $\{\Lambda_1, \Lambda_2, \Lambda_3\}$) for each instantaneous tree-wide event. Each branch is broken into branch segments that starts (tipwards) at t_y and ends at t_o so that within each branch segment no epoch switch, speciation or fossilization event occurs. We compute the probability of observing each branch segment by $\frac{D(t_o)}{D(t_y)}$, which constitutes the core part of the probability density computation.

$\lambda(t)$, $\mu(t)$, $\phi(t)$, and $r(t)$ for speciation, extinction, fossilization/sampling and treatment (in a macroevolutionary setting, this can allow relaxation of the assumption that fossilization and extinction are independent). The instantaneous, tree-wide events are defined by a vector of pairs of times and probabilities (Λ_i, M_i, Φ_i, and R_i respectively). For a summary of the notation and model parameters, see Table 2.1.

2.3.2 The probability density of a phylogenetic tree

Our derivation of the probability density of a phylogenetic tree with extinct and extant taxa under the generalized episodic fossilized-birth-death process breaks up each branch of the tree into segments belonging only to one epoch (Figure 2.2). This derivation follows the same logic of Stadler et al. [137] and Gavryushkina et al. [49] but additionally includes mass

extinction and burst events. The probability density of a phylogenetic tree Ψ is

$$
\begin{aligned}
f(\Psi) = \quad & \frac{2^{I+H-||\mathcal{A}||-1}}{(I+H-||\mathcal{A}||)!} && \text{(i)} \\
& \times \prod_{t \in \mathcal{N}} \Big[\lambda(t) \Big] && \text{(ii)} \\
& \times \prod_{t \in \mathcal{F}} \Big[\phi(t) \big(r(t) + \big(1 - r(t)\big) E(t) \big) \Big] && \text{(iii)} \\
& \times \prod_{t \in \mathcal{A}} \Big[\phi(t) \big(1 - r(t)\big) \Big] && \text{(iv)} \\
& \times \prod_{i=1}^{l} \Big[\Lambda_i^{K_i} \big(2\Lambda_i E(s_i) + (1 - \Lambda_i) \big)^{L(s_i) - K_i} \Big] && \text{(v)} \\
& \times \prod_{i=1}^{l} (1 - M_i)^{L(s_i)} && \text{(vi)} \\
& \times \prod_{i=0}^{l} \Big[(1 - \Phi_i)^{(L(s_i) - I_i)} \Phi_i^{I_i} (1 - R_i)^{T_i} \\
& \qquad (R_i + (1 - R_i) E(s_i))^{I_i - T_i} \Big] && \text{(vii)} \\
& \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)} \Big] && \text{(viii)}
\end{aligned}
\tag{2.1}
$$

where (i) the probability of the topology, (ii) the probability of the observed serial speciation events, (iii) the probability of the fossil (serially-sampled) tips (Figure 2.1c-d), (iv) the probability of the sampled ancestors (fossils) (Figure 2.1c), (v) the probability of the observed and unobserved speciation events at tree-wide speciation burst events (Figure 2.1e), (vi) the probability of all lineages surviving a tree-wide mass extinctions events (Figure 2.1f), (vii) the probability of all the observed fossil sampling times at given tree-wide fossilization/sampling events (Figure 2.1g), (viii) the probability of the observed branch segments (Figure 2.2).

Table 2.1: Model parameters and their interpretation

Parameter	Interpretation
t_{or}	the starting time of the process (can be the origin or the most recent common ancestor).
$\lambda(t)$	birth rate at time t.
$\mu(t)$	death rate at time t.
$\phi(t)$	fossilization or (serial) sampling rate at time t.
$r(t)$	probability that a ϕ-sampling event at time t causes a death event.
l	the number of breaks between intervals/episodes.
$\boldsymbol{s}$	the beginning times of the episodes/intervals, $s_0, \ldots, s_l$, with $s_0 = 0$.
$\boldsymbol{\Lambda}$	vector of l burst probabilities. At time s_i ($i \geq 1$) every lineage speciates with probability Λ_i.
$\boldsymbol{M}$	vector of l extinction probabilities. At time s_i ($i \geq 1$) every lineage dies with probability M_i.
$\boldsymbol{\Phi}$	vector of $l+1$ fossilization/sampling probabilities. At time s_i every lineage leaves a fossil/is sampled with probability Φ_i.
$\boldsymbol{R}$	vector of l treatment probabilities. At time s_i ($i \geq 1$) every sampled lineage dies with probability R_i.
$\mathcal{A}$	set of ages of all ϕ-sampled sampled ancestors.
$\mathcal{B}$	set of all branch segments, such that t_o is the beginning of the segment and t_y the end ($t_y < t_o$).
$\mathcal{F}$	set of ages of all ϕ-sampled tips.
$\mathcal{N}$	set of ages of all non-burst bifurcation (birth) events.
$L(t)$	the number of lineages in the reconstructed tree alive at time t.
H	total number of ϕ-samples.
I_i	number of Φ-samples at the ith Φ-sampling event.
I	total number of Φ-samples, with $I = \sum_{i=0}^{l} I_i$.
T_i	number of tips with sampled descendants in a given Φ-sampling event.
K_i	number of lineages with a burst event at the end of episode i.
K	total number of $K = \sum_{i=0}^{l} K_i$.

2.3.3 Probability of no sampled descendant $E(t)$

We begin our derivation by tracking $E(t)$, the probability that a lineage alive at time t has no sampled descendants when the process is stopped. Let us define the differential equation for E in a small time interval Δt to be,

$$E(t+\Delta t) = \underbrace{\mu(t)\Delta t}_{\text{a}} + \underbrace{(1-\mu(t)\Delta t)}_{\text{b}}\underbrace{(1-\phi(t)\Delta t)}_{\text{c}} \times \left[\underbrace{(1-\lambda(t)\Delta t)E(t)}_{\text{d}} + \underbrace{\lambda(t)\Delta t E(t)^2}_{\text{e}}\right] \tag{2.2}$$

which accounts for the three scenarios that can occur during the interval Δt, (1) the lineage dies (a), (2) no death (b) and no sampled fossil/serial sampling (c) and no birth (d), (3) no death (b) and no sampled fossil/serial sampling (c) and a birth event (e). Cases (2) and (3) require the extant lineage(s) to eventually die or remain unsampled, hence $E(t)$ and $E(t)^2$. Rearranging the equation, discarding all terms on the order of Δt^2, and dividing by Δt, we get,

$$\frac{dE}{dt} = \mu(t) - E(t)(\lambda(t) + \mu(t) + \phi(t)) + \lambda(t)E(t)^2, \tag{2.3}$$

where again we can see three cases, (1) the process dies, (2) nothing happens followed by $E(t)$, and (3) a birth followed by both lineages not being observed, $E(t)^2$.

We now make the assumption that $\lambda(t)$, $\mu(t)$, and $\phi(t)$ are piecewise-constant (see [136, 60]). Additionally, we assume that these rates share the same intervals, *i.e.* changes in rates may occur at the same time points. We keep track of these times with the vector $\boldsymbol{s}$, of which l elements must be specified, $s_1, \ldots, s_l$. We implicitly specify $s_0 = 0$ and $s_{l+1} = \infty$, such that the first interval begins at time 0 and extends to time s_1, and the last interval contains time up to infinity to avoid any orphaned events. In the ith interval, the birth rate is λ_i, the death rate is μ_i, the fossilization/serial sampling rate is ϕ_i. At the boundaries of the interval, we allow for each lineage to give birth, die, or be sampled with probabilities Λ_i, M_i, and Φ_i respectively. The intervals are left-inclusive so that the process is in interval i if

$s_{i-1} \leq t < s_i$. We assume that at any event time s_i, at most one of $\{\Lambda_i, M_i, \Phi_i\}$ is nonzero (that is, we forbid there to be multiple types of events at a single event time).

To compute $E(t)$, we need to compute $E_i(t)$ for all intervals, where $E_i(t)$ depends on all $E_j(t)$ for $j < i$. In this setup, we can analytically solve the differential equation for E and obtain,

$$E_i(t) = \frac{\lambda_i + \mu_i + \phi_i - A_i \frac{1+B_i - e^{-A_i(t-s_i)}(1-B_i)}{1+B_i + e^{-A_i(t-s_i)}(1-B_i)}}{2\lambda_i}, \tag{2.4}$$

where

$$A_i = \sqrt{(\lambda_i - \mu_i - \phi_i)^2 + 4\lambda_i \phi_i} \tag{2.5}$$

and

$$B_i = \frac{(1-2C_i)\lambda_i + \mu_i + \phi_i}{A_i}, \tag{2.6}$$

and where C_i is defined as

$$\begin{aligned} C_i = & \underbrace{\mathbb{I}_{(\Lambda_i=0, M_i=0, \Phi_i>0)}\big((1-\Phi_i)E_{i-1}\big)}_{\text{a}} \\ & + \underbrace{\mathbb{I}_{(\Lambda_i>, M_i=0, \Phi_i=0)}\big((1-\Lambda_i)E_{i-1}(s_i) + \Lambda_i E_{i-1}^2(s_i)\big)}_{\text{b}} \\ & + \underbrace{\mathbb{I}_{(\Lambda_i=0, M_i>0, \Phi_i=0)}\big((1-M_i)E_{i-1}(s_i) + M_i\big)}_{\text{c}} \\ & + \underbrace{\mathbb{I}_{(\Lambda_i=0, M_i=0, \Phi_i=0)}\big(E_{i-1}(s_i)\big)}_{\text{d}} \end{aligned} \tag{2.7}$$

for $i \geq 1$ and with $C_0 = 1$. C_i enables us to propagate $E_{i-1}(t)$ into $E_i(t)$, by accounting for the various tree-wide events that may take place at the end of each episode. In C_i, each element corresponds to one of the possible event types (or the absence of any event type), (a) this is an intensive sampling time where the lineage is unsampled, and the lineage is unobserved between time s_i and the present, (b) this is a burst time and either the lineage does not experience the burst event and is unobserved between time s_i and the present, or it experiences the burst without having either descendant observed between time s_i and the

present, (c) this is a mass extinction event and either the lineage survives the mass extinction but is unobserved between time s_i and the present or the lineage dies in the mass extinction, (d) there is no event at this time and the lineage is unobserved between time s_i and the present. Terms (a)-(d) are mutually exclusive as we forbid more than one event at any time.

2.3.4 Probability along an observed lineage

Next, we define the differential equation computing the probability of a branch-segment along an observed lineage (Figure 2.2). A branch-segment terminates at a birth event or a fossilization/sampling event, and over a small time Δt we have,

$$D(t+\Delta t) = \underbrace{(1-\mu(t)\Delta t)}_{\text{a}}\underbrace{(1-\phi(t)\Delta t)}_{\text{b}}\times \left[\underbrace{D(t)(1-\lambda(t)\Delta t)}_{\text{c}} + \underbrace{2D(t)\lambda(t)\Delta t E(t)}_{\text{d}}\right] \tag{2.8}$$

which accounts for the two scenarios that can apply over an interval, (1) no death (a) and no fossilization/serial sampling (b) and no birth (c), (2) no death (a) and no fossilization/serial sampling (b) and a birth event (d) but only one of the lineages is observed. As before, we can rearrange the above equation into a differential equation, yielding

$$\frac{dD}{dt} = -D(t)(\lambda(t)+\mu(t)+\phi(t)) + 2D(t)\lambda(t)E(t). \tag{2.9}$$

The analytical solution to this differential equation for piecewise-constant rates is

$$\begin{aligned} D_i(t) = \quad & D_{i-1}(s_i) \\ \times \quad & \frac{4e^{-A_i(t-s_i)}}{(1+B_i+e^{-A_i(t-s_i)}(1-B_i))^2} \end{aligned} \quad . \tag{2.10}$$

Note that this definition of $D(t)$ includes only the continuous-time portions of the model, in other words we leave out the probabilities of all the tree-wide events. This requires us to keep

track of these separately, but makes our likelihood easier to compare to that of Gavryushkina et al. [49].

2.3.5 Conditioning

When employing birth-death process models, we may wish to condition our model on having some tip samples (extinct or extant) in the tree or the tree surviving to the present day. Both options are dependent on whether we presume the tree begins at the origin (t_{or}, with a single lineage) or at the most recent common ancestor (t_{MRCA}, with two lineages). We provide a full description of the different conditioning options in the Appendix A.10.

2.3.6 Validation of Theory and Implementation

Our generalized episodic fossilized-birth-death process is implemented in the open source software `RevBayes` [63]. We performed simulation based calibration to validate our implementation (see Appendix A.16 and Figure A.16). We additionally validated our implementation within our simulation study testing the power and false-discovery rate of mass extinction events as well as the bias in speciation, extinction, and fossilization rates. We provide more details of our validation of the underlying theory and implementation in the Appendices A.11 and A.16.

2.3.7 Priors and parameterization

To account for background variation in the rates of speciation, extinction, and fossilization, we apply horseshoe Markov random field (HSMRF) prior distributions [90], which have been shown to be able to both detect rapid shifts in speciation rates and to reject time-varying models in favor of effectively constant-rate models. In a HSMRF model, we must specify prior distributions on the initial rates λ_1, μ_1, and ϕ_1. We follow May et al. [95] in taking an empirical Bayes approach, first estimating parameters of a constant-rate fossilized-birth-death model, then fitting Gamma distributions to the posterior distributions of λ, μ, and ϕ

and choosing the prior distributions to have twice the variance of the posterior distributions.

In our analyses, we discretize time into 100 evenly spaced epochs of 2.435 million years ($s_0, s_1, s_2, \ldots = 0, 2.435, 4.870, \ldots$). At every time greater than 0, we allow for the possibility of a mass extinction event. We use reversible-jump Markov chain Monte Carlo [54, 48] in order to infer whether there is evidence for a mass extinction at each of these times. To determine the prior probability p_i of a mass extinction at each event time, we first choose a prior expectation on the number of mass extinction events, $\mathbb{E}[n_{\text{ME}}]$, and set $p_i = \mathbb{E}[n_{\text{ME}}]/99$. To account for prior sensitivity, we perform analyses with a range of prior expectations, $\mathbb{E}[n_{\text{ME}}] = \{0.1, 0.5, 1.0, 2.0, 5.0\}$. For the (per-lineage) probability of death at a mass extinction event, M_i, we use a Beta(18,2) distribution, which has a prior mean corresponding to 90% of lineages going extinct. As having no mass extinction at time s_i is mathematically equivalent to having a mass extinction with $M_i = 0$, this makes the mixtures a $p, (1-p)$ mix of a mass extinction probability of 0 and a Beta(18,2) distribution. To determine whether our results are robust to phylogenetic uncertainty, we also perform these 5 analyses on 6 distinct trees inferred by Wilberg et al. [162], for a total of 30 analyses.

2.3.8 Convergence diagnostics

For all analyses (including those of simulated data), we ran 4 independent replicate MCMC simulations. To avoid convergence-related issues, we keep only the converged subset of these chains that pass a potential scale reduction factor (PSRF) cutoff [12]. As heavy-tailed distributions like the horseshoe distribution can cause problems for standard convergence diagnostics, we use the rank-stabilized PSRF of Vehtari et al. [153]. After applying these standards (requiring PSRF < 1.01), we retained at least 3 chains for each empirical replicate analysis.

2.3.9 Simulation study to test power and false-discovery rate

To determine the power of our method to detect mass extinctions, and its false positive rate, we performed analyses on 400 simulated trees. Of these, 200 are simulated from the

posterior distribution on $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\phi}$, and $\boldsymbol{M}$ with a prior expectation of 0.5 mass extinctions (power analysis). The other 200 trees are simulated from a similar analysis but where no mass extinctions were allowed (false-positives analysis). By estimating diversification rates for a phylogeny that experiences a mass extinction and mis-specifying the model by disallowing mass extinctions in the analysis, we allow for the possibility that the inferred diversification rates produce trees that appear to have experienced mass extinctions, providing a robust test of the false positive rate. The inference settings for these simulated datasets was identical to the empirical analyses.

2.3.10 Code and data availability

The episodic fossilized-birth-death model is implemented in the software `RevBayes` [63], available at https://revbayes.github.io. Data and scripts for analyses and simulations are available at https://github.com/afmagee/gefbd and a tutorial is available at https://revbayes.github.io/tutorials/divrate/efbdp_me.html.

2.4 Results

2.4.1 The Generalized Episodic Fossilized-Birth-Death Process

We derived the probability density function of a phylogeny with extinct and extant tips under the generalized episodic fossilized-birth-death process (see Equation 2.1 in our Methods). Our generalized model combines previous special cases for macroevolution (*e.g.* [136, 49, 60, 95, 132]) and phylodynamic models (*e.g.* [140, 49]) and extends the existing models to include mass extinction events as well as bursts of birth events (see Appendix A.13 for a discussion between our and previous models). Our generalized episodic fossilized-birth-death process collapses to exactly the same processes and probability density function of these special cases (see Appendix A.14). Thus, our model provides a flexible framework to explore new and previously described models within the same framework. Most importantly, our model enables robust estimation of variation in diversification rates using phylogenies with

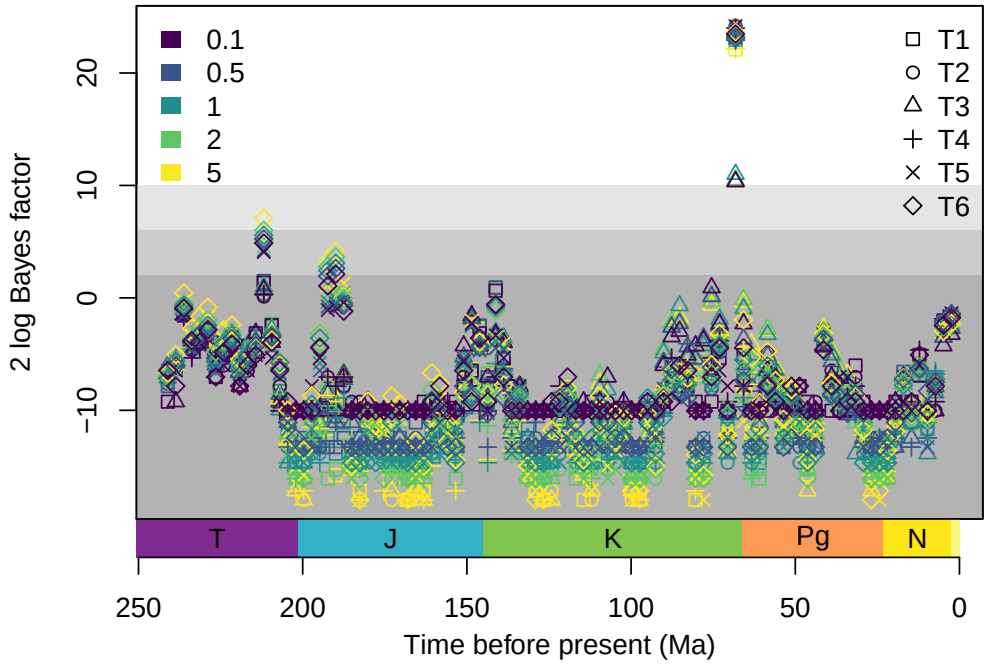


Figure 2.3: Per-interval support for mass extinctions in the Crocodylomorpha dataset for all 30 combinations of dataset and mass extinction prior. Mass extinctions are allowed only at the end of each of the 100 intervals, and all interval times are fixed. Different mass extinction priors are shown by color, and different datasets (six different phylogenetic trees inferred by Wilberg et al. [162], denoted here T1-T6) by symbols. Shaded regions denote 2ln Bayes Factor cutoffs of 2, 6, and 10, which correspond to weak support, support, and strong support for a mass extinction[71]. Across all analyses there is a strongly supported mass extinction that corresponds to the time of the K-Pg boundary. Geological periods are shown for convenience with standard abbreviations (from left to right, Triassic, Jurassic, Cretaceous, Paleogene, Neogene, and Quaternary (unlabeled)).

extant and/or extinct taxa. The probability density function is implemented in the software `RevBayes` [63]. This implementation allows both for flexible Bayesian parameter estimation and use of the generalized episodic fossilized-birth-death process as a prior distribution for divergence time estimation (*e.g.* see [57]).

2.4.2 K-Pg mass extinction in Crocodylomorpha

We tested for the effect of mass extinctions in Crocodylomorpha using the dataset of Wilberg et al. [162]. The tree has a root age of 243.5 million years, and includes 14 of the 23 known extant species and 128 fossil tips [162, 106]. To account for the possible effect of background

rate variation, and for the possibility that a mass extinction is followed by a rapid burst of speciation [23], we employed priors for time-varying rates on the speciation and extinction rates through time [90]. Similarly, we employed time-varying priors on the fossilization rate through time to account for fossil sampling biases and other temporal factors. We computed Bayes factors to test for the signal of a mass extinction event anytime along the phylogeny. We explored the robustness of results using six published phylogenies by Wilberg et al. [162] (we refer to these as trees T1-T6) and five different priors on the expected number of mass extinction events.

We find very strong evidence (2ln Bayes Factor > 10 for all dataset and prior combinations) for a mass extinction event approximately 68 Ma, which is almost certainly the K-Pg mass extinction (Figure 2.3, see Appendix Figure A.13 for an example of the estimated mass extinction probability). We find weak positive evidence for a signal of the Tr-J (end-Triassic) mass extinction, though the timing is less certain. The weaker signal for this ancient mass extinction event is most probably due to the age of the mass extinction event, as the signal of a mass extinction event within the first 25% of the age of the phylogeny is often erased [95]. We also find a very weak signal for the J-C (end-Jurassic) mass extinction. Overall, we recovered some signature for all four major mass extinctions that crocodylomorphs have survived and our method shows no notable support for any other mass extinction between the known events. Interestingly, there is a very weak signature of a mass extinction towards the present. This indicates that living crocodilians are in decline and might face yet another mass extinction, which is further corroborated by our estimated speciation and extinction rates (Figure 2.4).

Approximately 25 Ma, the speciation rates began a sharp decline, with extinction rates exceeding speciation rates in the recent past (Figure 2.4). This decline in net-diversification rate matches the previously observed late Cenozoic peak in crocodylian diversity [128]. The estimated fossilization rates exhibit substantial variation. First, the spike in fossilization around the K-Pg boundary is likely the result of the intense interest in studying the K-Pg boundary [148, 4, 157]. Thus, our flexible time-varying fossilization rate picks-up on

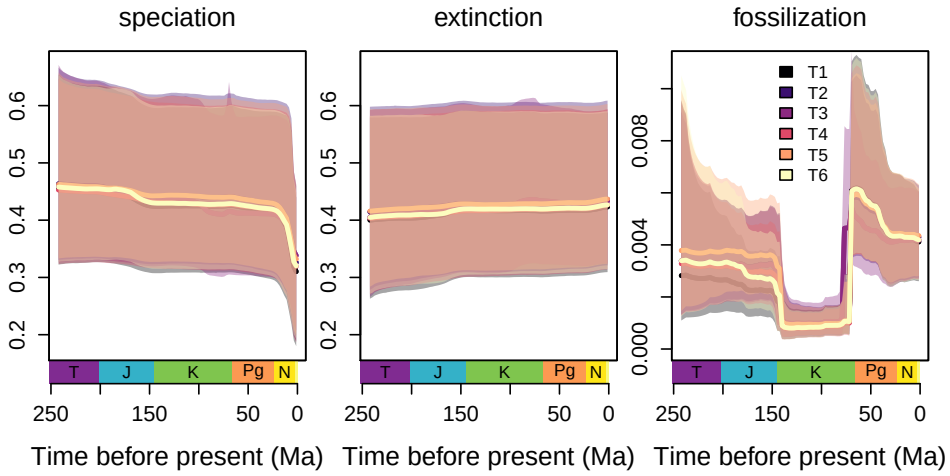


Figure 2.4: The rates of speciation, extinction, and fossilization through time for one mass extinction prior ($\mathbb{E}[n_{\text{ME}}] = 0.5$) and all datasets (six different inferred by Wilberg et al. [162], denoted here T1-T6). The solid line depicts the posterior median estimate and the shaded areas the 95% credible intervals, colored by dataset. Speciation rates began to decline steeply approximately 25 Ma, leading to a negative rate of net diversification. Estimates are broadly similar across all datasets and mass extinction priors, full plots are available in Figure A.1. Geological periods are shown for convenience with standard abbreviations (from left to right, Triassic, Jurassic, Cretaceous, Paleogene, Neogene, and Quaternary (unlabeled)).

this known fossil sampling bias and takes this bias automatically into account. The lower fossilization rate during the Cretaceous does by no means imply a lower number of fossil observations from that epoch. Instead, the expected number of observed fossil can be computed by the product of the estimated historical diversity (see Appendix A.5 and Figure A.9) and the estimated fossilization rate. Thus, although the fossilization rate during the Cretaceous is estimated to be lower, we infer high number of observed fossils during the Cretaceous with a peak at the K-Pg boundary (see Appendix A.6 and Figure A.10). Reassuringly, the estimated speciation, extinction, and fossilization rates are in strong agreement regardless of the specific phylogeny and prior probability on the number of mass extinction events (Figure A.1).

2.4.3 Power and False Positive Rate

We explored the power and false-discovery rate of detecting mass extinction events using simulations. We find that our method has a high power to detect mass extinctions, as the

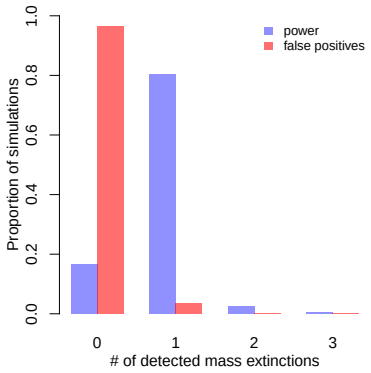


Figure 2.5: The number of detected mass extinctions across our 400 simulated data analyses. Detection is defined as a mass extinction for which the 2ln Bayes Factor is at least 10[71, 95]. The "power" simulations (blue, left bars) are simulations where in truth there was a mass extinction at the K-Pg boundary. In most of these simulations, a mass extinction is detected (and at the correct time), indicating good power to detect the K-Pg mass extinction. The "false positive" simulations (red, right bars) are simulations where in truth there was *no* mass extinction at the K-Pg boundary. In most of these simulations, no mass extinction is detected, indicating a low rate of false positives.

simulated K-Pg mass extinction is detected in 82% of simulations (Figure 2.5). We also find that our method has a low false positive rate, as 4% of simulations with mass extinctions detect a mass extinction at an incorrect time, and 4% of simulations without mass extinctions detect any mass extinction (Figure 2.5). This simulation shows that our inferred signal of the K-Pg mass extinction is indeed a true signal.

2.4.4 Model Adequacy

Our empirical results provided very strong evidence for an impact of the K-Pg mass extinction on crocodylomorphs. Nevertheless, even the best supported model might be inadequate for a given dataset [6, 14]. We tested model adequacy of both a model with and without the K-Pg mass extinction using posterior predictive simulations. Our model with mass extinction was slightly more adequate than our model without mass extinction.

Across 17 posterior predictive tests, both models were generally adequate and inadequate according to the same summary statistics respectively (Appendix Figure A.8). Overall, our generalized episodic fossilized birth-death model with mass extinction showed a very good fit

except for 5 of the 17 summary statistics: Colless' imbalance statistic [19], Tippyness [123] (as well as tippyness restricted to either extant or extinct tips) and the number of sampled ancestors. The deviation from the expected tree balance shows that the observed crocodilian phylogeny was most likely produced by a process of lineage heterogeneous speciation and extinction rates. The empirical number of sampled ancestors, which is zero, is most likely an artifact of the tree construction method which did not allow for fossils to be sampled ancestors [162]. Our model can be modified to enforce zero sampled ancestors by assuming that a fossilization event is always combined with an extinction event. Our results (Appendix Figure A.5) show that the inference of the K-Pg mass extinction is robust to the systematic bias of not allowing for sampled ancestors.

The only notable difference in adequacy between the two models comes from one test statistic devised specifically to examine the effect of the mass extinction, which focuses on the lineage through time curve at the K-Pg boundary. Simulations based on a model with mass extinctions produce a sharp decline in the number of lineages around the K-Pg boundary, which is also present in the crocodylomorph tree. However, this sharp decrease is not present in trees simulated from a model with no mass extinction (Appendix Figure A.8). The drop corresponds to an observable band of fossils in these (simulated and empirical) trees, a visual signal of the mass extinction. In this case, the model with mass extinctions is adequate while the model without is inadequate.

2.4.5 Signal from Extant vs Extinct Species

Previous research debated whether molecular phylogenies from extant taxa need the fossil record to reliably infer historical diversification patterns [118, 100]. Our generalized episodic fossilized birth-death process bridges analyses from the fossil record and molecular phylogenies. Nevertheless, we performed analyses using only the extant tips (pruning all fossils) and only using extinct tips (pruning all extant taxa). In the case of crocodilian diversification, the strongest signal is contained in the fossil taxa (see Appendix Figure A.3). Thus, molecular phylogenies may indeed need the fossil record to obtain a complete picture of his-

torical diversification processes. However, the importance of the fossil taxa for our analyses is not surprising because the phylogeny by Wilberg et al. contains ten times more fossil taxa than extant taxa and the crown age of the extant taxa is too recent to show a signal of the K-Pg mass extinction or other older mass extinctions. On the other hand, the phylogeny constructed of fossil crocodylomorphs alone was not able to pick up the sharp decline in diversification in the Neogene (see Appendix Figure A.2). This results confirms our prediction that fossil taxa provide more information about deep time historical diversification patterns whereas extant taxa provide more information about recent time diversification patterns, and the full picture can only be seen with a combination of both data sources.

2.5 Discussion

In this chapter, we have described the generalized episodic fossilized birth-death model, which includes three previously ignored parameters and constitutes the most parameter-rich time-dependent birth-death model to date. The new parameters $\boldsymbol{\Lambda}$ (which models tree-wide bursts of births), and $\boldsymbol{M}$ (which models tree-wide bursts of deaths, *i.e.* mass extinctions) will primarily be of interest in macroevolutionary studies, though they may be of use to phylodynamic situations for modeling superspreader events and culling. The new parameter $\boldsymbol{R}$ (which models different treatment probabilities between continuous sampling and intensive sampling will primarily be of interest in phylodynamic applications. The generalized episodic fossilized birth-death model is implemented in `RevBayes` [63], and is therefore available for inferring diversification parameters on fixed trees and for use in full Bayesian joint inference of divergence times and diversification rates.

We applied the generalized episodic fossilized birth-death model to a phylogeny of 142 extant and extinct crocodylomorphs. In our Bayesian analysis, we find strong evidence for the K-Pg mass extinction in crocodylomorphs, a signal which we show is robust to phylogenetic uncertainty, the prior on mass extinctions, and the exclusion of extant taxa. Using post-hoc analyses of simulated data, we show that we have good power to detect this signal and a low risk of false positives. Further post-hoc analyses of the crocodylomorph dataset show that

the primary signal for the K-Pg mass extinction is found in the fossil taxa. In general, we expect that the use of fossil samples should greatly increase the power of phylogenetic birth-death models to detect mass extinctions. An extant tip only contains information about mass extinctions through the branching time that it adds to the tree. Sampled ancestors contain information only through the sampling time. Fossil tips, on the other hand, contain both sources of information (branching time and sampling time) and their inclusion should prove useful to the detection of mass extinctions and variation in diversification rates. Uncertainty in the fossil ages is unlikely to erase the signature of the K-Pg (Appendix A.7 and Figure A.11). Our results do show that diversification-rate estimates change when extant tips are excluded, and thus we advocate for using the largest possible dataset when inferring diversification rates through time and/or mass extinctions.

Previous studies have considered the effect of the K-Pg on crocodylomorphs to be minor or non-existant [94, 150, 10, 11], though the extinction of several lineages of crocodylomorphs at the K-Pg boundary has been noted [92], as has elevated turnover of crocodylomorph communities in Europe at the K-Pg boundary [24, 114]. The most important difference of our study to previous studies is that we developed and applied a fully likelihood based approach using a phylogeny of extinct and extant taxa. A full likelihood based approach should, in general, be more powerful and more robust, which we also observe in our simulation study. Additionally, our phylogenetic approach utilizes information from both the speciation times and the fossil sampling times. If the phylogeny is integrated out, then the information of the speciation times becomes lost [131, 138, 158]. As a phylogenetic approach, our approach makes use of all lineages irrespective of geography, ecotype, and other differentiation criteria; allowing a holistic view of the effect of the K-Pg. In future model extension, one could extend our model to allow for more nuanced mass extinction survival probabilities by, for example, assuming survival probabilities are linked to geography [114] and/or ecotype [162] and using state-dependent diversification processes [88]. Furthermore, it is possible to incorporate additional information about fossil sampling biases into our model by explicitly specifying epochs with higher or lower fossilization rate. Nevertheless, our HSMRF prior approach is

more conservative and recovered known fossil sampling biases [148, 4, 157]. This should be considered as additional evidence that our estimated rates are biologically meaningful.

Recently, it was shown that time-varying speciation, extinction, and fossilization rates cannot be identified from phylogenies alone [86, 85]. That is, if we would allow for any (arbitrary) continuous speciation, extinction, and fossilization rate function, then there are infinitely many combinations of rate functions with exactly the same likelihood for an observed phylogeny. However, our usage of shrinkage priors which guide the inference to smoother and less variable rate estimates [90] impose identifiability, there is a single combination of rates with the highest posterior density. Non-identifiability of speciation, extinction and mass extinction has been known [136, 60, 95], and our use of empirically motivated priors on the strength of mass extinction makes these models identifiable [95]. Biologically, a number of our findings match features found in other studies. Most notably, our fossilization rate recovers a clear signature of the increased interest in studying the K-Pg [148, 4, 157]. The declining diversification rates inferred in the Neogene mirrors a drop in non-marine crocodylomorph diversity seen by Mannion et al. [92] in both Africa and South America in that time span. The diversity through time predicted by our model (Figure A.9) shares several similarities with Markwick's [94] findings, including exponential diversification through the late Cretaceous and relatively constant diversity after the K-Pg through at least the middle Eocene. Taken together, we are confident that our inferred signal of the K-Pg mass extinction event on crocodylomorphs is robust and true. Interestingly, although speciation and extinction rates as well as mass extinctions might not be identifiable by the likelihood function alone, our posterior predictive simulations show that a model with mass extinction was slightly more adequate (Figure A.8/S12). Our approach to validate the robustness using posterior predictive simulations might proof fruitful to overcome parameter non-identifiability.

Chapter 3

LOCALLY ADAPTIVE BAYESIAN BIRTH-DEATH MODEL SUCCESSFULLY DETECTS SLOW AND RAPID RATE SHIFTS

3.1 Abstract

Birth-death processes have given biologists a model-based framework to answer questions about changes in the birth and death rates of lineages in a phylogenetic tree. Therefore birth-death models are central to macroevolutionary as well as phylodynamic analyses. Early approaches to studying temporal variation in birth and death rates using birth-death models faced difficulties due to the restrictive choices of birth and death rate curves through time. Sufficiently flexible time-varying birth-death models are still lacking. We use a piecewise-constant birth-death model, combined with both Gaussian Markov random field (GMRF) and horseshoe Markov random field (HSMRF) prior distributions, to approximate arbitrary changes in birth rate through time. We implement these models in the widely used statistical phylogenetic software platform `RevBayes`, allowing us to jointly estimate birth-death process parameters, phylogeny, and nuisance parameters in a Bayesian framework. We test both GMRF-based and HSMRF-based models on a variety of simulated diversification scenarios, and then apply them to both a macroevolutionary and an epidemiological dataset. We find that both models are capable of inferring variable birth rates and correctly rejecting variable models in favor of effectively constant models. In general the HSMRF-based model has higher precision than its GMRF counterpart, with little to no loss of accuracy. Applied to a macroevolutionary dataset of the Australian gecko family Pygopodidae (where birth rates are interpretable as speciation rates), the GMRF-based model detects a slow decrease whereas the HSMRF-based model detects a rapid speciation-rate decrease in the last 12 million years.

Applied to an infectious disease phylodynamic dataset of sequences from HIV subtype A in Russia and Ukraine (where birth rates are interpretable as the rate of accumulation of new infections), our models detect a strongly elevated rate of infection in the 1990s.

3.2 Introduction

Studying variation in the rates of speciation and extinction enables researchers to examine the patterns and processes that shape the diversity of life on earth. Birth-death processes have given biologists a model-based framework in which questions about the birth rate, death rate, or net diversification (birth minus death) rate of species can be studied [117]. For example, the question, "are nectar spurs a key innovation in plant evolution leading to a rapid radiation?" can be rephrased as, "is the rate of diversification in plant lineages correlated with the presence of nectar spurs?" and this hypothesized association can be tested statistically. In infectious disease phylodynamics, the question "was this intervention effective in containing disease spread?" can be rephrased as, "after the intervention, did the birth rate (effective reproduction number) decrease?" In general, causation is difficult to establish, but the presence of correlation can lend support to hypotheses regarding causes of diversification. Questions involving variation in diversification rates can generally be broken down into two categories. The first class of questions, including the question about nectar spurs, concerns variation in diversification rates across lineages. In these scenarios, models are built that allow the birth and death rates to vary across the branches of the phylogenetic tree [88, 1]. The second class of questions, including the question about intervention efficacy, concerns temporal variation in diversification rates shared by all lineages [68, 141, 9]. In these scenarios, the birth and death rates are modeled as functions of time, but at any instant in time all branches of the tree share a common birth rate and a death rate. This second class of questions and models is our focus in this chapter. Our aim is to develop flexible Bayesian nonparametric methods for accurately estimating changes of birth and death rates over time without sacrificing precision.

Birth-death models [72, 105] define a probability distribution on time-calibrated

phylogenies—phylogenetic trees where branch lengths are measured in time rather than in evolutionary distances. Early approaches to inferring variability of birth and/or death rates required the use of a time-calibrated phylogeny as data. This involved estimating parameters of birth-death models and then either statistically testing for violations of constant birth and death rates [115] or choosing the best functional form (*e.g.*, two-piece piecewise constant or exponential curves) for birth and death rate trajectories from a set of candidate models via likelihood-ratio tests or the AIC [109, 120]. These early methods had the downside of not accounting explicitly for missing taxa, requiring the use of Monte Carlo simulation in order to determine if the rejection of a constant-rate (or other) model in favor of a more complex model was an artifact of incomplete sampling of phylogenetic lineages [65, 26, 59]. However, the underlying theory and likelihood function for arbitrary functions of birth and death rates including unsampled taxa was already introduced by Nee *et al.* (1994) [105]. Later, the introduction of the piecewise-constant, or episodic, birth-death model (EBD) [136] enabled biologists to perform likelihood-based comparison of birth and death rates' functional forms while accounting for incomplete taxon sampling (see Höhna (2015) for a review of the EBD and comparison to the work by Nee *et al.* (1994) [60, 105]). The EBD model was extended to work in contexts with serial samples (*e.g.*, fossils) and possibly sampled ancestors [140, 49].

The EBD model divides time into a finite number of intervals and assigns each interval its own set of birth and death rates. The first uses of the EBD model assumed that *a priori* birth and death rates in each interval are independent and identically distributed (iid) [140, 49]. This assumption means that the number of intervals (or epochs) needs to be kept small to keep estimation reasonably precise and to avoid overfitting. Further work on Bayesian modeling using the EBD employed temporally-autocorrelated models derived from discretizing Ornstein-Uhlenbeck and Brownian motion processes [33, 21, 132], which provides smoothing and allows the number of episodic intervals to be larger. May *et al.* (2016) propose another EBD model, where birth and death rates change at an unknown, Poisson distributed number of change-points [95]. Wu (2014) uses a similar change-point model [164]. These random change-point models drastically increase the dimensionality of

the parameter space and make it variable, requiring complicated reversible-jump Markov chain Monte Carlo (MCMC) [54] algorithms to sample from the posterior distribution of the number of change-points. However, many other Bayesian nonparametric approaches for estimating functional forms have not been applied to EBD modeling.

Parametric and nonparametric estimation of functional forms is not unique to birth-death processes. For example, population genetics researchers have developed a rich toolbox of Bayesian nonparametric approaches to estimate changes of the effective population size in a neighboring class of coalescent models [73]. In fact, EBD models closely resemble piecewise constant effective population size coalescent models [30, 97, 53]. However, EBD models still lack Bayesian regularization approaches that control the potentially high number of model parameters. For coalescent models, such Bayesian regularization is accomplished by Gaussian Markov random field (GMRF) prior distributions, which underly the skyride [97] and skygrid [53] methods, and by their recently developed analog, the horseshoe Markov random field (HSMRF) [43]. These models provide a rich framework for building more complicated models with covariates [52] and are amenable to computationally efficient MCMC sampling techniques. Our goal is to bring GMRF and HSMRF prior distributions to EBD models and to test their performance.

We implement birth-death models that use GMRF and HSMRF prior distributions for the birth and/or death rates in the statistical phylogenetic software platform `RevBayes` [63]. This implementation allows us to jointly estimate birth-death parameters, phylogeny, and other (nuisance) parameters in a Bayesian framework. We develop an efficient, tuning-parameter-free MCMC algorithm for sampling high dimensional parameter vectors associated with GMRF- and HSMRF-based models. We also devise a framework for setting the global scale parameter—the key parameter controlling the degree of parameter regularization (also called shrinkage)—for both models in terms of the implied prior on the number of "effective" rate shifts. We note that our GMRF-based model is closely related to the work of Duplessis (2016), Condamine *et al.* (2018), and Silvestro *et al.* (2019), who use prior distributions that fall into the class of GMRF distributions, but our work differs from

these approaches in important computational and statistical details [33, 21, 132]. Namely, we develop a tuning-parameter free MCMC algorithm that enables efficient exploration of the high dimensional parameter vectors associated with GMRF- and HSMRF-based models and introduce a framework for setting the key hyperprior in an interpretable manner. To the best of our knowledge, this is the first instance of applying HSMRF prior distributions to birth-death processes. We test both GMRF-based and HSMRF-based models on a variety of simulated diversification scenarios, and then apply them to a species-level and an epidemiological dataset. We find that both models are capable of inferring variable diversification rates and correctly rejecting variable models in favor of effectively constant models. In general, in line with previous analyses of HSMRF prior distributions [44, 43], we see that the HSMRF-based model has higher precision than its GMRF counterpart, with little to no loss of accuracy. In empirical applications, we show that these models are useful for detecting a speciation-rate decline in the Australian gecko clade Pygopodidae and a complex pattern of variation in the rate of infection of HIV subtype A in Russia and Ukraine.

3.3 Methods

Our data, $\mathbf{y}$, take the form of a multiple sequence alignment. We assume that the alignment $\mathbf{y}$ has come from the following probabilistic model. First, a tree is generated from a time-varying birth-death process governed by time varying birth rate $\lambda(t)$, death rate $\mu(t)$, serial sampling rate $\phi(t)$, conditional probability of death upon sampling $r(t)$ (primarily for phylodynamic applications to represent becoming noninfectious when diagnosed and/or treated), and vector of sampling probabilities $\mathbf{\Phi}$ (with associated sampling times $\boldsymbol{t}_{\mathbf{\Phi}}$, we refer to these as event sampling times). Time starts at 0 at the most recent event (or serial) sampling time and increases into the past, such that the oldest bifurcation in the tree is t_o, the time of origin (here also the time of the most recent common ancestor) [105]. We call the resulting reconstructed tree T, and it consists only of lineages whose descendants were sampled. For more details on this model, including helpful figures and derivations, see Gavryushkina *et al.* (2014) [49]. On each branch of T, evolution proceeds at a rate gov-

erned by a molecular clock model [167, 147, 29]. Columns in the sequence alignment evolve independently under a continuous-time Markov chain (CTMC) model, which commonly is referred to as the substitution model. We use the generalized time reversible substitution model [145] with discretized gamma-distributed rate variation across sites (GTR+G) [165]. For notational simplicity we refer to the vector of substitution and clock model parameters as $\boldsymbol{\theta}$, and we discuss the specifics of these models on a case-by-case basis. We can write the phylogenetic likelihood—probability of the alignment under the CTMC substitution model—as $\Pr(\mathbf{y} \mid \boldsymbol{\theta}, T)$. All major statistical phylogenetic software platforms can efficiently compute a phylogenetic likelihood via a dynamic programming algorithm, known as the Felsenstein pruning algorithm [45]. We will use the `RevBayes` implementation of this algorithm [63].

In Bayesian inference we need prior distributions for $\lambda(t)$, $\mu(t)$, $\phi(t)$, $r(t)$, $\boldsymbol{\Phi}$, and t_o, as well as prior distributions on $\boldsymbol{\theta}$. We assume that $\boldsymbol{t_\Phi}$ is fixed *a priori* by the user. For our purposes there will only be one time at which event sampling may occur: the present day, making $\boldsymbol{t_\Phi}$ a scalar $t_\Phi = 0$. Notice that the prior on T, conditional on $\lambda(t)$, $\mu(t)$, $\phi(t)$, $r(t)$, $\boldsymbol{\Phi}$, and t_o, is already specified by the birth-death process. The choice of $\Pr(t_o)$ depends on the particular group of taxa studied, and the form of $\Pr(\boldsymbol{\theta})$ on the specifics of the group and the data, so we discuss these on a case-by-case basis. The posterior distribution takes the following form:

$$\begin{aligned}\Pr(\boldsymbol{\theta}, T, t_o, \lambda(t), \mu(t), \phi(t), r(t), \boldsymbol{\Phi} \mid \mathbf{y}) \propto\ & \Pr(\mathbf{y} \mid \boldsymbol{\theta}, T)\ \Pr(\boldsymbol{\theta})\ \Pr(t_o) \\ & \times\ \Pr(T \mid \lambda(t), \mu(t), \phi(t), r(t), \boldsymbol{\Phi}, \boldsymbol{t_\Phi}, t_o) \\ & \times\ \Pr(\lambda(t))\ \Pr(\mu(t))\ \Pr(\phi(t))\ \Pr(r(t))\ \Pr(\boldsymbol{\Phi}).\end{aligned}$$

In macroevolutionary analyses including extant species, there is a single event sampling at the present ($t = 0$) with known probability Φ_0 [166]. In phylodynamic analyses, there may be no event sampling, thus we set $\Phi_0 = 0$. We make the simplifying assumptions that the serial sampling rate is a constant, $\phi(t) = \phi$, and that the conditional probability of becoming

noninfectious upon sampling is a known constant, $r(t) = r$. For any macroevolutionary dataset $r = 0$, and in our phylodynamic application we assume $r = 1$. Additionally, in our macroevolutionary example there are no serial samples, hence $\phi = 0$ (in which case r is not a parameter of the simplified model). In all analyses we make the additional simplifying assumption that the death rate is a constant $\mu(t) = \mu$, and place a mean-0.1 exponential prior on μ. In phylodynamic applications, there is often prior information that enables the use of informative prior distributions on ϕ and μ, which we discuss in a later section. The remaining piece of the puzzle, and our contribution in this chapter, is in the specification of $\Pr(\lambda(t))$, for which we use Markov random field models. (Note that our implementation and theory of the GMRF and HSMRF can be applied to all time-varying rates; we focused on the birth rate only for simplicity.) Our simplified posterior distribution takes the following form:

$$\begin{aligned}\Pr(\boldsymbol{\theta}, T, t_o, \lambda(t), \mu, \phi, r, \mid \mathbf{y}) \propto\ & \Pr(\mathbf{y} \mid \boldsymbol{\theta}, T)\ \Pr(\boldsymbol{\theta}) \\ & \times\ \Pr(T \mid \lambda(t), \mu, \phi, r, \Phi_0, t_o) \\ & \times\ \Pr(\lambda(t))\ \Pr(\mu)\ \Pr(\phi)\ \Pr(r)\ \Pr(t_o).\end{aligned}$$

We note that historically Φ_0 has been called ρ, and $\boldsymbol{\Phi}$ has sometimes been called $\boldsymbol{\rho}$. However, $\boldsymbol{\rho}$ has been used to refer to both sampling probabilities [49] and mass extinction probabilities [136, 60, 95], which creates room for confusion.

3.3.1 Horseshoe Markov random field prior

We define the birth rate on the log scale, $\lambda^*(t) = \ln(\lambda(t))$. Following Stadler (2011), we discretize time into n intervals and assume that $\lambda^*(t) = \lambda_i^*$ when t is in the ith time interval, using the parameterization $\boldsymbol{\lambda^*} = (\lambda_1^*, ..., \lambda_n^*)$ [136]. We find $n = 100$ works well in practice and refer readers to S1 Text for a more detailed discussion of the grid size n. An HSMRF is a model in which $\lambda_{i+1}^* \mid (\lambda_i^*, \gamma) \sim \text{Horseshoe}(\lambda_i^*, \gamma)$, where γ is a global scale parameter that controls the smoothness of the overall field. The horseshoe is a distribution used as a

shrinkage prior, a statistical tool designed to discern signal from noise [17]. In our case, the HSMRF exerts strong prior belief that $\lambda^*_{i+1} \approx \lambda^*_i$; in other words, we do not expect much change in the birth rate between adjacent intervals. However, the horseshoe distribution also has fat (Cauchy-like) tails that allow the HSMRF to behave like a spike-and-slab mixture model [149], giving the HSMRF a property known as local adaptivity. The horseshoe distribution employs an auxiliary variable σ_i and is represented as a scale mixture of normal distributions

$$\sigma_i \sim \text{halfCauchy}(0, 1),$$
$$\lambda^*_{i+1} \mid \sigma_i, \gamma \sim \text{Normal}(\lambda^*_i, \sigma_i^2\gamma^2).$$

This mixture representation helps explain the local adaptivity of the HSMRF: one or a few (relatively) large changes can be handled by large σ_i without increasing γ. We place a Normal($\ln(\hat{\lambda}), \xi$) prior distribution on λ^*_1, where $\hat{\lambda}$ is a rough estimate of net diversification rate. When there are extant lineages in the tree, $\hat{\lambda}$ is the maximum likelihood estimator for the net diversification rate, d, from Magallon and Sanderson (2001) [89]. When there are no extant lineages in the tree, we put a lower bound on the net diversification rate using the number of births in the tree (excluding the origin or root as appropriate), B_{obs}. The expected net number of births observed in a tree by time t is given by $\mathbb{E}(B_t) = 2e^{t \cdot d} - 2$ if starting at the time of the most recent common ancestor, and $\mathbb{E}(B_t) = e^{t \cdot d} - 1$ if starting with a single lineage. By the method of moments, we can obtain (for the case of starting with the MRCA) $\hat{d} = (\ln(B_{obs} + 2) - \ln(2))/t$, where t is the age of the tree. As not all lineages that are born will be sampled in our tree, the number of observed births will be an underestimate of the number of births and our rate will be underestimated, but it will suffice. In practice, when setting the prior for the first birth rate, we use $\xi = 1.17481$, producing, *a priori*, $\Pr(\hat{\lambda}/10 \leq \lambda_1 < 10\hat{\lambda}) = 0.95$. We use a halfCauchy(0, ζ) prior on γ, where ζ is the global shrinkage prior, and we discuss how to set it in a later section. A list of all parameters in the model and their prior distributions can be found in Table 3.1. The full

posterior distribution of our HSMRF-based model parameters is,

$$\Pr(T, t_o, \boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mu, \gamma, \boldsymbol{\sigma} \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \boldsymbol{\theta}, T)\ \Pr(\boldsymbol{\theta})\ \Pr(T \mid \boldsymbol{\lambda}^*, \mu, t_o, \rho)\ \Pr(\boldsymbol{\lambda}^* \mid \gamma, \boldsymbol{\sigma}) \times\ \Pr(\boldsymbol{\sigma})\ \Pr(\gamma)\ \Pr(\mu)\ \Pr(t_o).$$

We approximate the above posterior distribution using the following MCMC strategy. We use standard Metropolis-Hastings kernels available in `RevBayes` to update the tree T, time of the root t_o, substitution model parameters $\boldsymbol{\theta}$, and extinction rate μ, and the first log-speciation rate, λ_1^* (see Höhna *et al.* (2017) for a description of the standard `RevBayes` Metropolis-Hastings kernels [62]). Since vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\sigma}$ can be high dimensional, we update the vectors in blocks. First, we reparameterize the model to work with the first order differences $\boldsymbol{\Delta}$ instead of $\boldsymbol{\lambda}^*$ (see Figure 3.1). This allows us to sample the vector $\boldsymbol{\Delta}$, where all elements are *a priori* independent, instead of the vector $\boldsymbol{\lambda}$ where the adjacent values are highly correlated, greatly increasing the efficiency of the MCMC. Further, under the GMRF and the hierarchical representation of the HSMRF, all the $\boldsymbol{\Delta}$ are Normal random variables, enabling us to employ an elliptical slice sampler [102] for $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_{n-1})$. The (conditional) normality of the $\boldsymbol{\Delta}$ also allows us to employ a Gibbs sampler for γ and $\boldsymbol{\sigma}$, which allows us to adequately sample the tails of the posterior distribution. Without this elliptical slice sampler and Gibbs sampler combination, MCMC for these models fails to converge to the posterior distribution. We defer a more thorough discussion of our MCMC strategy to S1 Text. We also note that there are directionless specifications of MRF models which make it implicit that information is shared across adjacent intervals in both directions (that is that the full conditional distribution of λ_i^* depends on both λ_{i+1}^* and λ_{i-1}^*). For more on this subject, we direct readers to Rue and Held (2005) [126].

3.3.2 Gaussian Markov random field prior

Our GMRF-based model can be seen as a special case of an HSMRF-based model where $\sigma_1 = \cdots = \sigma_n = 1$, meaning $\lambda_{i+1}^* \mid \lambda_i^*, \gamma \sim \text{Normal}(\lambda_i^*, \gamma^2)$. The lack of local scale parameters

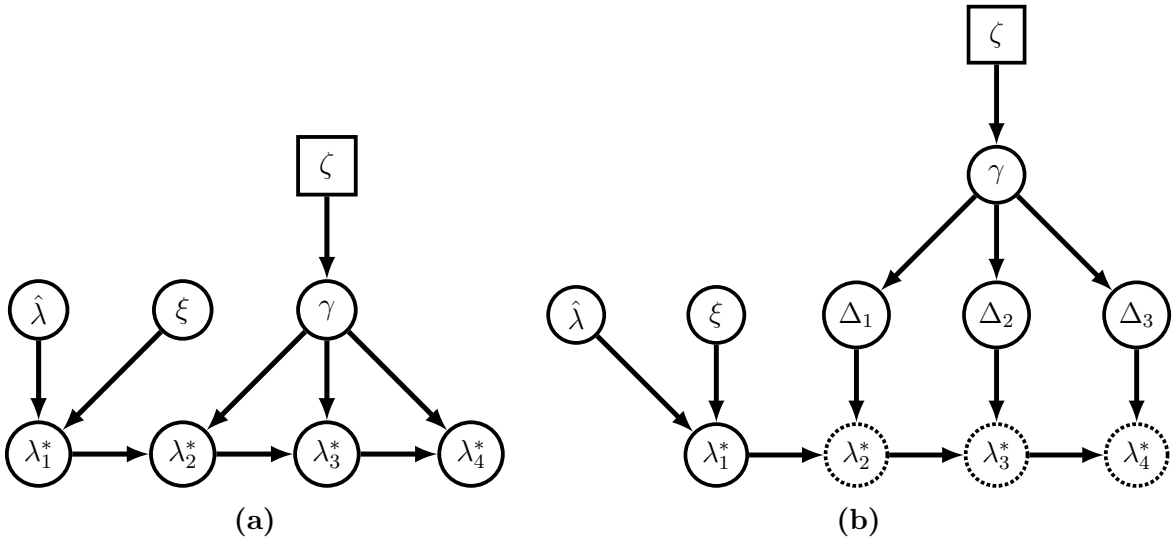


Figure 3.1: Simplified versions of our MRF-based models, shown as a grid of size 4. To highlight the structural similarities between the GMRF- and HSMRF-based models, we draw the directed acyclic graph (DAG) as if we had an analytical form of the horseshoe distribution (that is, we omit the local scale parameters of the HSMRF). In (a), we show the idealized general MRF model, while in (b), we show how we can reparameterize the model in terms of a vector $\boldsymbol{\Delta}$ of independent random variables. From $\boldsymbol{\Delta}$, we can recover $\boldsymbol{\lambda}^*$ as $\lambda^*_{i+1} = \lambda^*_i + \Delta_i$, $i = 1, \ldots, n-1$. This reparameterization greatly improves the efficiency of MCMC sampling. When drawing the model as a DAG, squares represent constant values, closed circles stochastic values, and open circles deterministic transformations of other nodes.

makes the GMRF-based model a non-locally-adaptive model. For the GMRF-based model, the posterior distribution is

$$\begin{aligned}\Pr(T, t_o, \boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mu, \gamma, \mid \mathbf{y}) \propto\ & \Pr(\mathbf{y} \mid \boldsymbol{\theta}, T)\ \Pr(\boldsymbol{\theta})\ \Pr(T \mid \boldsymbol{\lambda}^*, \mu, t_o, \rho)\ \Pr(\boldsymbol{\lambda}^* \mid \gamma) \\ & \times \Pr(\gamma)\ \Pr(\mu)\ \Pr(t_o).\end{aligned}$$

The MCMC algorithm to approximate the above posterior distribution is the same as for the HSMRF-based model, except we do not need to update the vector $\boldsymbol{\sigma}$.

3.3.3 Setting the prior on the global scale parameter

In both HSMRF- and GMRF-based models, the global scale parameter, γ, controls the smoothness of the overall field, with smaller values favoring less variability. Following

Faulkner *et al.*, we take a hierarchical approach and place a prior distribution on the global scale parameter, such that $\gamma \sim \text{halfCauchy}(0, \zeta)$ [43]. We choose ζ in terms of s_e — the number of "effective shifts" in the birth rate and we define an effective shift to be an event $\{\lambda_{i+1}/\lambda_i < 1/\epsilon \text{ or } \lambda_{i+1}/\lambda_i \geq \epsilon\}$. That is, an effective shift is the event where two adjacent birth rates are different by more than ϵ-fold. We set $\epsilon = 2$, reflecting that a 2-fold change in the birth rate is biologically meaningful and statistically detectable. Setting ζ is then accomplished implicitly by setting the prior expected number of effective shifts, $\mathbb{E}[s_e]$, which is more interpretable than ζ. In this setup, $\mathbb{E}[s_e]$ is the expectation of a binomial random variable with probability p that there is an effective shift between λ_{i+1} and λ_i. Since we can compute p given a particular value of ζ, and since there is no obvious closed form solution, we use numerical methods to solve for ζ. Code to calculate ζ from $\mathbb{E}[s_e]$ is available at github.com/afmagee/hsmrfbdp.

We find that in practice $\mathbb{E}[s_e] = \ln(2)$ produces a prior that is reasonably conservative yet flexible. This yields $\zeta_{HSMRF} \approx 0.0021$ for HSMRF-based models and $\zeta_{GMRF} \approx 0.0094$ for GMRF-based models. An alternative approach to specify ζ examines the implied prior distribution on λ_n/λ_1, *i.e.*, the prior distribution on the fold change across the entire process. *A priori*, for the HSMRF on a grid size $n = 100$, $\mathbb{E}[s_e] = \ln(2)$ leads to $\Pr(0.5 \leq \lambda_n/\lambda_1 < 2) \approx 0.76$ and $\Pr(0.1 \leq \lambda_n/\lambda_1 < 10) \approx 0.9$. While we do not use this approach to set ζ, it shows that our chosen value for ζ focuses the prior mass on reasonable regimes while leaving room for rather substantial amounts of change. For completion, in S1 Text we provide more context for these choices of prior, including an alternative choice of $\mathbb{E}[s_e]$ following Drummond and Suchard (2010), and examine two additional frameworks, bounding the marginal variance of the GMRF and HSMRF (explored by Sørbye and Rue (2014) and Faulkner *et al.* (2018)), and bounding the effective number of parameters in the model (explored by Piironen and Vehtari (2017)) [31, 133, 43, 111].

Table 3.1: Model parameters, their prior distributions, and their role in the model. In most of our analyses, we assume a constant death rate, μ, with an Exponential prior with rate parameter $\eta = 10$. In phylodynamic applications, there may be more information to set η, while in macroevolutionary examples one could instead employ an empirical Bayes approach. When there is serial sampling, we adopt an empirical Bayes approach to setting the prior on the sampling rate, ϕ, using a guess at the tree age and the number of tips to obtain $\hat{\phi}$. In practice we set $\omega = 1.17481$. In analyses without serial sampling, $\phi = 0$. For details on computing $\hat{\phi}$, see S1 Text. The sampling fraction at present, Φ_0 and the probability of death upon sampling, r are taken to be known *a priori*. The age of the tree, t_{or} is fixed to the observed height if the tree is data, else it is a variable with the prior determined by the user. For models with $n = 100$ intervals, we set $\zeta = 0.0021$ for HSMRF-based models and $\zeta = 0.0094$ for GMRF-based models, while for models with other n, we provide code for setting ζ. The GMRF-based model lacks local scale parameters $\boldsymbol{\sigma}$. We adopt an empirical Bayes approach to setting the prior on the first log-birth-rate using a guess at the tree age and the number of tips to obtain $\hat{\lambda}_1^*$. In practice we set $\xi = 1.17481$. In models where the death rate varies, the previously discussed prior on μ serves as the prior on μ_1, and the rest of the prior is accomplished via an MRF model exactly as with the birth rate.

Parameter	Prior	Role
μ	Exponential(η)	death rate
ϕ	Lognormal($\hat{\phi}$, ω)	serial sampling rate
Φ_0	Fixed	sampling fraction at present
r	Fixed	probability of death upon sampling
t_{or}	User choice	age of tree
ζ	Fixed	global scale hyperparameter
γ	halfCauchy(0,ζ)	global scale of the MRF
σ_i	halfCauchy(0,1)	local scale of HSMRF
λ_1^*	Normal($\ln(\hat{\lambda}), \xi$)	log-scale birth rate at present
$\lambda_{i>2}^*$	Normal(λ_{i-1}^*,$\gamma^2\sigma_i^2$)	log-scale birth rates

3.4 Results

3.4.1 Simulation study

To understand statistical properties of both random field birth-death models, we perform a (nonexhaustive) simulation study. Some of the most debated questions in species diversification concern diversification-rate decreases [130, 82, 115, 110, 37, 98, 38], and the ability to detect effective epidemiological interventions hinges on the ability to accurately estimate decreases in the rate of infection, so we consider simulation scenarios where the birth-rate declines through time. We devise a series of piecewise-linear functions $\lambda(t)$ in which the birth rate decreases through time. For each model, we use the R package `TESS` [64] to simulate

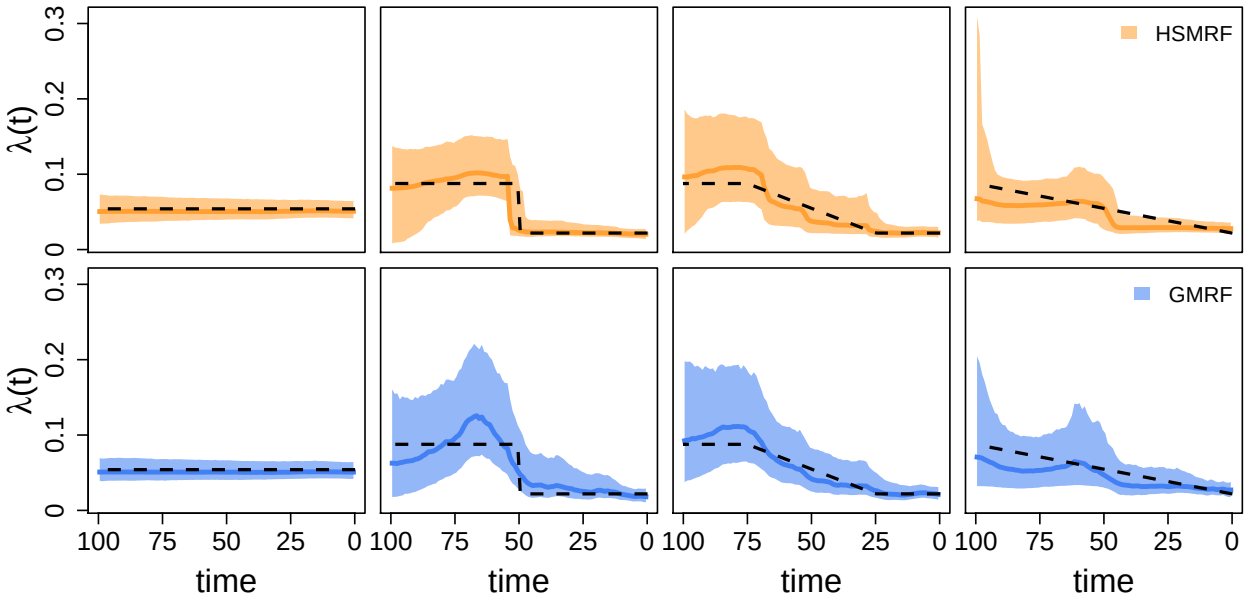


Figure 3.2: Inferred birth-rate trajectories from four individual simulations. The dashed line is the true simulating birth rate, the dark colored line is the posterior median trajectory (the median is taken separately for each grid cell), and the shaded region show the 90% Credible Interval (CI) for the rate. The leftmost column is from the constant-rate simulations, and the right three columns demonstrate the effect of changing the shift duration (the length of the tree over which the birth rate changes), from an instantaneous shift to a constant change model. When we focus instead on the location of the shift, all simulations are piecewise-constant as in the second column. In each column, we show the simulation with the most average performance measured in terms of the Mean Absolute Deviation of both the GMRF and HSMRF.

100 trees conditioned on the tree age ($t_o = 100$), with complete species sampling ($\Phi_0 = 1$), and choosing values for $\lambda(t)$ and μ to give an expectation of 200 species/tips at the present. Given the underdeveloped infrastructure for simulating serially-sampled trees, we focus on trees where all samples are at the present day ($\phi = 0$), but see Barido-Sottani *et al.* (2019) for recent developments [3]. When analyzing these simulations, we take the tree and sampling fraction to be known. Treating the tree as data allows us to focus on the performance of the random field birth-death models without worrying about potential sources of bias during time-calibrated tree estimation [15]. Taking the tree as data also mirrors the predominant historical usage of models of rate variation, detecting variation in trees previously inferred [120, 88, 1, 136].

We assess model performance by looking at four summaries of the inferred birth-rate trajectories. We take as our estimate of $\lambda(t)$ the birth-rate trajectory defined by the median posterior of each birth rate λ_i. First, to quantify bias we use the Mean Absolute Deviation (MAD) of the estimated birth-rate trajectory, given by $(1/n)\sum_{i=1}^{n}|\hat{\lambda}_i - \lambda_i|$. Second, we look at the Mean Absolute Sequential Variation (MASV) of the estimated trajectory, the gross change inferred, given by $(1/(n-1))\sum_{i=1}^{n-1}|\hat{\lambda}_{i+1} - \hat{\lambda}_i|$. Where the simulated trajectory is variable, it is more useful to consider the relative MASV (RMASV), MASV($\hat{\lambda}$)/MASV(λ). If RMASV > 1, the inferred trajectory is more variable than the true trajectory, and if RMASV < 1, it is less variable. Third, we look at the fold change (FC) of the estimated trajectory, $\hat{\lambda}_n/\hat{\lambda}_1$. This will show us if we capture the presence of an overall change in the birth rate, even if we fail to capture the specific pattern. Finally, we look at the average width (across all estimated birth rates, λ_i) of the 90% posterior credible interval as a measure of precision, $(1/n)\sum_{1}^{n}(\hat{\lambda}_i^{0.95} - \hat{\lambda}_i^{0.05})/\hat{\lambda}_i$. This measure, which we call relative precision (RP), is both more interpretable than the raw credible interval and more comparable across simulations.

Constant-rate simulations

Our first simulations are from a constant-rate diversification model, such that $\lambda(t) = \lambda$ (e.g., the first column in Figure 3.2). This allows us to test the tendency towards what could be termed "false positives," the detection of spurious rate variation. Both the GMRF and HSMRF birth-death models can produce effectively constant-rate trajectories, though their flexibility is not without minor drawbacks. Ridge plots of performance measure histograms across all simulations are shown in Figure 3.3. The trajectories estimated by both models have low MAD performance measure, FC ≈ 1, and small RP performance measure, indicating generally good performance. Further, compared to fitting constant-rate models, the increase in the MAD performance measure from inferring the variable-rate models is negligible, and the decrease in precision is small (S1 Text Figure A). Thus, the primary drawback to using these models to fit constant-rate trajectories is that there are false positives. In other words, the models occasionally fit trajectories where the inferred change between the beginning and

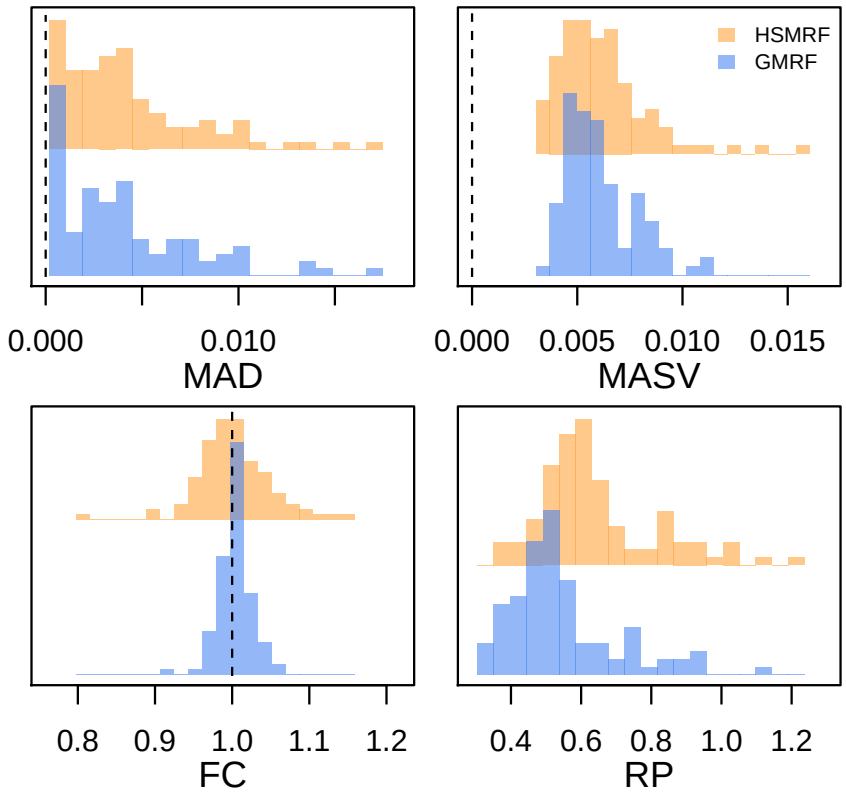


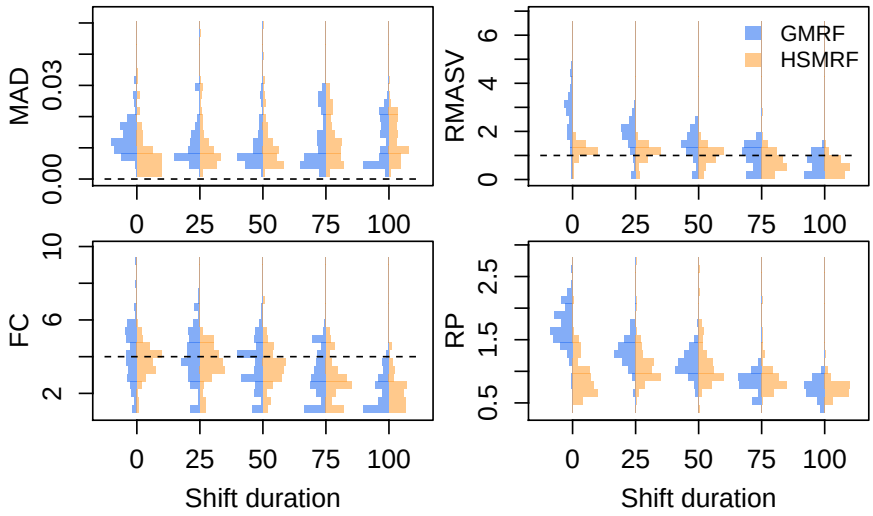
Figure 3.3: Performance of the models on simulated constant-rate datasets. MAD (Mean Absolute Deviation) measures the error in the estimated trajectory. MASV (Mean Absolute Sequential Variation) measures the total amount of change relative in the trajectory, horizontal line at true value for reference. FC (Fold Change) measures the fold change from present to past, dashed line at true value for reference. RP (Relative Precision) is a measure of precision, the average width of the 90% Credible Interval relative to the birth rate.

end of the process does not appear negligible. However, comparisons to the prior make it quite clear that both the GMRF and HSMRF are fitting effectively constant-rate trajectories. For both models roughly 99% of the prior MASV is greater than 0.01, while roughly 5% of the posterior MASV is greater than 0.01, further indicating that the models are producing effectively constant trajectories. Computing such a significance threshold to reject a constant rate model can be computed using Monte Carlo simulation [18]. The HSMRF and GMRF produce very similar average error and fold change, though the distributions of these metrics

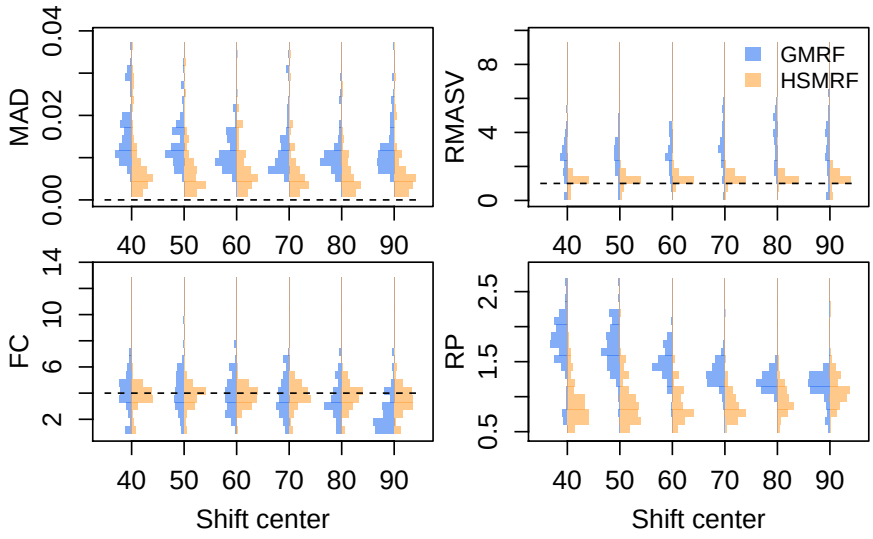
for the GMRF are more tightly focused around the target values (MAD of 0, FC of 1), and the GMRF has slightly tighter credible intervals. The GMRF generally estimates narrower credible intervals, while the HSMRF generally estimates trajectories with lower MASV.

Piecewise linear simulations

Our primary time-variable simulations examine the impact of the shift duration, *i.e.*, the amount of time over which the birth-rate changes. To examine this, we build a piecewise-linear birth-rate function, where the birth rate is λ_1 for $100 \geq t > t_1$, λ_2 for $t \leq t_2$, and a linear interpolation for $t_1 \geq t > t_2$. We center the shift at 50 ($(t_1 + t_2)/2 = 50$), and simulate shift durations ($t_2 - t_1$) of 0%, 25%, 50%, 75%, and 100% of the age of the tree. All simulation parameters are recorded in S1 Text Table A. The HSMRF-based model performs better when the shift is fast (when $t_2 - t_1$ is small), and the GMRF-based model performs better when the shift is slow (when $t_2 - t_1$ is large, Figure 3.4a). For the HSMRF-based model, the MAD performance measure is lower and both the MASV and FC performance measures are closer to the truth when the true shift is shorter. In contrast, for the GMRF-based model, the MASV performance measure gets closer to the truth, *i.e.*, the error decreases, and the RP performance measure gets narrower as the shift duration increases. In some simulations, both models effectively fit constant-rate trajectories. The HSMRF-based model also has a tendency towards fitting steep shifts even in cases where the true shift is slow (Figure 3.2). Both models though have difficulty with continuous, slow declines where they have a tendency to underestimate the total change. However, comparisons to the prior show that both models tend to detect some change. For both the HSMRF- and GMRF-based models, the median FC is approximately 2 for the continuous decline simulations (52% and 54% of FC are greater than 2, respectively). Simulations from the prior of the HSMRF-based model show that 12% of trajectories are expected to have a larger fold change than 2, while for the GMRF-based model 3% are expected to be greater than 2. Thus, compared to the prior, the posterior median trajectories have shifted to larger fold changes. Further, while the prior median fold change is 1, only 3% and 2% of FC inferred by the HSMRF- and GMRF-based



(a)



(b)

Figure 3.4: The effect on parameter inference of (a) changing the (four-fold) rate shift from instantaneous to the entire length of the trajectory and (b) changing the center of an instantaneous (four-fold) rate shift. MAD measures the error in the estimated trajectory. RMASV (Relative MASV) measures the total amount of change relative to the true MASV, horizontal line at 1 for reference. FC measures the fold change from present to past, dotted line at true value for reference. RP is a measure of precision, the average width of the 90% Credible Interval relative to the birth rate.

models are below 1. All of this indicates evidence for rate variation.

Piecewise constant simulation

We also examine the effect of varying the location of the instantaneous birth-rate shift. To do this, we build a piecewise-constant birth-rate function, where the birth rate is λ_1 for $100 \leq t < t_{shift}$, λ_2 for $t \leq t_{shift}$; we simulate $t_{shift} = 90, 80, ..., 40$ (*e.g.*, Figure 3.2 second panel). The location of the rate shift should influence the capacity for detection by altering the expected number of births in the pre-shift portion of the tree. As the shift moves from past to present, for the HSMRF-based model the MAD performance measure decreases and the RP performance measure gets smaller (Figure 3.4b). However, for the GMRF-based model, as the shift becomes more recent, the RMASV performance measure becomes increasingly inflated, indicating trajectories that are too variable. This is due to the GMRF-based model estimating rather substantial variation in the more ancient portions of the tree. The HSMRF-based model outperforms the GMRF-based model in most performance measures and for most shift locations.

Shift magnitude

The magnitude of the birth-rate shift should also impact the capacity for detection, so we simulate shifts of two magnitudes for all scenarios outlined above. For our low magnitude shift, we simulate a two-fold change, and for our larger shift, we simulate a four-fold change. Unsurprisingly, it is harder to detect smaller shifts. Results for different functional forms are qualitatively similar between shift magnitudes, so we present only the results for the four-fold shifts in Figure 3.4. In many cases with two-fold shifts, the inferred trajectory is effectively constant. Thus in general the MAD performance measure is higher, the RMASV, FC, and RP performance measures are lower. S1 Text Figures D and E give full simulation results for the two-fold case comparable to Figure 3.4.

3.4.2 Time-varying death rates

We also investigated the ability of our models to simultaneously infer both time-varying birth and death rates. We devised a piecewise-constant simulation scenario loosely based on a dataset of Hepatitis C infections in Egypt, which is often used as a benchmark for assessing phylodynamic methods [140, 43]. While the rate of recovery/removal can be relatively constant over time during spread of an infectious disease agent, this rate will vary if interventions are implemented (e.g., isolation of identified infectious individuals). Since recoveries/removals are represented by deaths in the phylodynamic birth-death models, it is of interest to be able to infer changes in the death rate over time. In the first applications of the birth-death skyline model to real data, the total death rate ($\mu(t) + r(t)\phi(t)$) was observed to vary by roughly three-fold in two different datasets [140]. In our scenario, the death rate experiences an approximately five-fold increase, while the birth rate first undergoes a five-fold increase and then a ten-fold decrease. Changes in birth and death rates are asynchronous. We simulate 100 trees with isochronous sampling ($\Phi_0 = 1$), targeting an expectation of 200 tips as in the main simulation study. We additionally simulate 100 trees with heterochronous sampling ($\phi = 0.009, r = 1$) using `TreePar` [136], producing trees with an average of 175 tips. For comparability (in terms of the expected number of extant taxa at any time in the tree), in the heterochronous simulations we adjust $\mu(t)$ such that the total death rate ($\mu(t) + \phi r$) is the same as the death rate ($\mu(t)$) in the isochronous simulations.

We analyze both tree sets with our GMRF-based and HSMRF-based models in two different analysis setups. First, we attempt to infer both $\lambda(t)$ and $\mu(t)$. Second, we misspecify the true model by inferring $\lambda(t)$ while inferring a constant μ. In Figure 3.5 we show representative estimates of birth and death rates for both models and both analysis setups. We also summarize our results using the MAD, RP, and RMASV performance measures in Figure 3.6. We do not report fold change here as it is not a useful measure here. This is because the birth-rate trajectory has more than one shift, and in some intervals $\mu(t) = 0$, making the fold change infinite (this choice of $\mu(t)$ allows us to keep the total death rate the same

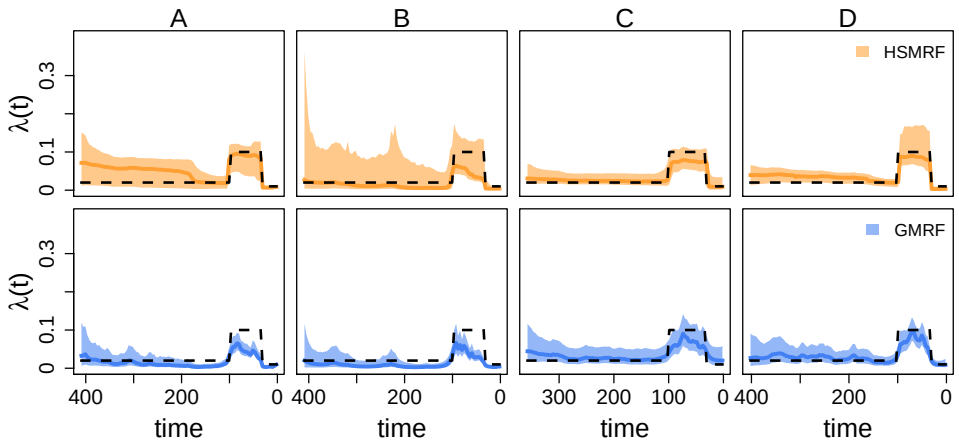
between the heterochronously sampled and isochronously sampled simulations).

First, we examine the ability of our models to infer time-varying death rates. Our simulations reveal that estimating time-varying death rates can be quite difficult, especially without serial samples. Without heterochronous sampling, there is very little signal in any analysis for time-varying death rates; the performance is essentially equivalent to fitting a constant death rate. When heterochronous samples are present, it is possible to detect time-varying death rates and estimates become more accurate. However, even with heterochronous samples both models frequently underestimate the variability of the death-rate trajectory as measured by the RMASV performance measure. The RMASV performance measure also shows that the HSMRF-based model is much better at detecting appropriate amounts of variation than the GMRF-based model.

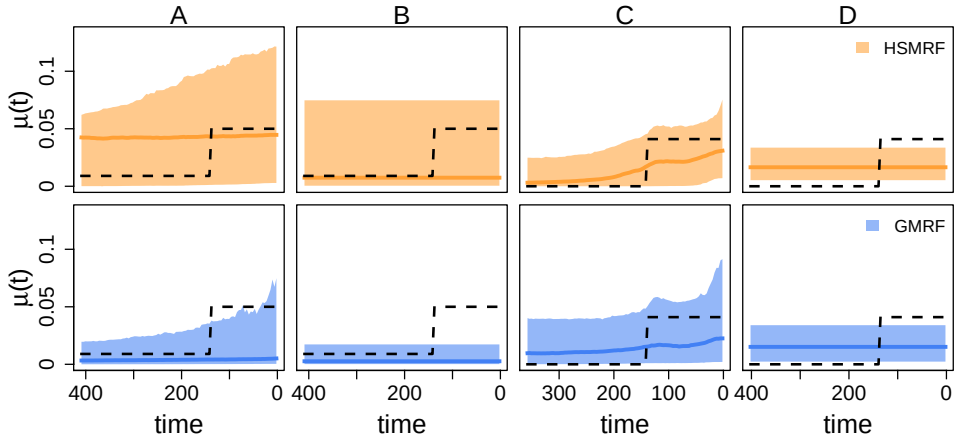
We additionally investigate the ability of our models to infer time-varying birth rates in the presence of a varying death rate, and find that estimating time-varying birth rates is more difficult when the death rate varies. The MAD performance measure is higher in the time-varying death simulations than in the main simulation study and the RP performance measure is larger. In the absence of serial samples, the RMASV performance measure shows that the amount of variability is generally underestimated, though there is still clear evidence for variation. As the trajectories have similar birth rates to the constant-death simulations, the MAD performance measures should be largely comparable, despite not being relative measures like the RP and RMASV performance measures. The presence of serial samples greatly improves estimation of the total variability of the trajectory as measured by the RMASV performance measure, and greatly improves accuracy and precision. Fitting a model with a constant death rate generally has little effect on how well the birth rate trajectory is estimated.

3.4.3 Empirical analysis of Pygopodidae

Pygopodidae is a clade of approximately 46 legless geckos [9]. Recently, Brennan *et al.* (2017) used several birth-death models to investigate the history of diversification in this group, ex-



(a)



(b)

Figure 3.5: Inferred (a) birth-rate trajectories and (b) death-rate trajectories from four individual simulations with time-varying birth and death rates. The dashed line is the true simulating rate, the dark colored line is the posterior median trajectory (the median is taken separately for each grid cell), and the shaded region show the 90% Credible Intervals (CIs) for the rate. In each column, we show the simulation with the most average performance measured in terms of the Mean Absolute Deviation of the birth- and death-rate trajectories from both the GMRF and HSMRF (columns are shared across birth-rate and death-rate subfigures). The column labels A, B, C, and D identify the different combinations of tree simulations and analysis setup. A and B are analyses of trees with isochronous sampling, C and D heterochronous sampling. A and C are analyses where time-varying death rate, $\mu(t)$ is inferred, B and D where a constant death rate, μ, is inferred.

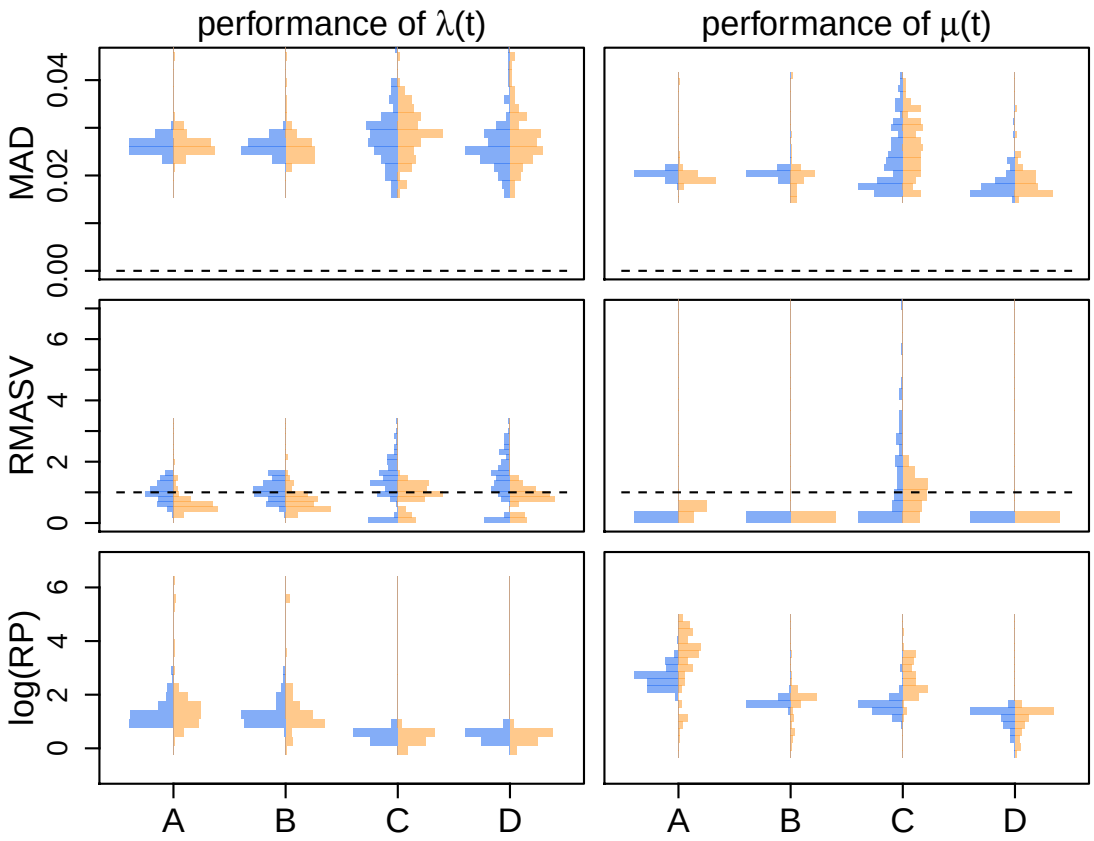


Figure 3.6: Performance of the models on simulated datasets where both the birth- and death-rate trajectories. MAD measures the error in the estimated rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the true rate. RMASV measures the total amount of change relative to the true MASV, horizontal line at 1 for reference. The column labels A, B, C, and D identify the different combinations of tree simulations and analysis setup. A and B are analyses of trees with isochronous sampling, C and D heterochronous sampling. A and C are analyses where time-varying death rate, $\mu(t)$ is inferred, B and D where a constant death rate, μ, is inferred.

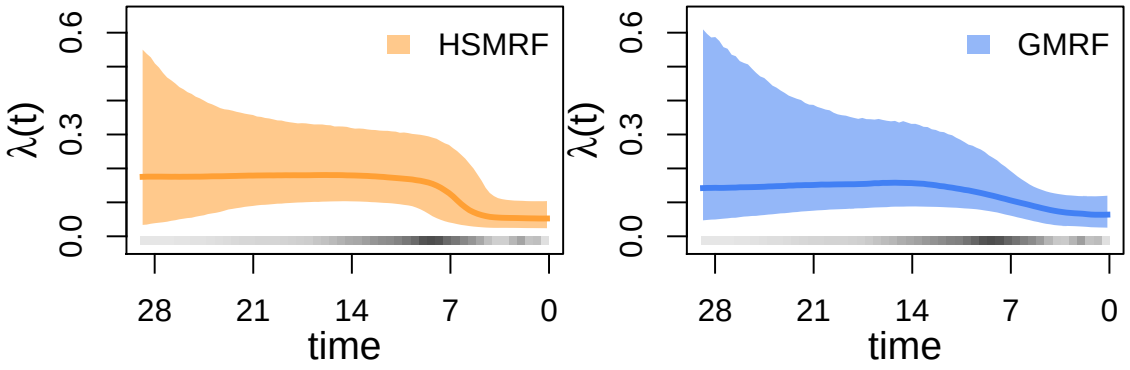


Figure 3.7: Analyses of the Pygopodidae dataset. Plotted are posterior median trajectories (dark lines) and 90% credible intervals (shaded regions). Time is in millions of years before the present day. In grey is a heatmap of the inferred divergence times.

amining trends in speciation over time using a posterior sample of 100 phylogenies estimated via BEAST 1.8.3 [9, 32]. The majority of their analyses revealed a drastic speciation-rate decrease in the recent (2-5 million years) past, though there was some disagreement between methods over the significance and timing of the shift. Here we revisit the question of significance and timing of the birth-rate shift in full joint analyses of phylogeny and both our GMRF and HSMRF birth-death models from molecular sequence data. In these analyses, we assume a constant death rate, μ. Details of the substitution and clock models are available in S1 Text, as are details of MCMC convergence diagnostics performed.

Our dataset includes 41 out of 46 representatives of Pygopodidae, which we use to set the species sampling fraction, $\Phi_0 = 0.89$. We employ calibrations on the same nodes as Brennan *et al.* (2017), resulting in a calibration for the root node and one each on the genera *Delma* and *Apprasia* [9]. Following Brennan *et al.* (2017), we place a Uniform(19.5, 29.0) prior on the root age [9]. To set up our grid, we thus choose to divide the interval $[0, 29]$ into 100 intervals/epochs of equal length.

GMRF- and HMRF-based models produce a clear visual signature of a diversification-rate decrease (Figure 3.7), with a higher rate from the origin of the clade up until at least 12 Ma, and a lower rate afterwards. The HSMRF-based model favors a steeper decrease ending

approximately 6 Ma, while the GMRF-based model favors a much slower decline that starts approximately 14 Ma and lasts until approximately 2 Ma. Over the range [2 Ma, 12 Ma], the HSMRF estimates a 3.43-fold decrease (90% Credible Interval (CI) [1.12, 8.49]), while the GMRF estimates a 2.41-fold decrease (90% CI [1.00, 7.58]). The HSMRF produces 90% credible intervals for the speciation rate that are generally narrower than the GMRF-based model's intervals. The behavior of both models is in line with the simulation results for fast to intermediate shifts, with the HSMRF inferring a faster shift of larger magnitude with tighter credible intervals than the GMRF-based model.

Given that the posterior distributions of adjacent birth rates are highly correlated, testing for a shift in a specific interval from λ_i to λ_{i+1} could suggest there is no shift even when there is clearly a shift present in the overall trajectory. However, we can avoid this issue by testing hypotheses over longer timespans. The Bayes factor [71] in support of an s-fold decrease between t_{start} and t_{end} is given by,

$$\frac{\Pr(\lambda(t_{start})/\lambda(t_{end}) < s \mid \mathbf{y})}{\Pr(\lambda(t_{start})/\lambda(t_{end}) \geq s \mid \mathbf{y})} \Big/ \frac{\Pr(\lambda(t_{start})/\lambda(t_{end}) < s)}{\Pr(\lambda(t_{start})/\lambda(t_{end}) \geq s)}.$$

For an s-fold increase, the inequalities are reversed. If we are interested in the evidence of a shift over the range [2 Ma, 12 Ma], we can compare the speciation rates in the appropriate intervals for a given shift size s. For our grid, the 7th interval ends at 2.03 Ma, while the 43rd starts at 12.18 Ma, and we would test hypotheses regarding λ_7/λ_{43}. Then all we need to know are the posterior and prior probabilities of observing a shift of at least s (or less than s if testing a decrease). If we were instead interested in testing simply for the presence of a shift, then we choose $s = 1$. Under both our HSMRF- and GMRF-based models, the prior probability $\Pr(\lambda_i/\lambda_j < 1 \mid HSMRF) = 0.5$ (for any $i \neq j$), making the denominator (the prior odds) 1 and only requiring us to compute the numerator (the posterior odds). For the HSMRF-based model, $\Pr(\lambda_7/\lambda_{43} < 1.0 \mid \mathbf{y}, HSMRF) = 0.98$, and the 2ln(BF) in favor of a birth-rate shift over this interval is 7.73 (using the nomenclature of Kass and Raftery (1995), "strong" support [71]). For the GMRF-based model, equivalent calculations produce

a 2ln(BF) in favor of a birth-rate shift of 5.53 ("positive" support). If we had instead been interested in testing for a shift of a particular magnitude, we could simulate under the prior to estimate the prior odds.

3.4.4 HIV Dynamics in Russia and Ukraine

In Eastern Europe and Asia, the use of injected drugs was a driving force in HIV epidemics for many years and continues to be an important factor in the spread of HIV [27]. Russia and Ukraine have a particularly high number of people who inject drugs, 2 million individuals combined, and a total of 1 million HIV-infected individuals [151]. These factors, plus a limited effort to reduce the scope of the problem in the beginning of the epidemic, make Russia and Ukraine a good source of data for estimating how HIV spreads among those who inject drugs. Vasylyeva *et al.* (2016) used a number of approaches, including phylodynamic methods, to study the course of the epidemic from the 1980s through 2011 [151]. They estimated that half of all secondary infections take place during the first month post-infection. They further identified a massive increase in the size of the infected population during the 1990s, and estimated that during this period each newly infected individual transmitted to at least 5 others.

When using birth-death models for infectious disease phylodynamics, the primary parameter of interest is the effective reproductive number at time t, $R_e(t)$. This is defined as the average number of individuals who will be infected by a single infectious individual introduced into a population with the same numbers of susceptible and removed individuals as are present in the population of interest at time t [42]. In a constant-rate birth-death-sampling-treatment process, the average duration of an infection is the inverse of the total rate of becoming noninfectious, or $(\mu + r\phi)^{-1}$. The expected number of infections an individual will cause over a timespan t is given by $\lambda \cdot t$, approximately for small t. Thus, in the constant-rate case, the expected number of secondary infections caused by an individual is $R_e = \lambda/(\mu + r\phi)$. In the time-varying case, if an individual becomes infected at time t, we use the rates at that time to compute the expectation and obtain $R_e(t) = \lambda(t)/(\mu(t) + r(t)\phi(t))$.

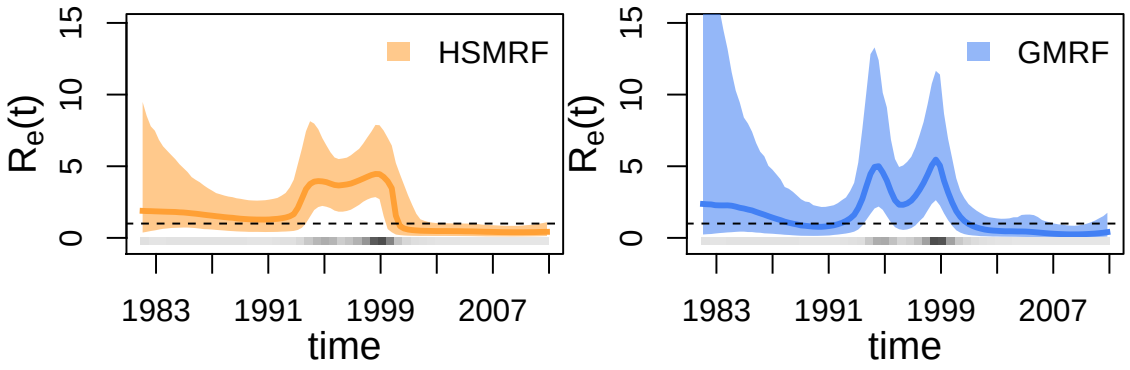


Figure 3.8: Analyses of the HIV dataset. Plotted are posterior median trajectories (dark lines) and 90% credible intervals (shaded regions). The upper CI for the GMRF-based analysis extends to ≈26, we have truncated the figure for a clearer view of the rest of the trajectory. Time is plotted as calendar time. A line at $R_e = 1$ is provided for convenience, as below this threshold the epidemic cannot be sustained. In grey is a heatmap of the inferred divergence times.

To understand the dynamics of HIV in Russia and Ukraine in the time period of interest, we use the sequence alignment for the *env* region from Vasylyeva *et al.* (2016) [151]. We analyze this dataset (457 sites for 92 sequences) under both the HSMRF-based and GMRF-based models, defining 2011 (the time of the most recent sample) to be the present day and dividing the range [0,29.1] into 100 evenly-sized intervals. We employ a Normal(29.1,5.0) root age prior, truncated to be older than the oldest sample age of 18. We fix $r = 1$, corresponding to the assumption that an individual, once sequenced and diagnosed, will not cause any further infections because they will be provided treatment and will have undetectable viral load. As there is information about the duration of infection in HIV, and thus the death rate, we replace our usual Exponential(10) prior with a Lognormal(-2.272,0.073) prior on the death rate, μ (the rate of becoming noninfectious in the absence of sampling and treatment). This corresponds to an *a priori* 95% probability that an untreated individual will be infectious for between 8.4 and 11.2 years [151, 84]. Details of the substitution and clock models are available in S1 Text, as are details of MCMC convergence diagnostics performed.

While the model we fit only has a time-varying birth rate, we plot the more informative $R_e(t)$ instead in Figure 3.8. Both the HSMRF-based and GMRF-based models show evidence

for a spike in $R_e(t)$ in the early 1990s and a sharp decrease at the end of the 1990s. We quantify support for shifts in $\lambda(t)$ instead of $R_e(t)$, as we do not directly parameterize the effective reproductive number. The 2ln(BF) in favor of an increase between 1992 and mid-1994 (of any magnitude) are 6.20 (strong support) for the HSMRF-based model and 8.06 (strong support) for the GMRF-based model. Similarly, the 2ln(BF) in favor of a decrease between 1999 and 2001 are 7.63 (strong support) for the HSMRF-based model and 9.91 (strong support) for the GRMF-based model. However, where the HSMRF-based model largely shows evidence for a consistently elevated rate in this period, the GMRF-based model shows a sharp dip midway through the decade, with the 90% CI including $R_e(t) = 1$. The HSMRF-based model estimates an average rate in this interval of 3.99, with rates that may be as low as 1.83 or as high as 7.88, and the GMRF-based model estimates an average rate of 3.61 with rates possibly as low as 0.55 or as high as 11.65.

The results of our HSMRF-based model analysis are largely consistent with those of Vasylyeva *et al.* (2016), who also observed an increased rate of infection from 1995 to 2000 [151]. Our GMRF-based analysis, with its large decrease in $R_e(t)$, does not align with either the prevalence data or any analysis performed by Vasylyeva *et al.* (2016) [151]. While both our HSMRF-based and GMRF-based models estimated $R_e(t) < 1$ throughout the 2000s, there is no evidence from HIV prevalence that the epidemic is decreasing [28, 36]. Examining the posterior distribution on phylogenies provides some insight into this apparent conflict: there are few infections inferred to have happened post-2000, and thus there is no information suggesting that $R_e(t) > 1$ in this period. Previous coalescent analyses have favored a higher $R_e(t)$ persisting with no sign of a decrease, however such models can have difficulty inferring decreases in the absence of coalescent events [30]. This highlights the fact that while birth-death process and coalescent models are good at peering into the past, without birth (coalescent) events there is little to no information from which to infer birth (coalescent) rates and thus the posterior distribution is largely determined by the prior distribution. On the other hand, there is outside evidence that the epidemic has slowed since 2005 [154], so it is possible that our models are picking up on a real signal and simply

exaggerating it.

3.5 Discussion and Conclusion

In this chapter, we use a piecewise-constant birth-death model, combined with both GMRF and HSMRF prior distributions, to approximate arbitrary changes in both the birth and death rates through time. We implement these models in the statistical phylogenetic software platform `RevBayes`, allowing for both inference of birth-death process parameters using a phylogeny as data and for joint inference of BDP parameters, phylogeny, and nuisance parameters directly from molecular sequence data. Additionally, we present an intuitive scheme for setting the key hyperparameter for these models, the global shrinkage parameter, and provide an efficient and tuning-parameter free inference framework that enables inference for these high-dimensional models. We find that both GMRF- and HMRF-based models are capable of inferring variable birth rates and correctly rejecting variable models in favor of effectively constant models. When estimating birth rates, we see that in general the HSMRF-based model has higher precision than its GMRF counterpart, with little to no loss of accuracy. Applied to a macroevolutionary dataset of the Australian gecko family Pygopodidae (where birth rates are interpretable as speciation rates), our models detect a speciation-rate decrease in the last 12 million years. Applied to an infectious disease phylodynamic dataset of sequences from HIV subtype A in Russia and Ukraine, our models detect a complex pattern of variation in the rate of infection.

Through simulations we find that different functional forms of birth-rate variation produce unique challenges in estimating these forms, even if they share the same magnitude of change. Slow changes are easy to miss, intermediate shifts are largely detectable, and fast shifts are generally hard for the GMRF-based model but easy for the HSMRF-based model to estimate. It is likely that slow changes are difficult to detect because both priors prefer piecewise-constant trajectories to continuous variation. As there are relatively fewer births in the older part of the tree, the prior can more easily overwhelm the likelihood, leading to an effectively constant model being fit. This also likely contributes to the increase in

uncertainty through time in the estimated rate. Fast shifts cause issues for the GMRF-based model because they require the global scale parameter γ to be large, which results in noisy and imprecise inference of slowly changing parts of the birth trajectory. At the same time, the GMRF-based model has a tendency to over-smooth the rapid changes. More recent changes are generally easier to detect than older changes, although very recent changes are often missed. Larger magnitude shifts are easier to detect than smaller magnitude shifts for both models, regardless of the functional form. In general, factors that make detecting shifts easier also exacerbate the poor behavior of the GMRF-based model. The HSMRF-based model often favors a trajectory with one or a few steeper shifts, even when the truth is a more gradual change. However, even if the duration of the shift is not accurately estimated, the HSMRF can recover the presence of rate variation even when the GMRF fails. Overall, we find that the performance of the HSMRF is quite good, and it is only clearly outperformed by the GMRF on a few types of birth rate trajectories.

Our simulations with time-varying death rates show that birth rates can be estimated quite well even in the face of difficulties inferring death rates. Serial sampling greatly improves the precision of the estimated birth rates and seems to mitigate the tendency of uncertainty to increase farther into the past. The increasing variability in the estimated rate is likely a function of both the increase in prior variability (due to the directional nature of our prior) and the reduced number of birth events in the past. The presence of serial samples can increase the number of observed birth events early in the process and thus improve both accuracy and precision in the estimated rate in older intervals. One theme that becomes even more evident in these simulations is that when the true birth-rate trajectory includes large jumps, the GMRF-based model will tend to infer spurious variability in the trajectory in regions where the birth rate should be small. When the death rate varies, the GMRF-based model is no more accurate than the HSMRF-based model in inferring the birth rate, but it infers substantially more variability than the HSMRF-based model, suggesting that much of this variability is in fact spurious. On the whole, though, we find that birth rates are generally well-estimated and that when there is serial sampling, the HSMRF-based

model can capture the presence of variability in the death rate. This may at first seem contradictory with the findings of Louca and Pennell (2020) [86], who have shown that a fully-variable $\lambda(t)$ and $\mu(t)$ cannot fully be identified from an extant phylogeny. For any tree there is an infinitely large "congruence class" of diversification-rate histories that are equally likely. We think that our results show the potential that priors provide for mitigating this problem. Bayesian inference introduces regularization to this problem in the form of prior distributions, which in general should reduce the size of the congruence class. Our results suggest that, at least in some instances, our priors are strong enough to ensure that there is only one set of birth- and death-rate trajectories that are plausible in light of observed data and imposed priors.

There are several avenues by which random field birth-death processes might be extended. It would be useful to devise models that can accurately infer slower declines, situations where the models we have put forth here have difficulty. One option for this would be to build second order Markov random field models, which can more easily collapse to linear models. These models have shown promise in coalescent modeling [43], but they have a higher risk of over-smoothing than first order models. Extending the models to include time-varying sampling rates may prove useful. Covariates may be added to time-varying birth-death models; previous work on birth-death models for macroevolution allowed for climate-dependent rates [21], while previous coalescent-based models considered the size of the region in which infections were found [52]. Adding covariates to the HSMRF-based model may allow for better success in inferring time-varying death rates by providing additional information. Models that allow for the serial sampling rate to vary may have better success (with or without covariates), as there is more direct information about this rate. However, cases where a number of samples have the same recorded age but there is not a sampling event (such as when some epidemiological sampling dates are available only to the year), may prove difficult. In such a case, the apparent variation in the sampling rate will likely overwhelm any signal of true variation in the sampling rate and may lead to erroneous estimates of the birth rate. Finally, for phylodynamic applications like HIV, it is clear

that GMRF-based and HMRF-based birth-death models would benefit strongly from the inclusion of epidemiological data, which has been incorporated into time-homogenous birth-death-sampling-treatment process by Gupta *et al.* (2019) [55].

In this work, we have developed and explored the performance of an HSMRF-based birth-death model for time-varying birth and death rates. This model is capable of detecting slow or rapid shifts in birth rates, and can infer the timing of rapid shifts quite accurately. Detecting variability in birth and death rates simultaneously is problematic, but possible. The HSMRF-based models are is extensible, and incorporating covariates or variable sampling rates will widen the range of potential applications of these models. The GMRF-based models may also prove useful, but they have issues with both spurious inference of variability in birth rates and under-detection of variability in death rates. Therefore, we recommend using the HSMRF as Bayesian nonparametric priors for birth-death models.

Chapter 4

HOW TRUSTWORTHY IS YOUR TREE? BAYESIAN PHYLOGENETIC EFFECTIVE SAMPLE SIZE THROUGH THE LENS OF MONTE CARLO ERROR

4.1 Abstract

In phylogenetics, Bayesian inference is a popular and widely-used approach to infer phylogenies (evolutionary trees). However, despite decades of widespread application, it remains difficult to judge how well a given Bayesian Markov chain Monte Carlo (MCMC) run explores the space of phylogenetic trees. In this chapter, we investigate the Monte Carlo error of phylogenies, including variability in estimated edge/branch (known in phylogenetics as split) and tree probabilities, and variability in the estimated summary tree. Specifically, we ask if there is any measure of effective sample size (ESS) applicable to phylogenetic trees which is capable of capturing the Monte Carlo error of these three quantities. We find that some there are ESS measures capable of capturing the error inherent in using MCMC samples to estimate quantities of the posterior distributions. We term these tree ESS measures, and identify a set of three which are useful in practice for assessing the Monte Carlo error. Lastly, we present visualization tools that can improve comparisons between multiple independent MCMC runs by accounting for the Monte Carlo error present in each chain. Our results indicate that common post-MCMC workflows are insufficient to capture the inherent Monte Carlo error of the tree, and highlight the need for both within-chain mixing and between-chain convergence assessments.

4.2 Introduction

Bayesian inference via Markov chain Monte Carlo (MCMC) is widely used in phylogenetic estimation. MCMC enables the generation of samples according to arbitrary distributions, such as the posterior distribution of a phylogenetic model, though it must draw autocorrelated samples. In fact, it is only in the limit of running the analysis infinitely long that we are guaranteed that summaries of our MCMC samples will converge to the true values of the corresponding posterior summaries. Short of finding a way to compress time, users are thus left with the task of determining whether inference from their MCMC samples are trustworthy. In practice, this entails assessing whether any given chain appears to be stationary, and assessing how well it is mixing. Generally, multiple MCMC chains are run, enabling users to also compare whether those chains appear to be sampling from the same distribution. In this chapter, we focus on the issue of mixing as embodied in the notion of the effective sample size (ESS). The effective sample size is closely related to the notion of Monte Carlo error, which describes the error in parameter estimation due to using sampling-based approaches.

Phylogenetic posterior distributions are complex objects, and it is known that sampling them via MCMC can be quite difficult [75, 61, 161, 96]. Overall performance of the MCMC chain is often diagnosed by looking at the trace (collection of samples through time) of the log-likelihood or the log-posterior density, though this should only be the first step in a more rigorous pipeline [81]. `Tracer` is a popular tool for summarizing and visualizing MCMC samples from phylogenetic software, which will automatically compute the ESS for all continuous model parameters and the densities of the log-likelihood and log-posterior [121]. `Tracer` flags parameters if the ESS is below 200, commonly taken as a rule-of-thumb minimum [77]. This can be useful for determining whether continuous model parameters have been sampled appropriately, though [39] argue for a more stringent cutoff of 625.

These tools and guidelines do not address a central question: how well did a given MCMC run sample from the posterior distribution of tree topologies? Tree sampling is challenging, and previous theoretical [101] and empirical [56] work has demonstrated decoupling between

the mixing of the log-likelihood and the sampling of the tree. For assessing MCMC convergence of trees, standard practice involves running multiple chains and comparing the estimated split probabilities [81]. Splits correspond to edges in an unrooted phylogeny, and represent bipartitions of taxa (tips of the tree), and can be useful for comparing tree distributions even when there are many sampled tree topologies [81]. This approach only addresses whether multiple runs (or different windows of a single run) are sampling similar distributions. Thus, split-based between-chain comparisons are not sufficient for ensuring adequate tree sampling via MCMC.

One promising approach to answering the question of MCMC performance for the phylogeny is to extend the notion of effective sample size to trees. [77] present two approaches for computing a single ESS for tree topologies which they term the approximate ESS and the pseudo-ESS (both available in the `RWTY` package [159]). They additionally present several simulation-based validations of their ESS measures, but do not address how those ESS measures capture Monte Carlo error. [50] and [39] alternatively consider taking the ESS individually for each split by representing it as a 0/1 random variable, which allows standard ESS computations. However, this approach has drawbacks, in that it does not account for correlation in presence/absence of splits due to the (shared) tree topology and it cannot provide uncertainty about unobserved splits or splits present in every tree. Further, trees can be summarized in many ways, and there is no obvious way to link a vector of per-split-probability ESS values to the Monte Carlo error of other key quantities, like the probabilities of different tree topologies or the summary tree.

In this chapter, we seek to understand the relationship between an ESS for phylogenies and phylogenetic Monte Carlo error. This requires special consideration because trees are complex and high-dimensional objects. Thus, classical means of linking ESS to Monte Carlo error are not directly applicable, and a considerable amount of the paper will be dedicated to making this link. We note that in this chapter, we will often use "tree" as synonymous with the topology of the phylogenetic tree (the set of nested relationships describing which lineages are most closely related), as we focus on the challenges posed by the discrete tree

structure. In this chapter, we will refer to ESS measures for phylogenies interchangeably as either tree or topological ESS measures. The goal of such measures is to adequately describe the mixing and autocorrelation of the MCMC samples of phylogenies such that the Monte Carlo error of phylogenetic quantities can be addressed.

Classically, the ESS is defined using the variance of the sample mean. Imagine we wish to estimate the mean, $\bar{\mu}$, of some distribution with known (true) variance σ^2 given n samples from this distribution. If the n samples were independent, we would have a direct link between the variance of our sample mean and the number of samples, given by $\text{Var}(\hat{\mu}) = \sigma^2/n$, where $\hat{\mu}$ is the sample mean. However, when samples are dependent this will underestimate the true variance of the sample mean. The ESS is a hypothetical number of independent samples which corrects for this and yields the correct variance (or standard error) of our estimator of the mean [152], which can be defined by $\text{Var}(\hat{\mu}) = \sigma^2/\text{ESS}$. The ESS is important to MCMC-based Bayesian inference because we must use correlated samples to estimate quantities of the posterior distribution. Run infinitely long, it is guaranteed that the posterior mean of a parameter calculated from MCMC samples will be infinitesimally close to the true posterior mean. Given finite run lengths we must account for the resultant error to understand how precise our estimates are [103, 70]. The Markov chain central limit theorem establishes that the sampling distribution of a mean converges asymptotically to a normal distribution [69]. Thus, if we know the ESS of a (real-valued) model parameter, we can construct confidence intervals for it, making the ESS a key quantity for Bayesian inference.

In this chapter, we focus on two questions: for what aspects of estimated phylogenies might a tree ESS capture Monte Carlo error, and how might we compute an ESS of phylogenies? Specifically, we consider whether an ESS of phylogenies is informative about the sampling variability of (1) the probabilities of splits in the tree, (2) the probabilities of tree topologies, and (3) a summary tree. Then, we describe several different ways to compute putative measures of tree ESS, including approaches of [77] and newly-derived approaches. These methods can be broken down into three categories: (A) generalizing continuous variable ESS identities to trees, (B) computing the ESS on a reduced-dimensional representation

of the trees, and (C) *ad-hoc* methods. We test these putative definitions and approaches to a tree ESS via simulations. In these simulations, we take probability distributions on phylogenetic trees inferred from real data and run an MCMC sampler targeting this known posterior. This allows us to compute brute-force estimates of sampling variability to which to compare the estimates based on our ESS measures.

We conclude with a case study, applying all ESS measures to six datasets of frogs and geckos from Madagascar [127]. Time-calibrated phylogenies of these groups were originally used to test hypotheses about adaptive radiations, and the datasets were revisited by [77] as benchmarks for their ESS measures. Using these datasets, we demonstrate how tree ESS can be used to construct confidence intervals on split probabilities. When combined with a common visual multi-chain convergence diagnostic, the split probability plot, this allows us to decompose between-chain disagreement into disagreements that can be attributed to low sample size, and disagreements that cannot. This highlights both the importance of within-chain measures of Monte Carlo error, but also how they form only a part of the larger picture of MCMC convergence. Indeed, our results show that multiple chains are always necessary to be confident in phylogenetic MCMC convergence. Furthermore, these multiple-chain runs show that even with confidence intervals derived using our best ESS estimates, pairs of runs have distinctly different estimates of split probabilities. Taken together, our empirical and simulated results make clear the importance of directly assessing how well the tree topology mixes using a tree-specific ESS.

4.3 Methods

We first offer a brief overview of this section before diving into the methods. In the first subsection, we present a brief summary of Bayesian phylogenetic inference as a point of reference. Next, we review background information on the ESS of one-dimensional Euclidean random variables, including how it is linked to the estimator variance (of the sample mean) definition above and how it is computed. Then, we detail three Markov chain Monte Carlo standard error (MCMCSE) measures which we will use to assess whether any tree ESS

measure works as intended. Specifically, we investigate Monte Carlo variability in estimated split probabilities, estimated tree probabilities, and the estimated summary tree. Lastly, we consider four tree ESS measures in several different families of approaches, including two methods from [77].

4.3.1 Phylogenetic inference

Phylogenetic inference is primarily concerned with the estimation of a phylogenetic tree topology, τ, from character data such as DNA sequences [81]. Standard inference approaches require a number of continuous parameters, $\boldsymbol{\xi}$, and compute the likelihood, $\Pr(y \mid \tau, \boldsymbol{\xi})$ using the Felsenstein pruning algorithm [45]. In Bayesian phylogenetic inference, we use MCMC to sample from the $\Pr(\tau, \boldsymbol{\xi} \mid y)$ and marginalize out $\boldsymbol{\xi}$ to obtain the distribution of interest, $\Pr(\tau \mid y)$. In this article, we consider unrooted tree topologies, where the topology depicts a set of nested evolutionary relationships without directionality [81]. Unrooted trees can be defined by the collection of edges, which are often referred to as splits or bipartitions of taxa. Every fully resolved tree with n_{taxa} tips contains $n_{\text{taxa}} - 3$ non-trivial splits (internal edges) and n_{taxa} trivial splits (pendant edges which exist in all trees). One common summary of the posterior distributions on trees is the majority-rule consensus tree. The (strict) MRC tree includes every split estimated to posterior probability greater than 0.5, and only those splits [93, 81], and is widely implemented in software for both Bayesian and maximum likelihood inference (where it summarizes bootstrap support). The set of all these splits correctly defines a tree [129], though it may not be completely resolved. The posterior probabilities of splits (henceforth simply split probabilities) themselves are useful as measures of posterior support for particular evolutionary relationships.

4.3.2 Classical definitions of ESS

Before we dive into the classical definition of ESS, we briefly review the Markov chain central limit theorem. Assume we have a set of MCMC samples (with any burnin period previously removed as needed) $X_1, \ldots, X_n$, a function g, and use $\bar{g}_n$ to denote the sample mean using

n samples:

$$\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i). \tag{4.1}$$

Let π denote the posterior distribution and $\mathbb{E}_\pi[g] = \int_X g(x)\pi(dx)$ the true posterior expectation. Using the notation of [47], the Markov chain central limit theorem tells us that the asymptotic distribution of $\bar{g}_n$ is,

$$\bar{g}_n \sim \text{Normal}(\mathbb{E}_\pi[g], \sigma_g^2/n). \tag{4.2}$$

Following [47], we can write σ_g^2 as the sum of a series of autocovariances at increasing time lags,

$$\sigma_g^2 = \text{Var}_\pi(g(X_1)) + 2\sum_{i=2}^{\infty} \text{Cov}_\pi(g(X_1), g(X_i)), \tag{4.3}$$

where we assume $X_1 \sim \pi$ and use π as a subscript to denote this. For the rest of this section, we will assume that g is the identity function, such that $g(x) := x$, and $\bar{g}_n$ becomes the average of the MCMC samples, $\bar{g}_n = \bar{X}$. For clarity of terminology, we will distinguish between the two variance quantities of interest. We will use σ_π^2 to represent the variance of the posterior distribution,

$$\sigma_\pi^2 = \text{Var}_\pi(X_1).$$

We will refer to σ_g^2 as the limiting variance and represent it as σ_{lim}^2. In this simpler case, where we only care about the sample mean, Equation 4.2 tells us that

$$\frac{\sigma_{\text{lim}}^2}{n} = \text{Var}(\bar{X}). \tag{4.4}$$

This reveals why σ_{lim}^2 is an important value: it is tied directly to the variance of our estimate of the mean. Thus, it allows us to construct confidence intervals for the posterior mean computed from MCMC samples.

The effective sample size of a set of MCMC samples is the (hypothetical) number of independent samples which would have the same variance of the sample mean [83]. Written

as an equation, this yields $\sigma_\pi^2/\text{ESS} = \sigma_{\text{lim}}^2/n$, which we can rearrange as

$$\text{ESS} = n\frac{\sigma_\pi^2}{\sigma_{\text{lim}}^2}. \tag{4.5}$$

While in practice we only know n, we can easily estimate σ_π^2 from the MCMC samples as $1/n\sum_{i=1}^{n}(X_i - \bar{X})^2$. The difficulty in estimating the ESS lies in estimating either σ_{lim}^2 or $\sigma_\pi^2/\sigma_{\text{lim}}^2$. In practice, there are a number of different approaches to estimating these quantities, which do not always yield the same answers [39].

4.3.3 Assessing performance of candidate ESS measures

Before discussing our approach to assessing whether tree ESS measures work, let us discuss the simpler case of a single Euclidean parameter and the classical ESS. Classical ESS measures "work" in the sense that they allow us to correctly compute the standard error of our estimate of the posterior mean $\hat{\theta}$. First, let us review the Monte Carlo standard error of a Euclidean variable in the context of Bayesian inference via MCMC. We want to estimate the true posterior mean with the mean of our MCMC samples, $\hat{\theta}$. The standard deviation of our estimator $\hat{\theta}$ is $\text{SD}(\hat{\theta}) = \sqrt{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}$. This is the standard deviation of the sampling distribution, also called the standard error (SE). In general, there is not a known closed form solution for the actual sampling distribution of $\hat{\theta}$, and thus for $\text{SE}(\hat{\theta})$. But if we were to run m independent replicate analyses, we could use a brute-force approach to estimate the *true* Markov chain Monte Carlo SE (MCMCSE) as,

$$\widehat{\text{SE}}_{\text{MCMC}}(\hat{\theta}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \widehat{\mathbb{E}}[\hat{\theta}])^2},$$

where $\hat{\theta}_i$ is the estimate of θ using the ith set of MCMC samples and $\widehat{\mathbb{E}}[\hat{\theta}]$ is estimated using all mn MCMC samples drawn. This is (an estimate of) the true standard error of the mean given n MCMC samples. Imagine that we had the ability to draw samples identically and independently distributed (iid) from the true posterior. Then, as a comparison, for each

run $i = 1, \ldots, m$, we could compute the effective sample size, ESS_i, and draw ESS_i samples iid according to the posterior distribution $\Pr(\theta \mid y)$. This gives us a second set of samples, for which we can repeat our above brute-force calculation of the Monte Carlo error. These m sets of ESS-equivalent iid samples allow us to estimate $\widehat{\text{SE}}_{\text{ESS}}(\hat{\theta})$, the standard error of the mean when drawing ESS samples iid from the true posterior distribution. We should expect to find that $\widehat{\text{SE}}_{\text{ESS}}(\hat{\theta}) \approx \widehat{\text{SE}}_{\text{MCMC}}(\hat{\theta})$, because the classical ESS is derived to give us the hypothetical number of independent samples such that we have the same variance of the posterior mean, and thus the same SD and same SE. We will use this approximate identity to judge the quality of various ESS measures.

When running MCMC targeting multivariate distributions, the usual approach in high dimensions is to compute the ESS separately for each univariate variable. However, phylogenetic posterior distributions are more complex than typical high-dimensional vector-valued distributions, which precludes this approach. Instead, we will consider several different summaries of posterior distributions of trees, and the Monte Carlo SE in the estimates of those summaries. We now write out more concretely our testing approach, which is a generalization of the above idea of how ESS measures should work. We will assume that we have m phylogenetic MCMC runs, and that each has n samples. We will assume for the moment that we have both a topological ESS measure and a way to draw tree topologies iid from the posterior distribution, $\Pr(\tau \mid y)$. Let $S(\tau \mid y, n)$ be a summary of the posterior distribution of trees based on a set n of MCMC samples from the posterior distribution $\Pr(\tau \mid y)$. For example, $S(\tau \mid y, n)$ could be the probability of a particular split or tree topology, or a summary tree such as the MRC tree. Our approach to testing topological effective sample sizes can then be written out as follows:

1. Run m independent MCMC chains.

2. Using the chains from step 1, compute the brute-force estimate of the true Markov chain Monte Carlo SE, $\widehat{\text{SE}}_{\text{MCMC}}(S(\tau \mid y, n))$.

3. For each chain $i = 1, \ldots, m$, compute the ESS, ESS_i, and draw ESS_i trees independently from the true posterior distribution $\Pr(\tau \mid y)$. Round ESS_i as needed such that we draw an integer number of trees.

4. Using the set of ESS-equivalent trees drawn in (3), compute the ESS-equivalent estimate of the MCMCSE, $\widehat{\text{SE}}_{\text{ESS}}(S(\tau \mid y, n))$.

5. Compare $\widehat{\text{SE}}_{\text{ESS}}(S(\tau \mid y, n))$ and $\widehat{\text{SE}}_{\text{MCMC}}(S(\tau \mid y, n))$.

Generally speaking, if a putative measure of the topological ESS works with respect to a MCMCSE measure, then $\widehat{\text{SE}}_{\text{ESS}}(S(\tau \mid y, n)) \approx \widehat{\text{SE}}_{\text{MCMC}}(S(\tau \mid y, n))$, and the closer the ESS-based estimate of the SE is, the better that ESS estimator works. If we have overestimated the ESS, we will find $\widehat{\text{SE}}_{\text{ESS}} < \widehat{\text{SE}}_{\text{MCMC}}$ (as we have drawn too many independent samples and thus the SE of the estimate from them will be too small). On the other hand, if we have underestimated the ESS, we will find $\widehat{\text{SE}}_{\text{ESS}} > \widehat{\text{SE}}_{\text{MCMC}}$ (as we have drawn too few independent samples and thus the SE of the estimate from them will be too large). As the scales of the Monte Carlo SEs are not inherently meaningful, we instead measure the relative error in the estimated Monte Carlo error (the relative Monte Carlo error, RMCE),

$$\text{RMCE} = \frac{\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{ESS}}}{\widehat{\text{SE}}_{\text{MCMC}}}. \tag{4.6}$$

This quantity is negative when we have underestimated the ESS, positive when we have overestimated the ESS, and tells us the relative proportion by which we have mis-estimated the MCMCSE using our ESS measure. In some cases, however, a more useful measure is given by the inflation (or deflation) of the true Monte Carlo error (ITMCE), which is defined by the relation,

$$\widehat{\text{SE}}_{\text{MCMC}} = \widehat{\text{SE}}_{\text{ESS}} \times \text{ITMCE}. \tag{4.7}$$

This value measures how inflated the *true* Monte Carlo error is relative to the estimated Monte Carlo error, and is computed as $\text{ITMCE} = 1/(1 - \text{RMCE})$. In essence, the ITMCE

measures how much wider the true confidence intervals should be relative to our estimated intervals. The ITMCE is greater than one when we have over-estimated the ESS, and less than one when we have underestimated it. These two measures are complementary; the RMCE is useful to quantify when ESS is under-estimated, while the ITMCE clearly shows when ESS is over-estimated.

In the introduction, we outlined three quantities a tree ESS might reflect; now we will more rigorously define three MCMCSE measures based on these, noting that they are all measures of the Monte Carlo standard error (MCMCSE). Broadly, we contemplate whether a tree ESS measure might reflect the standard deviation of our estimates of split probabilities, tree probabilities, or a summary tree.

ESS and split probabilities

We may wish for ESS of trees to reflect the quality of our estimates of the split probabilities. We defer the question of aggregating split probabilities to the results. For now, we focus on assessing the Monte Carlo estimation error of a single split. Let us denote the probability of this split p, the estimate of the split probability from the ith MCMC run as $\hat{p}^i_{\text{MCMC}}$ (Step (1)), and the estimate of the split probability computed from ESS_i tree topologies drawn iid from the posterior distribution as $\hat{p}^i_{\text{ESS}}$ (Step (3), we discuss how to do this in the section "Faking phylogenetic MCMC"). Then, we can estimate the true MCMCSE of the split probability as (Step (2)),

$$\widehat{\text{SE}}_{\text{MCMC}}(\hat{p}^i) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}^i_{\text{MCMC}} - \widehat{\mathbb{E}}[\hat{p}_{\text{MCMC}}]\right)^2},$$

where $\widehat{\mathbb{E}}[\hat{p}_{\text{MCMC}}]$ is the average probability of the split across all m chains. Then, using the ESS-equivalent tree samples, we can compute $\widehat{\text{SE}}_{\text{ESS}}$ (step (4)),

$$\widehat{\text{SE}}_{\text{ESS}}(\hat{p}^i) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}^i_{\text{ESS}} - \widehat{\mathbb{E}}[\hat{p}_{\text{ESS}}]\right)^2},$$

and then we can compare $\widehat{\text{SE}}_{\text{MCMC}}(\hat{p}^i)$ and $\widehat{\text{SE}}_{\text{ESS}}(\hat{p}^i)$ (Step (5)).

ESS and tree probabilities

We may wish for ESS of trees to reflect the quality of our estimates of the tree topology probabilities. As these are probabilities, we can compute $\widehat{\text{SE}}_{\text{MCMC}}(\hat{p}^i)$ and $\widehat{\text{SE}}_{\text{ESS}}(\hat{p}^i)$ exactly as for split probabilities. We again defer the issue of aggregating the many different tree probabilities and their Monte Carlo SEs to a later section.

ESS and the summary tree

We may wish for ESS of trees to reflect the quality of our estimates of the summary tree. In this chapter, we focus on the majority rule consensus (MRC) tree as the summary tree. As our measure of variability between topologies, we consider the squared Robinson-Foulds (RF) distance [122] (which we describe in more detail below). Thus, our MCMCSE measure here is a given by,

$$\widehat{\text{SE}}_{\text{MCMC}}(\widehat{\tau}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} d(\widehat{\tau}^i_{\text{MCMC}}, \widehat{\mathbb{E}}[\widehat{\tau}_{\text{MCMC}}])^2},$$

where $d(\cdot,\cdot)$ is the RF distance. Here, we define $\widehat{\tau}_i$ to be the MRC tree for the ith MCMC run, and $\widehat{\mathbb{E}}[\widehat{\tau}_{\text{MCMC}}]$ to be the MRC tree obtained by pooling all m runs to estimate split frequencies. Then, using the ESS-equivalent tree samples (step (3)), we can compute $\widehat{\text{SE}}_{\text{ESS}}$ (step (4)) analogously,

$$\widehat{\text{SE}}_{\text{ESS}}(\widehat{\tau}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} d(\widehat{\tau}^i_{\text{ESS}}, \widehat{\mathbb{E}}[\widehat{\tau}_{\text{ESS}}])^2}.$$

As there is only one MCMCSE measure per MCMC run here, there is no need for aggregation. When we introduce the tree ESS measures, we describe how this can be interpreted as a Frechét-like generalization of the classical ESS definition.

Faking phylogenetic MCMC

Our testing setup requires running many independent MCMC chains on every given dataset, and that we are able to draw iid samples from the target distribution. Because it is impossible to draw iid samples from a Bayesian phylogenetic posterior—else we would not need MCMC in the first place—we set up a simulated (or "fake") phylogenetic MCMC. We start with an estimate of a real phylogenetic posterior distribution, comprising a vector of trees $\boldsymbol{\tau}$ and the associated estimate of the probability mass function $\widehat{\Pr}(\boldsymbol{\tau})$. In this chapter we re-use posterior distributions inferred by [160], though any phylogenetic MCMC run could be used. The posterior distribution can also be restricted to 95% highest posterior density (HPD) set of trees, or some other subset, if desired. We then take the set of trees sampled and their estimated probabilities to be the *true* probabilities for our fake MCMC (if we have truncated the set of trees, we re-normalize the probabilities, though this is not strictly necessary). We define any tree $\Psi \notin \boldsymbol{\tau}$ to have zero probability. Thus, we have a realistic known true target distribution and can draw samples iid from it.

To run an MCMC sampler on this known target distribution, we use nearest neighbor interchange (NNI, the interchanging of two subtrees across an edge in the tree) proposals. First, we construct a list of all NNI neighbors for each tree, such that $N(\Psi)$ is the set of all neighbors of Ψ. Then, we randomly draw a starting state according to $\widehat{\Pr}(\boldsymbol{\tau})$. At each step thereafter, if the current state for the MCMC is Ψ, we propose a tree Ψ^* uniformly at random from $N(\Psi)$ (excluding Ψ), and the acceptance probability for the move is $\min(1, \widehat{\Pr}(\Psi^*)/\widehat{\Pr}(\Psi))$. This move is symmetric and has a Hastings ratio of 1 [75].

Although this is by necessity an artificial setup, we believe that it is an improvement over previous benchmarking exercises. The simulation approach of [77] is based on accepting all proposed MCMC moves and assuming all trees were uniformly probable. They consider bimodal distributions by mixing sets of trees based on a small number of MCMC moves from distant starting points. In contrast, our approach is based on real-data posterior distributions of trees. This allows for multimodality, as well as uneven connectivity of trees, such that some

trees have many neighbors (with notable posterior probability) and others have few, which can make exploration difficult as some trees may be hard to reach. Additionally, using real posterior probabilities replicates unevenness, such that some datasets may be particularly rugged and others relatively flat [61]. Importantly, our approach includes rejected proposals, an important feature given that acceptance rates for phylogenetic MCMC proposals are notoriously low. Our setup allows for complete exploration of the target distribution with chains of sufficient length, whereas with a uniform distribution on all trees this is a practical impossibility. As we can initialize the chain from the target distribution, we can ignore burnin and focus on mixing while accommodating these realistic features. We can improve the speed and memory requirements by tweaking the proposal step and neighbor tracking, which we discuss in Appendix C.

4.3.4 Computing a tree ESS

In this chapter, we consider four different tree ESS measures, which fall into the following three categories of approaches. These categories, and their constituent ESS measures, are as follows:

- ESS measures based on Frechét-like generalizations of Equation 4.5 to trees (we discuss Frechét-like generalizations in the next section)

    1. The Frechét Correlation ESS (frechetCorrelationESS)

- ESS measures based on projecting the tree to a single dimension and computing the ESS of that using standard univariate approaches

    2. The median pseudo-ESS (medianPseudoESS)
    3. The minimum pseudo-ESS (minPseudoESS)

- *Ad-hoc* ESS measures

4. The approximate ESS (approximateESS)

In the next sections, we present short sketches of our new approaches and the medianPseudoESS of [77]. We refer readers to [77] for explanation of the approximateESS. In Appendix C, we describe six additional measures, as well as a more detailed derivation of the frechetCorrelationESS. We will use the notation $\boldsymbol{\tau}$ for a vector of phylogenies, $d(\tau_i, \tau_j)$ to denote the (RF) distance between trees i and j, and $\boldsymbol{D} = \{D_{ij}\} = \{d(\tau_i, \tau_j)\}$ for the distance matrix of all trees in $\boldsymbol{\tau}$.

All tree ESS measures presented depend on the distance matrix between all samples in the posterior. While in general any tree distance can be used, in this chapter we focus on the Robinson-Foulds (RF) distance [122]. The RF distance considers an unrooted tree as a collection of splits, and measures the number of splits by which two trees differ. Given trees τ_A and τ_B, with $\mathcal{A}$ the set of splits in τ_A and $\mathcal{B}$ the set of splits in τ_B, then the RF distance is $d_{\text{RF}} = |\mathcal{A} \cap \mathcal{B}^c| + |\mathcal{B} \cap \mathcal{A}^c|$. That is, the RF distance is the number of splits in τ_A but not τ_B plus the number of splits in τ_B but not τ_A. While there are some limitations of RF distance, it has many benefits, primarily that it is interpretable and easily computable in a wide variety of software. Additionally, RF distance is related to tree space exploration via NNI moves, as two trees separated by one NNI are separated by an RF distance of 2.

Calculating the ESS by generalizing previous definitions

In this section, we will attempt to generalize the definition of ESS in Equation 4.5 using concepts borrowed from the notions of Frechét mean and Frechét variance [see for example 34, and references therein]. The Frechét mean and variance generalize the concepts of means and variances to other complete metric spaces. Where the variance is the average squared deviation from the mean, the Frechét variance is the average squared distance from the Frechét mean. In the case where X is continuous and one-dimensional and the Euclidean distance is chosen, the Frechét mean is the classical mean and the Frechét variance is the classical variance. For unrooted phylogenetic trees with branch lengths, Billera-Holmes-

Vogtmann (BHV) space [5] is a complete metric space, and this is where the Frechét mean and variance of phylogenies have been previously defined [see 13, and references therein]. However, in this chapter we work with purely topological measures such as RF distance and split probabilities. Thus, we describe the approaches as Frechét-like.

The Frechét-like correlation ESS, or frechetCorrelationESS, is a generalization of what we will call the sum-of-correlations ESS, which we will now review in the case of a single continuous random variable X. The autocorrelation function defines the correlation between samples at times separated by lag s, and can be related to the autocovariance [156],

$$\rho_s = \mathrm{Cor}(X_t, X_{t+s}) = \frac{\mathrm{Cov}(X_t, X_{t+s})}{\sigma_\pi^2}.$$

This means we can re-write Equation 4.3 as,

$$\sigma^2_{\mathrm{lim}} = \sigma_\pi^2 \times \left(1 + 2\sum_{s=1}^{\infty} \mathrm{Cor}(X_0, X_s)\right) = \sigma_\pi^2 \times \left(1 + 2\sum_{s=1}^{\infty} \rho_s\right).$$

Re-arranged, this can be combined with Equation 4.5 to get,

$$\mathrm{ESS} = \frac{n}{1 + 2\sum_{s=1}^{\infty} \rho_s}.$$

In practice, we run into difficulties if we apply this definition naively. As there are fewer MCMC samples separated by large time lags, as s gets larger our estimates $\hat{\rho}_s$ get noisier, and for odd time lags it is possible to have $\rho_s < 0$, meaning we cannot smooth the estimates without care. Following [153], we can overcome these limitations in several steps. First, they suggest summing adjacent correlations, defining $\hat{P}_{s'} = \hat{\rho}_{2s'} + \hat{\rho}_{2s'+1}$, which guarantees that $\hat{P}_{s'} > 0$. The estimated $\hat{P}_{s'}$ may not be monotonically decreasing, but the actual curve $P_{s'}$ should be, so [153] smooth the curve by setting $\hat{P}^*_{s'} = \min(\hat{P}_{s'}, \hat{P}_{s'-1})$ for all $s' > 1$. We then find the largest non-zero time lag, $k = \underset{s'}{\mathrm{argmin}}\ \hat{P}^*_{s'} > 0$, and sum the series only up to this

time, yielding the final estimator,

$$\widehat{\text{ESS}} = \frac{n}{-1 + 2\sum_{s'=1}^{k} \hat{P}^*_{s'}}. \tag{4.8}$$

We will refer to this as the sum-of-correlations ESS.

Employing Equation 4.8 to estimate an effective sample size for trees requires us to define a correlation between vectors of trees. It can be shown that, for Euclidean variables X and Y, the covariance is related to Euclidean distances. If we denote the Euclidean distance between X and Y with $\Delta = d(X, Y)$, we have $\text{Cov}(X, Y) = 1/2 \times (\text{Var}(X) + \text{Var}(Y) - \mathbb{E}[\Delta^2] + (\mathbb{E}[X] - \mathbb{E}[Y])^2)$. In a setting where X and Y represent samples from the same MCMC run at some time lag t, we might expect $(\mathbb{E}[X] - \mathbb{E}[Y])^2$ to be negligible (if the chain is stationary, then the mean does not change notably over time). In this case, we can instead approximate the covariance as $\text{Cov}(X, Y) \approx \frac{1}{2}(\text{Var}(X) + \text{Var}(Y) - \mathbb{E}[\Delta^2])$. By replacing Euclidean distances with arbitrary distance metrics and means/variances with Frechét-like means/variances, we can use these relationships to define a generalization of the covariance and thus of the (auto)correlation. We propose to estimate the autocorrelation between MCMC samples of trees at time lag s as,

$$\hat{\rho}_s = \frac{\frac{1}{2}(\widehat{\text{Var}}(\boldsymbol{\tau}_t) + \widehat{\text{Var}}(\boldsymbol{\tau}_{t+s}) - \widehat{\mathbb{E}}[\Delta^2])}{\sqrt{\widehat{\text{Var}}(\boldsymbol{\tau}_t)\widehat{\text{Var}}(\boldsymbol{\tau}_{t+s})}}. \tag{4.9}$$

Here, $\boldsymbol{\tau}_t$ and $\boldsymbol{\tau}_{t+s}$ represent vectors of MCMC samples of trees separated by a time lag of s, $\widehat{\text{Var}}$ is the estimated Frechét-like variance based on RF distance, and $\widehat{\mathbb{E}}[\Delta^2]$ is the average squared RF distance between tree vectors $\boldsymbol{\tau}_t$ and $\boldsymbol{\tau}_{t+s}$. Once we obtain our estimates $\hat{\rho}_s$, we use the sum-of-correlations approach to estimate the ESS (Equation 4.8). A more thorough derivation of Equation 4.9, including how to compute $\widehat{\text{Var}}(\boldsymbol{\tau})$, is available in Appendix C.

Approaches to calculating the ESS by projecting the tree to a single dimension

The pseudo-ESS of [77] projects trees to a single dimension and computes the ESS of this transformed variable. It is computed in several steps. First, r reference trees are selected from the set of posterior trees. For the ith reference tree, we turn the sequence of trees sampled by the MCMC into a real-valued sequence by computing the distance between each such tree and the reference. We can then compute the ESS of this integer sequence to obtain an ESS_i. Lastly, we must summarize the vector of estimates, $\text{ESS}_1, \ldots, \text{ESS}_r$. [77] use the median as a point estimate and suggest reporting the 95 percentile range as well. We depart slightly from the original approach by iterating over all n trees in the posterior as reference trees. As the ESS is a single number (the hypothetical size of an equivalent set of iid samples), we do not examine ranges but rather focus on point estimates. What [77] refer to as the pseudo-ESS (the median of the n distinct ESS values), we will refer as the medianPseudoESS. In addition to the median, as a potentially conservative estimator we also investigate the performance of the minimum of the ESS over reference trees, which we call the minPseudoESS. For both of these approaches, we first compute the dimension-reduced representation and then use the `R` package `coda` [119, 112] to compute the ESS. In Appendix C, we present three additional approaches to computing the tree ESS using projection, and a discussion of the `coda` approch to computing the ESS.

4.4 Results

4.4.1 Tree ESS measures reflect Monte Carlo variability in split probabilities, tree probabilities, and the summary tree

Now that we have several possible summaries of how an ESS measure could work and a number of potential ESS measures, we can determine whether any of the potential tree ESS measures are useful with respect to any of the summaries. We use a suite of common test datasets in phylogenetics [*e.g.* 75, 61, 78, 161], commonly referred to as DS1-DS8 and DS10. We re-use previous MCMC analyses of these datasets consisting of 10 independent

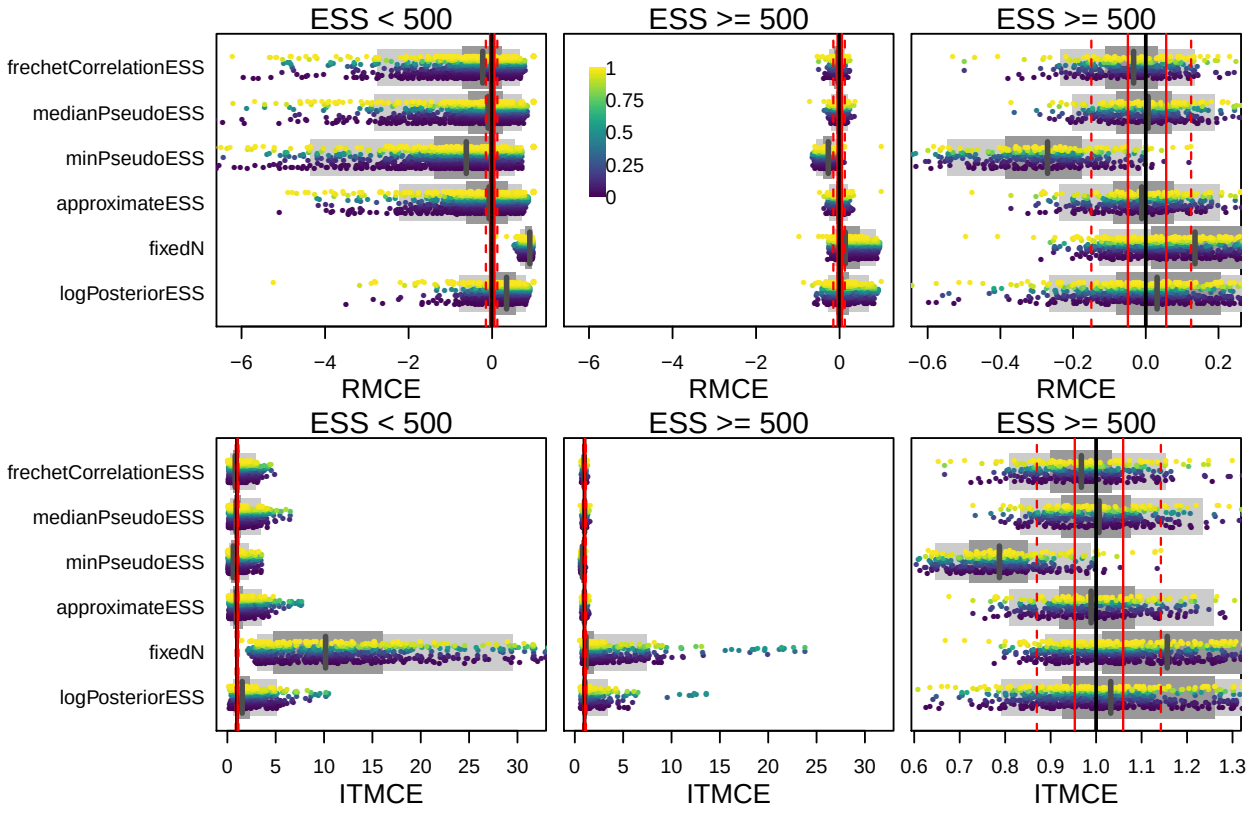


Figure 4.1: The RMCE $((\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{ESS}})/\widehat{\text{SE}}_{\text{MCMC}})$ and ITMCE $(1/(1-\text{RMCE}))$ for split probabilities for all topological ESS measures and all 45 DS by run-length combinations. Splits are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). The two right panels are the same except for the scale of the x-axis. The divide between the left and right panels is based on the estimated average ESS of each of the 45 simulations, such that all splits from a simulation with average frechetCorrelationESS of 100 would show up in the left panel, while all splits from a simulation with an average Frechét correlation ESS of 600 would show up in the right two panels. As fixedN always assumes ESS = 1000, for this row we split by the number of MCMC iterations run, with the left panel including 10^3 and 10^4, and the right panel 10^5, 10^6, and 10^7. The thinner light grey bar below the points shows the 95% quantile range, the thicker dark grey bar the 50% quantile range, and the grey line is the median. Ideal performance is RMCE = 0 and ITMCE = 1 (perfect estimation of the Monte Carlo SE). As references we have plotted a solid black line for perfect performance, while the dashed (solid) red lines represent the 95% quantile range (50% quantile range) from the univariate Normal(0,1) experiment. The best performance that might reasonably be expected of a tree ESS measure would match the Normal(0,1) experiment, and thus have the grey line on the solid black line, the the thinner light grey bar align with the dashed red lines, and the thicker dark grey bar align with the solid red lines. RMCE < 0 (ITMCE < 1) implies underestimating the ESS, while RMCE > 0 (ITMCE > 1) implies overestimating the ESS, thus the log-posterior ESS and assuming ESS $= n$ tend to overestimate the ESS for splits, often substantially, while most tree ESS measures are much closer to the truth.

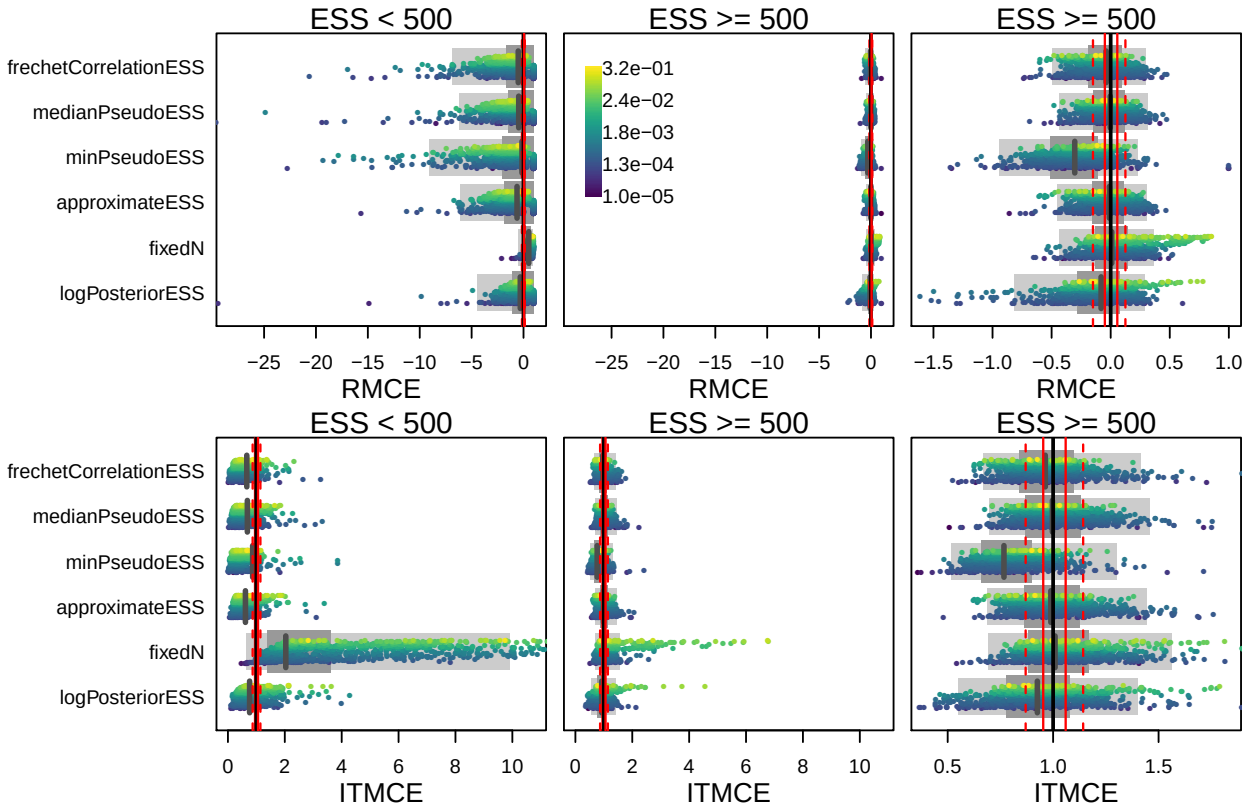


Figure 4.2: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{ESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ and ITMCE $(1/(1 - \mathrm{RMCE}))$ for topology probabilities for all topological ESS measures and all 45 DS by run length combinations. Tree topologies are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). As there are too many distinct topology probabilities (nearly 100,000 across all 45 simulations), we plot only 1000 per row, preferentially keeping the highest-probability trees as these are the ones that contribute most to summary trees. For more explanation, see Figure 4.1 caption.

MCMC chains of 1 billion iterations, sampled every 100 iterations and pooled into a set of 100 million tree samples [160]. As the full set of trees is too large to work with efficiently, we consider a subset of these trees which we term the "best connected trees." To obtain this subset, we start with either the 95% HPD or the first 4096 trees in the 95% HPD, whichever is smaller. Then, we ensure that every tree can be reached from every other tree using only NNI moves, and keep the largest connected subset. In practice this at most requires us to drop 20% of the trees. For each of these 9 real-data posterior distributions, we run simulated MCMC analyses (as described in the section "Faking phylogenetic MCMC") for

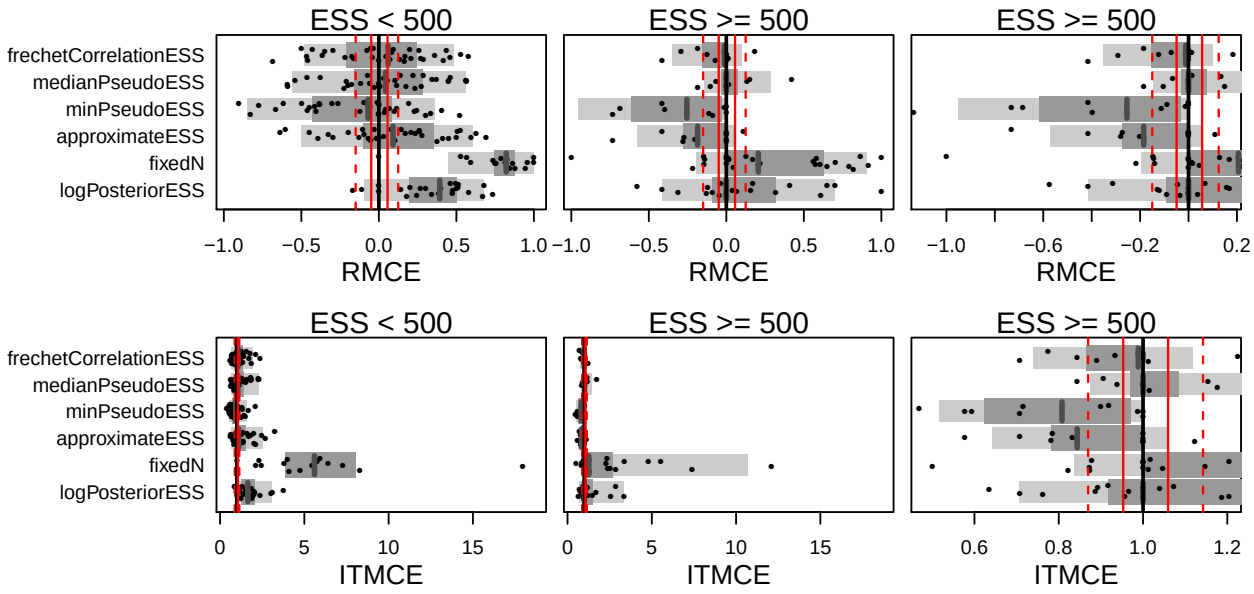


Figure 4.3: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{ESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ and ITMCE $(1/(1-\mathrm{RMCE}))$ for the majority-rule consensus (MRC) tree for all topological ESS measures and all 45 DS by run length combinations. The standard error for the MRC tree is a Frechét-like Monte Carlo SE, rather than a classical Euclidean Monte Carlo SE. For more explanation, see Figure 4.1 caption.

run lengths of $\{10^3, 10^4, 10^5, 10^6, 10^7\}$, in each case thinning in order to retain 1000 tree samples. Acceptance rates for the NNI moves are typical of those observed in real-data MCMC analyses, ranging from $\approx 3\%$ (DS3) to $\approx 11\%$ (DS6). Since we do not need to integrate out any of the continuous parameters such as branch lengths, and since we start each chain in the stationary distribution (avoiding burn-in), these shorter run lengths should be equivalent to longer runs on real datasets. For each dataset and run-length combination (of which there are 45), we run 100 replicate MCMC chains and use steps (1)-(5) to assess performance of the ESS measures for all summaries.

In the introduction, we mentioned that standard practice for phylogenetics often either ignores the mixing of the tree itself or uses the ESS of the log-posterior density as a proxy. Based on these approaches, we define two reference measures. One of these references, which we call fixedN, is to simply declare that the effective sample size is the sample size. This is akin to ignoring the mixing of the tree, in which case one essentially assumes that the n

tree samples are worth n independent samples. Secondly, we consider using the univariate ESS (as implemented in `coda`) of the trace of the log-posterior-density, which we call the logPosteriorESS.

As an additional reference for our simulation results, we also run our testing setup in a non-phylogenetic context: estimating the mean of a Normal(0,1) variable. We use 50 chain lengths between 1,000 and 50,000 and keep 1000 samples each time and use the `coda` univariate ESS. This procedure produces ESS values from nearly 0 to nearly 1000, a range comparable to the observed tree ESS values computed. We find a median RMCE (ITMCE) of 0 (1), with 50% quantiles spanning [0.05,0.06] ([0.95,1.06]), and 90% quantiles spanning [-0.15,1.2] ([0.87,1.14]). This establishes a baseline for how well we might expect our tree ESS measures to work, and provides a simulation-based validation that our procedure for testing ESS measures works.

Overall, performance of the tree ESS methods is variable across both simulated conditions and MCMCSE measures. We account for variable performance across simulated conditions by binning our simulations based on the estimated ESS. For each of the 45 simulated data analyses, and for each ESS measure, we compute the average ESS across all 100 replicate MCMC runs, and then bin these into the ESS < 500 and ESS ≥ 500 regimes. In the ESS < 500 regime, the RMCE and ITMCE are quite variable, and often far from the optimal values of 0 and 1 (Figures 4.1, 4.2, 4.3). This is not ideal, as it means that the performance is quite variable across different splits, trees, and datasets. In the ESS ≥ 500 regime, however, performance is much better for all methods. For both split probabilities (Figure 4.1) and the MRC tree (Figure 4.3), in both ESS regimes the tree ESS measures readily outperform the reference measures (fixedN and logPosteriorESS), and perform comparably to the univariate ESS of a Gaussian. For tree probabilities, however, none of these measures perform well as compared to standard univariate ESS of a Gaussian. Furthermore, the tree ESS measures perform almost as badly as the fixedN or logPosteriorESS approaches. It seems the tree ESS measures perform most usefully for split probabilities and the summary tree in the ESS ≥ 500 regime. In Appendix C, we present analogous figures for an ESS cutoff of 250, where the

performance is poor both below and above the cutoff.

The frechetCorrelationESS and medianPseudoESS are the best-performing tree ESS measures. Generally, the frechetCorrelationESS is a bit more conservative. For split probabilities, it avoids some of the over-estimation of the ESS evident in the medianPseudoESS in the ESS $\geq$ 500 regime. For the MRC tree, however, the medianPseudoESS is both less biased and performs about as well as the univariate ESS of a Gaussian. The approximateESS performs similarly to the frechetCorrelationESS and medianPseudoESS on split and tree probabilities, but notably worse for the MRC tree. The minPseudoESS is overly conservative, and even in the ESS $\geq$ 500 regime almost exclusively underestimates the ESS. In practice we recommend that practitioners consider multiple ESS measures, and note that any accurate quantification of Monte Carlo error requires an ESS of at least 500 for the ESS measure being used. However, which ESS measure to use depends on what question is being asked. If seeking the best description of error in the MRC tree, the medianPseudoESS is the clear choice. For split probabilities, the frechetCorrelationESS is the best choice. In both cases, the ESS measures perform about as well as the univariate ESS for a Normal(0,1) variable, indicating that they work well enough to be used in practice. Using the minPseudoESS alone is possible, though the cost is longer MCMC runs than strictly necessary to achieve a desired level of accuracy. Using all three in concert allows for the best estimation of error in split probabilities and the MRC tree, while also providing an upper bound on the error.

4.4.2 Case study

Now that we have some understanding of how tree ESS measures work (and do not work), we turn our attention to applying them to real-world datasets. Following [77], we apply our ESS measures to 6 datasets from [127]. These datasets are convenient because the same well-documented analysis methodology was applied to each, and because 4 independent replicate MCMC analyses for each dataset have been deposited at Dryad (`https://doi.org/10.5061/dryad.r1hk5`). While individual methods may often disagree, they all clearly agree that the Gephyromantis and Phelsuma datasets have low topological

ESS across all chains, and other than the approximateESS they agree that the 3rd Paroedura chain has a low ESS. In Figure 4.4, we show ESS measures computed for 1001 samples from each of the 4 replicate chains for all 6 datasets. On these datasets, the reference approaches (fixedN and logPosteriorESS) perform completely inadequately. Unlike in the simulations, though, here the logPosteriorESS says the ESS is much smaller than any of the actual tree ESS measures. That it is so low here, while the tree ESS measures are much larger, suggests that there were mixing problems with other model parameters than the tree, and highlights the need to address all parameters in the phylogenetic model, including the tree, on their own merits. That is, the logPosteriorESS is at best only loosely linked to the sampling of the tree topology, and its performance in our simulated scenarios is likely more closely linked to the actual Monte Carlo error in the tree than it ever will be in practice.

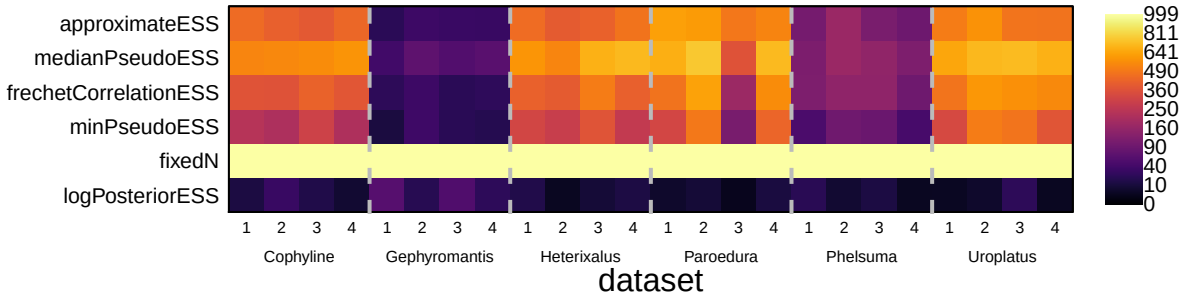


Figure 4.4: Tree ESS measures computed on 4 replicate chains for the 6 datasets from [127] as a heat map. To make differences clearer when ESS is low, the heatmap is spaced on the square-root scale. The ESS of the log-posterior and the fixedN approaches are included as references, though neither captures the meaningful between-dataset differences in topological ESS.

The classical Euclidean ESS can be used to construct frequentist confidence intervals for parameter estimates. In Figure 4.5, we use the tree ESS to construct confidence intervals for split probabilities, and update a standard MCMC convergence plot, the plot of split frequencies for two chains. Specifically, the confidence intervals are useful for determining whether the disagreement in split probabilities between two chains can be ascribed to a low sample size (CIs overlap), or whether it may be indicative of a convergence problem (CIs do not overlap). By plotting thresholds, we can also examine whether a given split can be

confidently assumed to be above a specific cutoff, such as 0.5 for inclusion in the MRC tree. In this case we find that, even accounting for the topological ESS, only chains 2 and 4 for the Paroedura dataset have complete agreement for all split probabilities. This highlights the importance of performing multiple independent MCMC runs and using between-chain convergence diagnostics to assess between-chain differences, in addition to assessing the ESS of parameters. On every dataset other than the Uroplatus dataset, regardless of ESS, at least two chains disagree about the probability of at least one split (Appendix Figures A.8-A.12). Topological ESS measures flag the Gephyromantis and Phelsuma datasets as having relatively low ESS, but despite the low ESS, all four Gephyromantis chains, and all Phelsuma chains except one, show levels of conflict in line with the other datasets. This further highlights the difference between low ESS (large uncertainty about parameter means) and between-chain convergence (two chains producing discordant estimates).

4.4.3 Data and code availability

All necessary functions for computing the ESS of phylogenetic trees have been implemented in an `R` package named `treess` available at `bitbucket.com/afmagee/treess`. Code for the simulation study and real-data analyses is available at available at `bitbucket.com/afmagee/tree_convergence_code`.

4.5 Discussion

In this chapter, we have investigated Monte Carlo error for phylogenetic trees. We present three summaries of Monte Carlo error that a topological ESS measure might capture, the MCMCSEs of split probabilities, tree probabilities, and summary trees. With simulations from real-data posterior distributions, we use these MCMCSE summaries to assess four putative tree ESS measures (as well as two additional references based on standard practice). We find that the performance of these measures varies across summary measures and within summaries based on the ESS. At their best, the frechetCorrelationESS and medianPseudoESS can capture the Monte Carlo error inherent in MCMC estimates of split probabilities and MRC

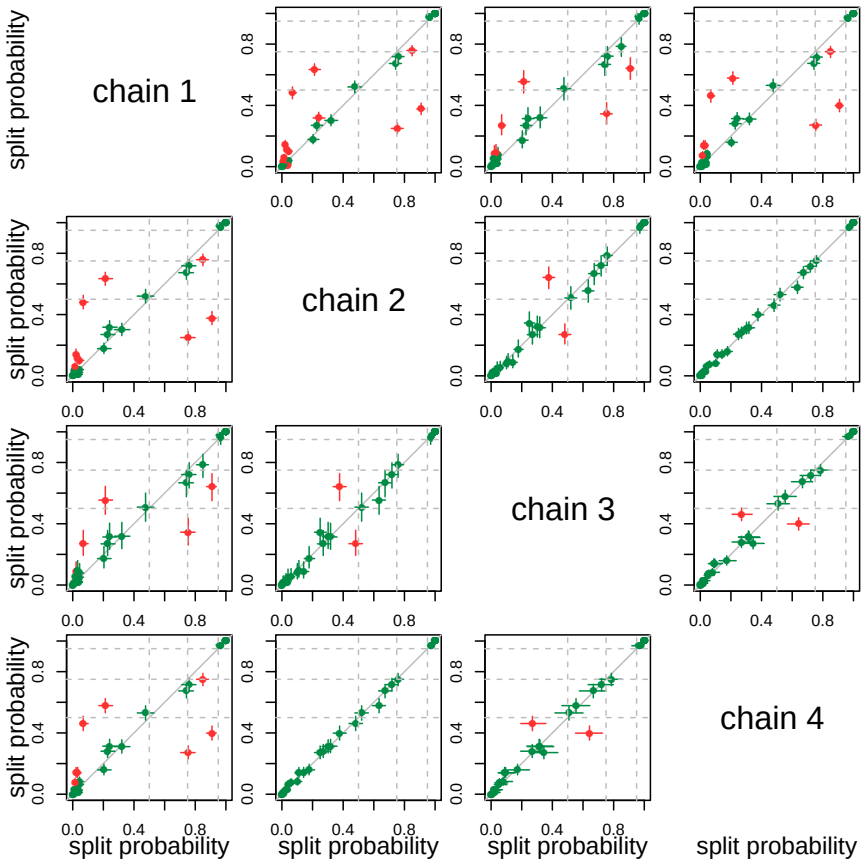


Figure 4.5: Split probabilities computed for all chains of the Paroedura dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. The upper diagonal uses the frechetCorrelationESS to compute confidence intervals, while the lower diagonal uses the minPseudoESS, which is generally smaller and thus leads to larger confidence intervals. Each confidence interval is colored by whether or not the 95% CI for the difference in split probability between chains i and j includes 0 (green for including 0, red for excluding 0). CIs for differences in probability that exclude 0 (or non-overlapping confidence intervals) are more likely to be indicative of convergence issues between chains, such that longer runs may still result in different estimated split probabilities. CIs for differences in probability that include 0 (or overlapping confidence intervals) suggest that longer runs will likely lead to identical split probabilities. Narrower confidence intervals from larger tree ESS estimates will flag more splits as problematic (as in chains 1 and 4). Dashed grey lines indicate posterior probabilities of 0.5 (threshold for inclusion in the MRC tree), 0.75 (moderate support for a split), and 0.95 (strong support for a split).

trees about as well as the univariate ESS captures the Monte Carlo error of the mean of a Normal(0,1) variable. That is, at their best, the frechetCorrelationESS and medianPseudoESS capture the Monte Carlo error of split probabilities and MRC trees about as well as could be hoped. When the estimated ESS is less than 500, however, performance is notably less than ideal, severely overestimating the ESS for some split probabilities and underestimating it for others (similar patterns hold for tree probabilities and the summary tree). For split probabilities and MRC trees, the frechetCorrelationESS and medianPseudoESS clearly outperform the standard practices of ignoring Monte Carlo error or using the ESS of the log-posterior density as a proxy. The tree ESS measures also capture Monte Carlo error in the estimated tree probabilities, though they do not clearly outperform the references. When assessing the ESS of tree topologies, we recommend that practitioners consider the minPseudoESS, frechetCorrelationESS, and the medianPseudoESS. If the estimated ESS is at least 500, then this combination provides an upper bound on the error (the minPseudoESS), and the best estimation of split probability error (the frechetCorrelationESS) and MRC tree error (the medianPseudoESS). If the estimated ESS is less than 500, then all three approaches may overestimate the ESS quite severely.

It is unexpected that the medianPseudoESS and the frechetCorrelationESS perform so similarly. The frechetCorrelationESS is derived as a generalization of the classical univariate ESS, and does not require a reference tree. The medianPseudoESS is the ESS of a reduced-dimensional representation of the MCMC samples and requires fixing a reference tree. We know of no obvious reason that they should work similarly.

While the simplicity of the purely topological measures we have considered are important as starting places for understanding MCMC for trees, extending this work to branch lengths via BHV space is an interesting possibility. This would have the advantage of additionally assessing the mixing of the branch lengths, which are currently unaddressed by most phylogenetic workflows. By comparing purely topological and BHV-based measures, one could also understand how differences in branch lengths contribute to the tree-to-tree distances relative to the topological distances. It is possible that most of the distance comes

from differences in branch lengths, and thus the BHV-based tree ESS measures miss key differences in topology. Alternately, BHV-based tree ESS measures could adequately capture both branch length and topological dynamics. It is also possible that these two regimes co-exist in different regions of parameter space. In either case, reducing the mixing of a very complex structure (a tree with branch lengths) to one or even two ESS measures would be a useful simplification. In BHV-space, it may also be possible to extend the work in [163] on confidence sets for phylogenetic trees to better describe the uncertainty in summary trees.

Recently, [56] performed an in-depth exploration of MCMC across thousands of empirical datasets. They computed the approximate ESS of [77] for all runs. This represents the largest use to date of tree ESS measures, though until now the properties and performance of these ESS estimators have been somewhat ambiguous. Our results show that effective sample sizes of at least 500 are generally capable of capturing the true mixing behavior of the tree, while smaller effective sample sizes are not. Since [56] found that most topological ESS were on the order of 750, in their usage the tree ESS is likely to adequately capture Monte Carlo error in the phylogeny.

[39] have recently published a paper on convergence diagnostics for phylogenetic MCMC, and an accompanying R package `convenience`. We agree with [39] (and [121]) that ESS $>$ 200 is an arbitrary cutoff, and appreciate the approach [39] take to deriving an ESS threshold which centers on Monte Carlo error. Our work departs from theirs, though, in considering the entire tree, where the tree-based diagnostics in `convenience` consider each split separately. As trees are inherently multivariate objects, we think it is important to attempt to consider them in their entirety. Consider, for example, that the fixedN approach estimates the Monte Carlo error for tree probabilities as well as the tree-based measures for longer runs (Figure 4.2), but severely underestimates the Monte Carlo error for split probabilities (Figure 4.1). In Appendix C we also show that while the split probabilities may appear to have converged over the course of an MCMC run, there can still be notable uncertainty in the MRC tree (Figure S1).

The tree ESS methods we have considered are applicable out-of-the-box to any data

structure where one can compute a distance. As distance measures exist for a wide range of objects, we hope that this work may prove useful for understanding Monte Carlo error in other non-Euclidean cases. In phylogenetic applications, one might seek to understand the Monte Carlo error in estimates of objects like ancestral state estimates or migration matrices.

We have shown that tree ESS measures exist which can sufficiently describe the Monte Carlo error of both split probabilities and the summary tree. Further, we have implemented these measures and approaches for construct confidence intervals for split probabilities (as in Figure 4.1) in the R package `treess`. Centering Monte Carlo error as we have in this chapter stands in contrast to the widespread use of ESS in phylogenetics, which is commonly treated as a "box that should be checked" by having ESS above a threshold before proceeding on to interpreting results. We hope that this work motivates phylogenetic community to take more seriously the quality of MCMC estimation of its focal parameter, and more broadly that it helps de-mystify the matter of how long to run phylogenetic MCMC. An analysis should be run long enough that the Monte Carlo error of the important quantities is small enough that conclusions are robust. Future work may enable auto-termination of MCMC when such a threshold is achieved and enable confidence intervals for the summary tree.

Chapter 5
DISCUSSION AND FUTURE DIRECTIONS

5.1 The future of birth-death models

5.1.1 The utility, and difficulties, of birth-death models

Birth-death models are potentially powerful tools for exploring the dynamics of speciation and extinction through time, but it has long been known that there are drawbacks. Extinction hides many events which occurred throughout evolutionary history. Species sampling can create data patterns resembling signatures of increasing extinction [115]. Even in simple constant-rate models, there are strong limits on what can and cannot be inferred [137]. Recent work [86, 85] has shown that there are an infinity of combinations of birth, death, and sampling rates that produce the same likelihood of a tree. And yet, despite this, the community persists in using, and trying to fix, birth-death modeling. This raises a question: why? Why not give up? The way I see it, the answer is, "because birth-death models are still the best tools available for many jobs." Birth-death models are directly linked to quantities we care about: rates of speciation and infection are of practical significance in addressing questions of biodiversity and disease spread.

It is not the case that other models can save us from the problems inherent in birth-death modeling. One alternative family of tree models includes the Kingman coalescent [73] and its offspring, derived as models for trees of a small sample of individuals (and a short stretch of the genome) in a much larger population. Coalescent models can be useful for phylodynamic and population-genetic applications, though important assumptions must be made to translate coalescent parameters into values like the effective reproductive number, R_e. From first principles, the coalescent is wildly inappropriate for modeling the tree of a group of species, and in practice these are very different tree priors which produce different

estimates of divergence times [134]. There are, of course, also non-mechanistic tree models [*e.g.* 124]. "Perhaps," we might hope, "such models can be of use if all we care about are estimating the divergence time." But we must then confront the fact that our phylogenetic likelihood can only truly estimate genetic distances, the product of evolutionary rate and time. And so we come full circle—even to simply date the tree, we need good mechanistic tree priors.

5.1.2 Extending time-varying birth-death-sampling models

The future of birth-death models is integrating multiple data sources. Data sources other than (genetic) sequences can provide us with additional information that allows us to tease apart rates of birth, death, and sampling. In the phylodynamic case, this is perhaps the easiest to see. Here, for example, we have information about the overall intensity of testing for the disease in question. This is distinct from the phylodynamic sampling parameter, but they should be related, and that relation would help us understand whether a sudden increase in lineages in the tree was due to increased sampling/testing or increased infection. In macroevolutionary applications, there are fewer outside sources of information to incorporate. If we have paleontological information about fossilization rates and what time periods are intensely targeted for fossil recovery, we may attempt similar integration, but the fossil record is sparse and there is no analog to other phylodynamic data streams like case counts. Regardless of field, there are many challenges in integrating these data sources, including how to specify the relationships between them. The biggest challenge, though, may lie in making this data integration easy, so that it can be used in practice by researchers.

A different, and complementary, approach is to build additional structure into the birth-death models. In phylodynamic contexts, even for a novel disease, we have a very good idea about the range in which R_e could plausibly sit, based on all the other diseases studied. Once it is known what kind of disease is at hand, this can be refined further; rhinoviruses do not spread nearly as rapidly as measles, for example. To build this information into models, we can modify the Markov random field models of Chapter 3 and make them have Ornstein-

Uhlenbeck-like properties, where there is a pull to an overall average value. An Ornstein-Uhlenbeck model, similar to the random walk model, has been explored in phylodynamic contexts [33] to good effect, but it has not been widely adopted. Given the good performance of the horseshoe Markov random field for birth-death inference, an Ornstein-Uhlenbeck-like extension to the horseshoe Markov random field seems quite promising. As with integrating other time-series information into models, integrating this kind of prior information into macroevolutionary birth-death processes may prove more difficult, as we have a weaker sense of what the rates should look like. However, as paleontologists, phylogeneticists, and population geneticists work together to unravel evolutionary processes, we may well begin to develop a clearer picture that will help us estimate these rates.

Integrating outside sources of information will enable researchers to get increasingly well-resolved pictures of how birth and death rates change through time. However, this quest for fine-scale resolution will bring new challenges. In the phylodynamic context, sampling dates are sometimes censored at the level of years or months, which can cause problems when trying to obtain resolution on the scale of months or weeks by creating artificial increases in the sampling rate. The obvious solution is to integrate out these sampling times, but we are left with two painful choices. On one hand, we could do it with MCMC, but this greatly increases the number of parameters in the model. On the other hand, any attempt at an analytical solution (or numerical approximation thereof) is a multivariate integral of a gnarly likelihood function. It may be possible to approximate an analytical solution by assuming independence of the sampling ages, though it remains to be seen if this can be done quickly and how well it works. Macroevolutionary issues with the sampling rate are many and varied, if somewhat different, and are being actively investigated [*e.g.* 138]

Additional problems will likely arise as phylodynamic datasets continue to grow in size. As we sample the infections more densely, we may run afoul of model violations (or perhaps simply be able to notice them where we currently do not). Phylodynamic inference often assumes a fixed death or recovery rate. In practice we may know how fast an individual recovers absent detection and treatment. We can account for some variability in the recovery

rate, $\delta(t)$, using what is sometimes called the treatment parameter, $r(t)$. Practically, $r(t)$ is the probability that an individual who is sampled becomes noninfectious, so it encompasses everything from drug-based treatment to self-isolation. In this case, we model the recovery rate as

$$\delta(t) = \mu + \phi(t)r(t),$$

where μ is a constant death rate (the rate of recovery absent medical intervention) and $\phi(t)$ the sampling rate. However, this approach ignores the possibility of detection and medical intervention in individuals who are not sequenced, and it is not clear whether this is safe to ignore. It is possible that this can be accounted for with a minimum of extra parameters. If we assume that the rate of such unsequnced interventions is perfectly proportional to the rate of sequenced interventions, we can employ a constant $k \geq 1$ and model the recovery rate as,

$$\delta(t) = \mu + k \times \phi(t)r(t).$$

This would be equivalent to defining a time-varying death rate $\mu(t)$ to be a function of the (known) rate of recovery absent intervention, μ_0, a different constant $k \geq 0$, and the rates of treatment and sampling,

$$\mu(t) = \mu_0 + k \times \phi(t)r(t).$$

and then defining the recovery rate as $\delta(t) = \mu(t) + \phi(t)r(t)$. A relatively simple test of whether this matters would be to analyze a few datasets and use a reversible-jump mixture model [54] to assess the posterior probability that $k = 0$. The identifiability of this model bears consideration [85].

As larger phylodynamic datasets grow more common, as with the incredibly dense sampling of the SARS-CoV-2 virus during the COVID-19 pandemic, we may also begin to note more dynamics of the infectious process. Delays between the time of infection and the time of becoming infectious (the incubation period), heterogeneity in transmission among hosts, and heterogeneity in transmissibility among different viral lineages are all features discussed

frequently in the SARS-CoV-2 literature. These issues could be addressed by adding lineage-specific rate variation to the birth-death models, but this is a large amount of complexity and finding a way to balance necessary realism against computational reality will not be easy.

5.2 Tree models and tree effective sample size

In my last research chapter, I departed from making inferences using trees, and revisited the question of inferring trees to ask how we can characterize Monte Carlo error in our estimated phylogenies. I, and my co-authors, found that there are several effective sample size (ESS) measures that can capture the Monte Carlo error in ways that matter if a researcher cares about the estimated tree itself. However, this tree ESS work focused entirely on unrooted phylogenies. As we have seen, time trees are of broad interest in macroevolution and epidemiology, leaving open the question of extending our methods to time trees. Also, as covered in Chapter 3, there are cases like estimating $R_e(t)$ for HIV for which we only care about the tree so that we can infer the parameters of a birth-death model. This raises an even more interesting question: does (and if so, how does) this notion of a tree ESS effect inference in the regime where one marginalizes over the tree? Let us speculate and wave our hands a bit.

On the matter of extending the tree ESS to time trees, there are a number of practical considerations. The biggest question is how to incorporate node age information into the ESS. For unrooted trees, we can incorporate branch lengths by using distances in BHV space, but time-calibrated trees live in a more complex and constrained space, and the geometry can be much more difficult even ignoring the continuous nature of time [20]. Moreover, the same question persists here as in unrooted trees: does moving to a distance on trees and branch lengths obscure topological differences? The answer is unclear in both cases, and this will need to be investigated. On a simpler note, it is possible that our purely topological distance approaches will work relatively well for time-calibrated trees. Nominally, one should consider clades rather than splits (in the 4-taxon case, with clades we care about both groupings AB

and CD whereas with splits we only care that AB splits from CD). However, unless there is considerable uncertainty in the rooting of the tree, the mapping between clades and splits should be nearly one-to-one and split-based approaches should be similar to clade-based approaches. The upshot here is that all our work used RF distances (which compare splits) or splits directly, so it may work well out-of-the-box for time trees.

The deeper question of how tree ESS matters when the tree is data (of interest only as a way to infer parameters of the tree-generating model) and we marginalize it out is somewhat more challenging. Using MCMC to integrate out a parameter requires that you've sampled it well enough to integrate over it, and so we should expect we need a good effective sample size. The question is, a good effective sample size *of what*? Many popular birth-death models, and all of the ones we looked at in this thesis, are purely time-varying models where the only information that matters is the divergence times. In this case, it may be good enough to track the sorted divergence times alone and make sure the lowest univariate ESS is good. But, preliminary analyses reveal evidence that the accuracy of the Bayes Factor support for the big jumps in $R_e(t)$ in the Ukrainian HIV dataset of Chapter 3 is related to purely topological ESS measures. More testing is clearly needed, as is a good framework for addressing the questions, as one must find a way to properly hold things constant and ask if, accounting for all other important features, the ESS of a specific feature (the topology, the divergence times) makes a difference.

Lastly, I note that there is some less-than-glamorous maintenance work that needs to be done on these tree ESS measures. All our testing was done with samples of 1,000 trees. With patience, one could probably run 10,000. But R cannot even handle the memory requirements to try 100,000. Subsampling approaches will solve some of this problem, but integration with `C` or `C++` will likely be needed for bigger datasets, as we will need faster distance computations. Relatedly, tree-to-tree distances are slower for bigger trees, making it harder to scale this approach up to the large datasets people wish to analyze. Overall, it is clear that using these tree ESS methods for big datasets will require `C` or `C++` code, and possibly some clever algorithmic design.

5.3 The take-home message

In this thesis, I have proposed new generative models for phylogenetic trees and new ways to assess how well our MCMC procedures sample trees. Despite known challenges with birth-death models, I believe they will be important in years to come, and so I hope that the work of these chapters is helpful for others to build upon, a next small step in decades of work on the birth-death processes. In Chapter 2, we use a Bayesian workflow that I am proud of, starting with reasonable priors, continuing with interrogating the MCMC output for convergence problems, and finally performing a number of posterior predictive model checks and post-hoc simulations to understand the quality of inference. In Chapter 3, I import Bayesian nonparametric techniques which have been quite successful in other contexts to time-varying birth-death models, building priors that match our biological intuition that birth and death rates should be correlated through time, and rigorously testing those priors in simulations. The matter of how well our MCMC procedures sample the tree is one that has proven elusive and difficult to answer for many years, despite a number of attempts. I hope that my work here helps ground future discussion in sound statistical thinking, regardless of whether the methods we propose gain any traction. If there is any overall theme to this thesis (other than "I like trees"), I think it is this: statistical rigor is important in statistical analyses.

BIBLIOGRAPHY

[1] Michael E Alfaro, Francesco Santini, Chad Brock, Hugo Alamillo, Alex Dornburg, Daniel L Rabosky, Giorgio Carnevale, and Luke J Harmon. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences, 106(32):13410–13414, 2009.

[2] Dahiana Arcila and James C. Tyler. Mass extinction in tetraodontiform fishes linked to the Palaeocene–Eocene thermal maximum. Proceedings of the Royal Society B: Biological Sciences, 284(1866):20171771, 2017.

[3] Joëlle Barido-Sottani, Walker Pett, Joseph E. O'Reilly, and Rachel C. M. Warnock. FossilSim: An r package for simulating fossil occurrence data under mechanistic models of preservation and recovery. Methods in Ecology and Evolution, 10(6):835–840, 2019.

[4] C Verity Bennett, Paul Upchurch, Francisco J Goin, and Anjali Goswami. Deep time diversity of metatherian mammals: implications for evolutionary history and fossil-record quality. Paleobiology, 44(2):171–198, 2018.

[5] Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics, 27(4):733–767, 2001.

[6] Jonathan P. Bollback. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution, 19(7):1171–1180, 2002.

[7] Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 15(4):e1006650, 2019.

[8] Ian G Brennan and Paul M Oliver. Data from: Mass turnover and recovery dynamics of a diverse australian continental radiation. Dryad Digital Repository, 2017.

[9] Ian G Brennan and Paul M Oliver. Mass turnover and recovery dynamics of a diverse Australian continental radiation. Evolution, 71(5):1352–1365, 2017.

[10] Christopher A. Brochu. Phylogenetic approaches toward crocodylian history. Annual Review of Earth and Planetary Sciences, 31(1):357–397, 2003.

[11] Mario Bronzati, Felipe C. Montefeltro, and Max C. Langer. Diversification events and the effects of mass extinctions on Crocodyliformes evolutionary history. Royal Society open science, 2(5):140385, 2015.

[12] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4):434–455, 1998.

[13] Daniel G Brown and Megan Owen. Mean and variance of phylogenetic trees. Systematic Biology, 69(1):139–154, 2020.

[14] Jeremy M Brown. Predictive approaches to assessing the fit of evolutionary models. Systematic Biology, 63(3):289–292, 2014.

[15] Joseph W Brown and Stephen A Smith. The past sure is tense: on interpreting phylogenetic divergence time estimates. Systematic Biology, 67(2):340–353, 2017.

[16] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. Statistical science, pages 101–117, 2001.

[17] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480, 2010.

[18] Ana Catalán, Adriana D. Briscoe, and Sebastian Höhna. Drift and directional selection are the evolutionary forces driving gene expression divergence in eye and brain tissue of Heliconius butterflies. Genetics, 213(2):581–594, 2019.

[19] Donald H Colless. Review of Phylogenetics: Theory and Practice of Phylogenetic Systematics. Systematic Zoology, 31(1):100–104, 1982.

[20] Lena Collienne, Kieran Elmes, Mareike Fischer, David Bryant, and Alex Gavryushkin. Geometry of ranked nearest neighbour interchange space of phylogenetic trees. BioRxiv, 2019.

[21] Fabien L Condamine, Jonathan Rolland, Sebastian Höhna, Felix AH Sperling, and Isabel Sanmartín. Testing the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo butterflies. Systematic Biology, 67(6):940–964, 2018.

[22] Samantha R Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for Bayesian models using posterior quantiles. Journal of Computational and Graphical Statistics, 15(3):675–692, 2006.

[23] Michael D Crisp and Lyn G Cook. Explosive radiation or cryptic mass extinction? interpreting signatures in molecular phylogenies. Evolution, 63(9):2257–2265, 2009.

[24] Zoltan Csiki-Sava, Eric Buffetaut, Attila Ősi, Xabier Pereda-Suberbiola, and Stephen L. Brusatte. Island life in the Cretaceous-faunal composition, biogeography, evolution, and extinction of land-living vertebrates on the Late Cretaceous European archipelago. ZooKeys, (469):1–161, 2015.

[25] Victoria Culshaw, Tanja Stadler, and Isabel Sanmartín. Exploring the power of Bayesian birth-death skyline models to detect mass extinction events from phylogenies with only extant taxa. Evolution, 73(6):1133–1150, 2019.

[26] Natalie Cusimano, Tanja Stadler, and Susanne S Renner. A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. Systematic Biology, 61(5):785–792, 2012.

[27] Jack DeHovitz, Anneli Uuskula, and Nabila El-Bassel. The HIV epidemic in Eastern Europe and Central Asia. Current HIV/AIDS Reports, 11(2):168–176, 2014.

[28] O Denisiuk, P Smyrnov, AMV Kumar, S Achanta, K Boyko, M Khogali, B Naik, and R Zachariah. Sex, drugs and prisons: HIV prevention strategies for over 190,000 clients in Ukraine. Public Health Action, 4(2):96–101, 2014.

[29] Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. PLoS Biology, 4(5):e88, 2006.

[30] Alexei J Drummond, Andrew Rambaut, Beth Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution, 22(5):1185–1192, 2005.

[31] Alexei J Drummond and Marc A Suchard. Bayesian random local clocks, or one rate to rule them all. BMC Biology, 8(1):114, 2010.

[32] Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29(8):1969–1973, 2012.

[33] Louis Du Plessis. Understanding the spread and adaptation of infectious diseases using genomic sequencing data. Ph.d. thesis, ETH Zurich, 2016.

[34] Paromita Dubey and Hans-Georg Müller. Fréchet analysis of variance for random objects. Biometrika, 106(4):803–821, 2019.

[35] Sebastian Duchene, Remco Bouckaert, David A Duchene, Tanja Stadler, and Alexei J

Drummond. Phylodynamic model adequacy using posterior predictive simulations. Systematic Biology, 68(2):358–364, 2019.

[36] Kostyantyn Dumchev, Yana Sazonova, Tetiana Salyuk, and Olga Varetska. Trends in HIV prevalence among people injecting drugs, men having sex with men, and female sex workers in Ukraine. International journal of STD & AIDS, 29(13):1337–1344, 2018.

[37] Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society of London B: Biological Sciences, 279:1300–1309, 2012.

[38] Rampal S Etienne, Alex L Pigot, and Albert B Phillimore. How reliably can we infer diversity-dependent diversification from phylogenies? Methods in Ecology and Evolution, 7(9):1092–1099, 2016.

[39] Luiza Guimarães Fabreti and Sebastian Höhna. Convergence assessment for Bayesian phylogenetics analysis using MCMC simulation. TBD, 2021.

[40] Yu Fan, Rui Wu, Ming-Hui Chen, Lynn Kuo, and Paul O Lewis. Choosing among partition models in Bayesian phylogenetics. Molecular Biology and Evolution, 28(1):523–532, 2011.

[41] Federico Fanti, Tetsuto Miyashita, Luigi Cantelli, Fawsi Mnasri, Jihed Dridi, Michela Contessi, and Andrea Cau. The largest thalattosuchian (Crocodylomorpha) supports teleosaurid survival across the Jurassic-Cretaceous boundary. Cretaceous Research, 61:263–274, 2016.

[42] Conor P Farrington and Heather J Whitaker. Estimation of effective reproduction numbers for infectious diseases using serological survey data. Biostatistics, 4(4):621–632, 2003.

[43] James R Faulkner, Andrew F Magee, Beth Shapiro, and Vladimir N Minin. Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. Biometrics, 76(3):677–690, 2020.

[44] James R Faulkner and Vladimir N Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. Bayesian Analysis, 13(1):225, 2018.

[45] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. Journal of Molecular Evolution, 17(6):368–376, 1981.

[46] Kent L. Fiala and Robert R. Sokal. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution, 39(3):609–622, 1985.

[47] James M Flegal, Murali Haran, and Galin L Jones. Markov chain Monte Carlo: Can we trust the third significant figure? Statistical Science, pages 250–260, 2008.

[48] William A. Freyman and Sebastian Höhna. Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. Systematic Biology, 67(2):1995–215, 2018.

[49] Alexandra Gavryushkina, David Welch, Tanja Stadler, and Alexei J Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Computational Biology, 10(12):e1003919, 2014.

[50] Ester Gaya, Benjamin D Redelings, Pere Navarro-Rosinés, Xavier Llimona, Miquel De Cáceres, and François Lutzoni. Align or not to align? resolving species complexes within the Caloplaca saxicola group as a case study. Mycologia, 103(2):361–378, 2011.

[51] Tanja Gernhard. The conditioned reconstructed process. Journal of Theoretical Biology, 253(4):769–778, 2008.

[52] Mandev S Gill, Philippe Lemey, Shannon N Bennett, Roman Biek, and Marc A

Suchard. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. Systematic Biology, 65(6):1041–1056, 2016.

[53] Mandev S Gill, Philippe Lemey, Nuno R Faria, Andrew Rambaut, Beth Shapiro, and Marc A Suchard. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular Biology and Evolution, 30(3):713–724, 2013.

[54] Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732, 1995.

[55] Ankit Gupta, Marc Manceau, Timothy Vaughan, Mustafa Khammash, and Tanja Stadler. The probability distribution of the reconstructed phylogenetic tree with occurrence data. bioRxiv, page 679365, 2019.

[56] Sean M Harrington, Van Wishingrad, and Robert C Thomson. Properties of markov chain monte carlo performance across many empirical alignments. Molecular Biology and Evolution, 2020.

[57] Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences, 111(29):E2957–E2966, 2014.

[58] Philip Heidelberger and Peter D Welch. A spectral method for confidence interval generation and run length control in simulations. Communications of the ACM, 24(4):233–245, 1981.

[59] Sebastian Höhna. Likelihood Inference of Non-Constant Diversification Rates with Incomplete Taxon Sampling. PLOS ONE, 9(1):e84184, 2014.

[60] Sebastian Höhna. The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. Journal of Theoretical Biology, 380:321–331, 2015.

[61] Sebastian Höhna and Alexei J Drummond. Guided tree topology proposals for Bayesian phylogenetic inference. Systematic Biology, 61(1):1–11, 2012.

[62] Sebastian Höhna, Michael J Landis, and Tracy A Heath. Phylogenetic inference using RevBayes. Current Protocols in Bioinformatics, 57:6–16, 2017.

[63] Sebastian Höhna, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology, 65(4):726–736, 2016.

[64] Sebastian Höhna, Michael R. May, and Brian R. Moore. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics, 32(5):789–791, 2016.

[65] Sebastian Höhna, Tanja Stadler, Fredrik Ronquist, and Tom Britton. Inferring speciation and extinction rates under different sampling schemes. Molecular Biology and Evolution, 28(9):2577–2589, 2011.

[66] John P. Huelsenbeck and Bruce Rannala. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. Systematic Biology, 53(6):904–913, 2004.

[67] John P. Huelsenbeck, Fredrik Ronquist, et al. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics, 17(8):754–755, 2001.

[68] Walter Jetz, Gavin H Thomas, Jeffrey B Joy, Klaas Hartmann, and Arne O Mooers. The global diversity of birds in space and time. Nature, 491(7424):444, 2012.

[69] Galin L Jones. On the markov chain central limit theorem. Probability surveys, 1:299–320, 2004.

[70] Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain Monte Carlo in practice: a roundtable discussion. The American Statistician, 52(2):93–100, 1998.

[71] Robert E Kass and Adrian E Raftery. Bayes factors. Journal of the American Statistical Association, 90(430):773–795, 1995.

[72] David G Kendall et al. On the generalized "birth-and-death" process. The Annals of Mathematical Statistics, 19(1):1–15, 1948.

[73] John Frank Charles Kingman. The coalescent. Stochastic Processes and their Applications, 13(3):235–248, 1982.

[74] Denise Kühnert, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. Journal of the Royal Society Interface, 11(94):20131106, 2014.

[75] Clemens Lakner, Paul Van Der Mark, John P Huelsenbeck, Bret Larget, and Fredrik Ronquist. Efficiency of markov chain monte carlo tree proposals in Bayesian phylogenetics. Systematic Biology, 57(1):86–103, 2008.

[76] Robert Lanfear, Brett Calcott, Simon YW Ho, and Stephane Guindon. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, 29(6):1695–1701, 2012.

[77] Robert Lanfear, Xia Hua, and Dan L Warren. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome Biology and Evolution, 8(8):2319–2332, 2016.

[78] Bret Larget. The estimation of tree posterior probabilities using conditional clade probability distributions. Systematic Biology, 62(4):501–511, 2013.

[79] Anders Larsson. Aliview: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics, 30(22):3276–3278, 2014.

[80] Nicolas Lartillot and Hervé Philippe. Computing Bayes factors using thermodynamic integration. Systematic Biology, 55(2):195–207, 2006.

[81] Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testi Cambridge University Press, 2009.

[82] Jeffrey S Levinton. A theory of diversity equilibrium and morphological evolution. Science, 204(4390):335–336, 1979.

[83] Jun S Liu. Monte Carlo strategies in scientific computing. Springer Science & Business Media, 2008.

[84] Ira M Longini Jr, W Scott Clark, Robert H Byers, John W Ward, William W Darrow, George F Lemp, and Herbert W Hethcote. Statistical analysis of the stages of HIV infection using a Markov model. Statistics in Medicine, 8(7):831–843, 1989.

[85] Stilianos Louca, Angela McLaughlin, Ailene MacPherson, Jeffrey B Joy, and Matthew W Pennell. Fundamental identifiability limits in molecular epidemiology. bioRxiv, 2021.

[86] Stilianos Louca and Matthew W. Pennell. Extant timetrees are consistent with a myriad of diversification histories. Nature, pages 1–4, 2020.

[87] Ailene MacPherson, Stilianos Louca, Angela McLaughlin, Jeffrey B Joy, and Matthew W Pennell. A general birth-death-sampling model for epidemiology and macroevolution. bioRxiv, 2020.

[88] Wayne P Maddison, Peter E Midford, and Sarah P Otto. Estimating a binary character's effect on speciation and extinction. Systematic Biology, 56(5):701–710, 2007.

[89] Susana Magallon and Michael J Sanderson. Absolute diversification rates in angiosperm clades. Evolution, 55(9):1762–1780, 2001.

[90] Andrew F. Magee, Sebastian Höhna, Tetyana I. Vasylyeva, Adam D. Leaché, and Vladimir N. Minin. Locally adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. PLoS Computational Biology, 16(10):e1007999, 2020.

[91] Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. IEEE Signal Processing Letters, 23(1):179–182, 2016.

[92] Philip D. Mannion, Roger B. J. Benson, Matthew T. Carrano, Jonathan P. Tennant, Jack Judd, and Richard J. Butler. Climate constrains the evolutionary history and biodiversity of crocodylians. Nature communications, 6(1):1–9, 2015.

[93] Timothy Margush and Fred R McMorris. Consensus n-trees. Bulletin of Mathematical Biology, 43(2):239–244, 1981.

[94] Paul J. Markwick. Crocodilian diversity in space and time: the role of climate in paleoecology and its implication for understanding K/T extinctions. Paleobiology, 24(4):470–497, 1998.

[95] Michael R May, Sebastian Höhna, and Brian R Moore. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. Methods in Ecology and Evolution, 7(8):947–959, 2016.

[96] Xavier Meyer. Adaptive tree proposals for Bayesian phylogenetic inference. BioRxiv, page 783597, 2019.

[97] Vladimir N Minin, Erik W Bloomquist, and Marc A Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution, 25(7):1459–1471, 2008.

[98] Daniel Moen and Hélène Morlon. Why does diversification slow down? Trends in Ecology & Evolution, 29(4):190–197, 2014.

[99] Hélène Morlon. Phylogenetic approaches for studying diversification. Ecology letters, 17(4):508–525, 2014.

[100] Hélène Morlon, Todd L. Parsons, and Joshua B. Plotkin. Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences, 108(39):16327–16332, 2011.

[101] Elchanan Mossel and Eric Vigoda. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science, 309(5744):2207–2209, 2005.

[102] Iain Murray, Ryan P Adams, and David JC MacKay. Elliptical slice sampling. In AISTATS, volume 13, pages 541–548, 2010.

[103] Radford M Neal. Probabilistic inference using Markov chain Monte Carlo methods. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.

[104] Sean Nee. Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics, 37:1–17, 2006.

[105] Sean Nee, Robert M May, and Paul H Harvey. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 344(1309):305–311, 1994.

[106] Jamie R. Oaks. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. Evolution: International Journal of Organic Evolution, 65(11):3285–3297, 2011.

[107] Jamie R. Oaks. Full Bayesian comparative phylogeography from genomic data. Systematic Biology, 68(3):371–395, 2019.

[108] Brian C. O'Meara. Evolutionary Inferences from Phylogenies: A Review of Methods. Annual Review of Ecology, Evolution, and Systematics, 43(1):267–285, 2012.

[109] Emmanuel Paradis. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. Proceedings of the Royal Society of London B: Biological Sciences, 264(1385):1141–1147, 1997.

[110] Albert B Phillimore and Trevor D Price. Density-dependent cladogenesis in birds. PLoS Biology, 6(3):e71, 03 2008.

[111] Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pages 905–913. AISTATS, 2017.

[112] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for MCMC. R News, 6(1):7–11, 2006.

[113] Dimitris N Politis. The impact of bootstrap methods on time series analysis. Statistical science, pages 219–230, 2003.

[114] Eduardo Puértolas-Pascual, Alejandro Blanco, Christopher A. Brochu, and José Ignacio Canudo. Review of the Late Cretaceous-early Paleogene crocodylomorphs of Europe: extinction patterns across the K-PG boundary. Cretaceous Research, 57:565–590, 2016.

[115] Oliver G Pybus and Paul H Harvey. Testing macro–evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London B: Biological Sciences, 267(1459):2267–2272, 2000.

[116] Oliver G Pybus, Andrew Rambaut, and Paul H Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics, 155(3):1429–1437, 2000.

[117] Robert Alexander Pyron and Frank T Burbrink. Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. Trends in Ecology & Evolution, 28(12):729–736, 2013.

[118] Tiago B Quental and Charles R. Marshall. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution, 25:434–441, 2010.

[119] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[120] Daniel L Rabosky. Likelihood methods for detecting temporal shifts in diversification rates. Evolution, 60(6):1152–1164, 2006.

[121] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Systematic Biology, 67(5):901, 2018.

[122] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. Mathematical Biosciences, 53(1-2):131–147, 1981.

[123] F. James Rohlf, W. S. Chang, R. R. Sokal, and Junhyong Kim. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. Evolution, 44(6):1671–1684, 1990.

[124] Fredrik Ronquist, Seraina Klopfstein, Lars Vilhelmsen, Susanne Schulmeister, Debra L Murray, and Alexandr P Rasnitsyn. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. Systematic Biology, 61(6):973–999, 2012.

[125] Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61(3):539–542, 2012.

[126] Havard Rue and Leonhard Held. Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC, 2005.

[127] Daniel P Scantlebury. Diversification rates have declined in the malagasy herpetofauna. Proceedings of the Royal Society B: Biological Sciences, 280(1766):20131109, 2013.

[128] T. M. Scheyer, Oscar A. Aguilera, Massimo Delfino, D. C. Fortier, Alfredo A. Carlini, R. Sánchez, J. D. Carrillo-Briceño, L. Quiroz, and M. R. Sánchez-Villagra. Crocodylian diversity peak and extinction in the late Cenozoic of the northern Neotropics. Nature communications, 4(1):1–9, 2013.

[129] Charles Semple, Mike Steel, et al. Phylogenetics, volume 24. Oxford University Press on Demand, 2003.

[130] J John Sepkoski Jr. A kinetic model of Phanerozoic taxonomic diversity I. Analysis of marine orders. Paleobiology, pages 223–251, 1978.

[131] Daniele Silvestro, Jan Schnitzler, Lee Hsiang Liow, Alexandre Antonelli, and Nicolas Salamin. Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. Systematic Biology, 63(3):349–367, 2014.

[132] Daniele Silvestro, Marcelo F Tejedor, Martha L Serrano-Serrano, Oriane Loiseau, Victor Rossier, Jonathan Rolland, Alexander Zizka, Sebastian Höhna, Alexandre Antonelli, and Nicolas Salamin. Early arrival and climatically-linked geographic expansion of New World monkeys from tiny African ancestors. Systematic Biology, 68(1):78–92, 2019.

[133] Sigrunn Holbek Sørbye and Håvard Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. Spatial Statistics, 8:39–51, 2014.

[134] Tanja Stadler. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. Journal of Theoretical Bbiology, 261(1):58–66, 2009.

[135] Tanja Stadler. Sampling-through-time in birth-death trees. Journal of Theoretical Biology, 267(3):396–404, 2010.

[136] Tanja Stadler. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences, 108(15):6187–6192, 2011.

[137] Tanja Stadler. How can we improve accuracy of macroevolutionary rate estimates? Systematic Biology, 62(2):321–329, 2013.

[138] Tanja Stadler, Alexandra Gavryushkina, Rachel CM Warnock, Alexei J Drummond, and Tracy A Heath. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. Journal of Theoretical Biology, 447:41–55, 2018.

[139] Tanja Stadler, Roger Kouyos, Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, Beda Joos, Philip Rieder, Dong Xie, et al. Estimating the basic reproductive number from viral sequence data. Molecular Biology and Evolution, 29(1):347–357, 2012.

[140] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences, 110(1):228–233, 2013.

[141] Tanja Stadler, Denise Kühnert, David A Rasmussen, and Louis du Plessis. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. PLoS Currents, 6, 2014.

[142] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9):1312–1313, 2014.

[143] Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. Virus Evolution, 4(1):vey016, 2018.

[144] Marc A Suchard, Robert E Weiss, Janet S Sinsheimer, Karin S Dorman, Megha Patel, and Edward R B McCabe. Evolutionary similarity among genes. Journal of the American Statistical Association, 98(463):653–662, 2003.

[145] Simon Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences, 17(2):57–86, 1986.

[146] Jonathan P. Tennant, Philip D. Mannion, and Paul Upchurch. Environmental drivers of crocodyliform extinction across the Jurassic/Cretaceous transition. Proceedings of the Royal Society B: Biological Sciences, 283(1826):20152840, 2016.

[147] Jeffrey L Thorne, Hirohisa Kishino, and Ian S Painter. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution, 15(12):1647–1657, 1998.

[148] Paul Upchurch, Philipp D Mannion, Roger BJ Benson, Richard J Butler, and Matthew T Carrano. Geological and anthropogenic controls on the sampling of the terrestrial fossil record: a case study from the Dinosauria. Geological Society, London Special Publication, 358(1):209–240, 2011.

[149] Stéphanie van der Pas, Botond Szabó, Aad van der Vaart, et al. Uncertainty quantification for the horseshoe (with discussion). Bayesian Analysis, 12(4):1221–1274, 2017.

[150] Dennis Vasse and Stéphane Hua. Diversité des crocodiliens du crétacé supérieur et du paléogéne. Oryctos, 1:65–77, 1998.

[151] Tetyana I Vasylyeva, Samuel R Friedman, Jose Lourenco, Sunetra Gupta, Angelos Hatzakis, Oliver G Pybus, Aris Katzourakis, Pavlo Smyrnov, Timokratis Karamitros, Dimitrios Paraskevis, et al. Reducing HIV infection in people who inject drugs is impossible without targeting recently-infected subjects. AIDS, 30(18):2885–2890, 2016.

[152] Dootika Vats and Christina Knudson. Revisiting the gelman-Rubin diagnostic. arXiv preprint arXiv:1812.09384, 2020.

[153] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. Bayesian Analysis, 2021.

[154] Charles R Vitek, Jurja-Ivana Čakalo, Yuri V Kruglov, Konstantin V Dumchev, Tetyana O Salyuk, Ivana Božičević, Andrew L Baughman, Hilary H Spindler, Violetta A Martsynovska, Yuri V Kobyshcha, Abu S Abdul-Quader, and George W Rutherford. Slowing of the HIV epidemic in Ukraine: evidence from case reporting and key population surveys, 2005–2012. PLOS ONE, 9(9):e103657, 2014.

[155] EM Volz and SDW Frost. Scalable relaxed clock phylogenetic dating. Virus Evolution, 3(2), 2017.

[156] Hans Von Storch and Francis W Zwiers. Statistical analysis in climate research. Cambridge university press, 2001.

[157] Peter J Wagner. On the probabilities of branch durations and stratigraphic gaps in phylogenies of fossil taxa when rates of diversification and sampling vary over time. Paleobiology, 45(1):30–55, 2019.

[158] Rachel C. M. Warnock, Tracy A. Heath, and Tanja Stadler. Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. Paleobiology, 46(2):137–157, 2020.

[159] Dan L Warren, Anthony J Geneva, and Robert Lanfear. RWTY (R We There Yet): an R package for examining convergence of Bayesian phylogenetic analyses. Molecular Biology and Evolution, 34(4):1016–1020, 2017.

[160] Chris Whidden, Brian C Claywell, Thayer Fisher, Andrew F Magee, Mathieu Fourment, and Frederick A Matsen IV. Systematic exploration of the high likelihood set of phylogenetic tree topologies. Systematic Biology, 69(2):280–293, 2020.

[161] Chris Whidden and Frederick A Matsen IV. Quantifying MCMC exploration of phylogenetic tree space. Systematic Biology, 64(3):472–491, 2015.

[162] Eric W. Wilberg, Alan H. Turner, and Christopher A. Brochu. Evolutionary structure and timing of major habitat shifts in Crocodylomorpha. Scientific Reports, 9(1):514, 2019.

[163] Amy Willis. Confidence sets for phylogenetic trees. Journal of the American Statistical Association, 114(525):235–244, 2019.

[164] Chieh-Hsi Wu. Bayesian approaches to model uncertainty in phylogenetics. Ph.d. thesis, University of Auckland, 2014.

[165] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution, 39(3):306–314, 1994.

[166] Ziheng Yang and Bruce Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution, 14(7):717–724, 1997.

[167] Emile Zuckerkandl and Linus Pauling. Molecular Disease, Evolution and Genetic Heterogeneity. Academic Press, 1962.

Appendix A

APPENDIX TO: IMPACT OF K-PG MASS EXTINCTION EVENT ON CROCODYLOMORPHA INFERRED FROM PHYLOGENY OF EXTINCT AND EXTANT TAXA

A.1 Estimated diversification rates incorporating phylogenetic uncertainty and prior sensitivity

To account for phylogenetic uncertainty in diversification rate estimates, we estimate diversification rates and mass extinctions on a total of 6 distinct phylogenetic trees from [162], which we refer to as T1 to T6. Simultaneously, we investigate the sensitivity of our estimate to the prior expectation on the number of mass extinctions. Thus, we analyze each tree with a prior expectation of $\mathbb{E}(n_{\text{ME}}) = \{0.1, 0.5, 1.0, 2.0, 5.0\}$ mass extinctions. This leads to a total of 30 empirical analyses, which produce largely congruent results. In Figure A.1 we plot the estimated rates of speciation, extinction, and fossilization, which is summarized in Figure 2 of the main text.

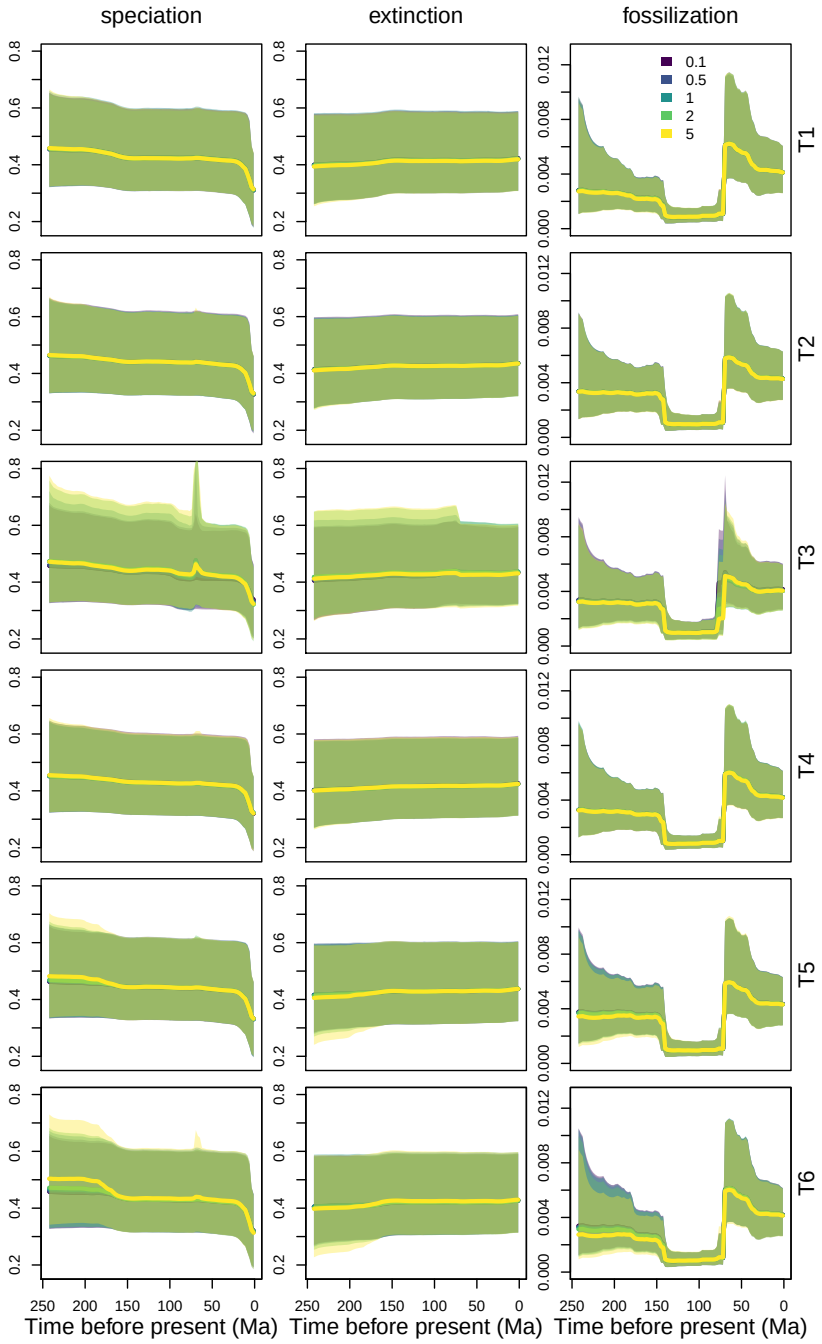


Figure A.1: Estimated rates of speciation, extinction, and fossilization through time across all datasets and all priors on the expected numbers of mass extinctions. Datasets are shown in rows, while different priors on the expected number of mass extinctions are denoted by color. Solid lines are the posterior median rates, while shaded regions are the 90% CIs. CIs for the speciation rate for the third tree extend above 0.8 and have been truncated for ease of viewing.

A.2 Additional empirical analyses

A.2.1 Extinct and extant only phylogenies

We perform two analyses of subsampled trees to examine the contribution of fossil taxa to the signature of the mass extinction. First, we analyze a subtree of tree T1 consisting of only the extant Crocodylomorph taxa. This analysis detects no signal of the K-Pg mass extinction (Figure A.3, top row), and the estimated speciation rate through time is effectively constant (Figure A.2, top row). Both speciation and extinction rates are estimated to be lower than using the combined dataset (without fossils there is no fossilization rate to be estimated). Second, we analyze a tree consisting only of extinct Crocodylomorph taxa. This analysis strongly detects the K-Pg mass extinction (Figure A.3, middle row). As with the extant-only analysis, the estimated speciation rate does not decrease towards the present, though it is otherwise similar to the diversification rate estimates obtained from the combined dataset (Figure A.2). The extinction and fossilization rates estimated are almost identical to the combined analysis. Thus, at least for the crocodylomorphs, the fossils provide the primary signal of the K-Pg mass extinction. Nevertheless, there is no harm in using a combined dataset. Diversification rates, however, do appear more sensitive to the exclusion of any taxa. Specifically, without the combined dataset it would not be possible to obtain the complete picture of historical diversification rates which includes the more ancient mass extinctions and recent decrease in speciation rate.

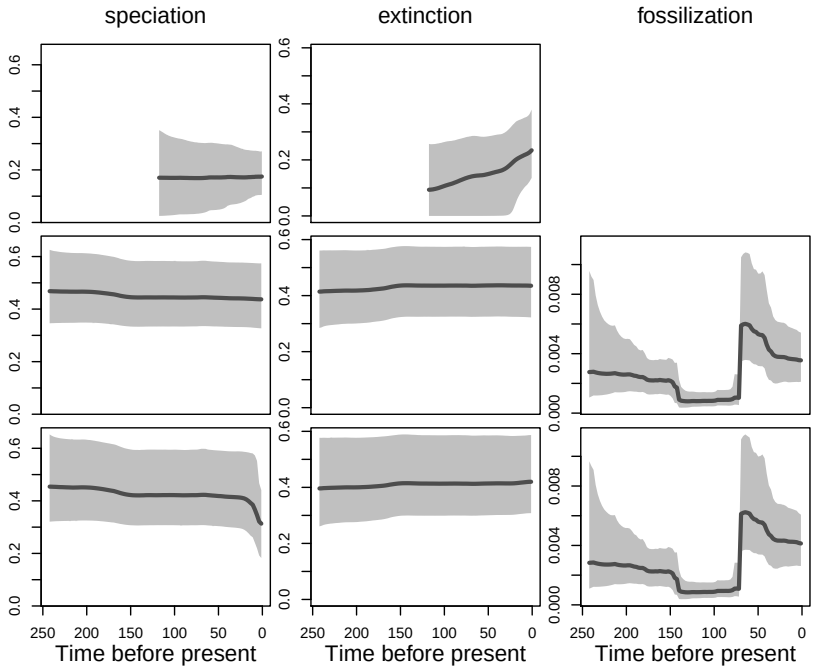


Figure A.2: Estimated rates of speciation, extinction, and fossilization through time. Top row: only extant taxa used in analysis (hence no fossilization rate). Middle row: only extinct taxa used in analysis. Bottom row: all taxa used in analysis (reproduced from our main empirical analysis).

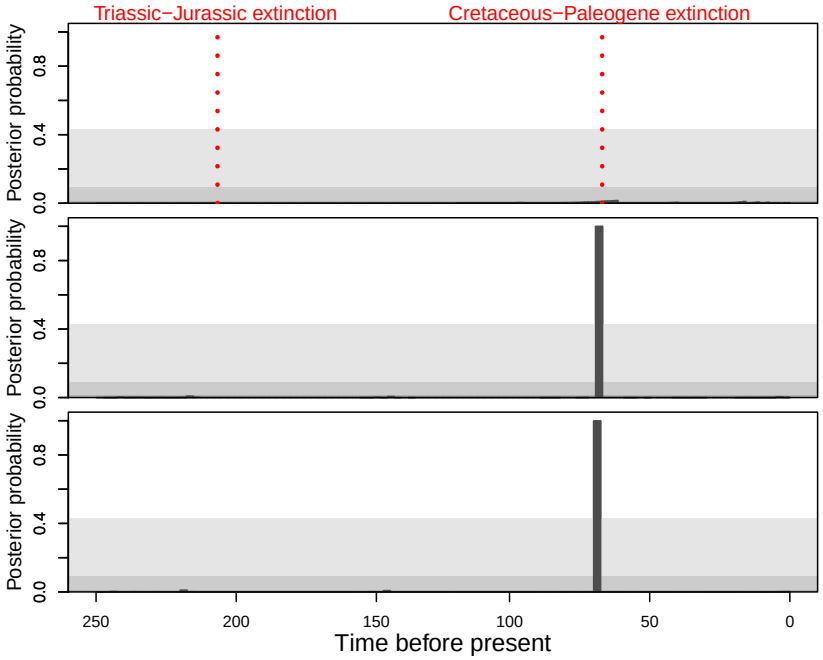


Figure A.3: Support for mass extinctions. Top row: only extant taxa use in analysis. Middle row: only extinct taxa used in analysis. Bottom row: all taxa used in analysis (reproduced from our main empirical analysis).

A.2.2 Assuming all fossils to be tips (treatment)

We used the phylodynamic treatment parameter to investigate the effect of assuming that all fossils are tips and not sampled ancestors. Specifically, we re-analyze the tree with $r_1 = r_2 = \cdots = r_{100} = 1.0$, which forces all tips to be fossils. This is not biologically meaningful, as leaving a fossil does not enforce the species to go extinct (there is no macroevolutionary equivalent to the phylodynamic treatment), but this analysis provides insight into the effects and systematic bias of forcing fossils to all be tips.

We find that the estimated signal of the K-Pg mass extinction is robust to assuming that all fossil taxa must be tips (Figure A.5). However, the estimated diversification rates are noticeably different (Figure A.4). Estimated speciation and extinction rates are much lower (by a factor of two), while the fossilization rate is overall higher (by a factor of three). The speciation rate does not display any decrease towards the present, and the extinction rate increases towards the present day. Note however, in cases with $r > 0$ that fossilization also implies the death of a lineage, and the total death rate is actually $\mu(t) + \phi(t)r(t)$, which explains why the estimate extinction rate is lower when assuming that all fossils are tips.

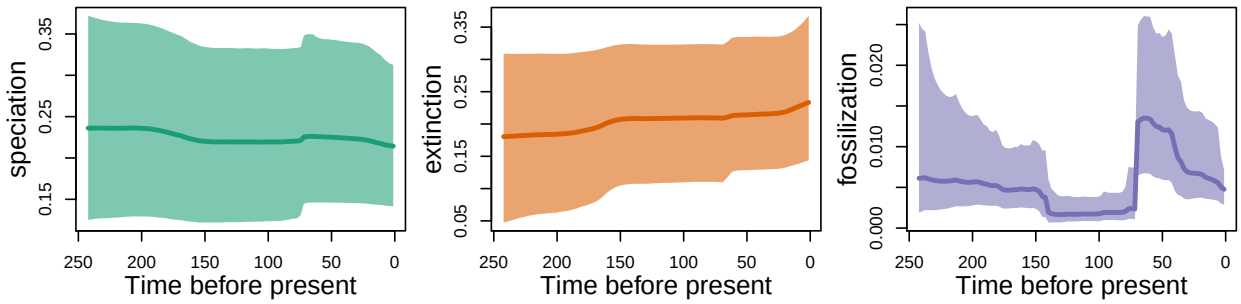


Figure A.4: Estimated rates of speciation, extinction, and fossilization through time when assuming that all fossils are tips (treatment).

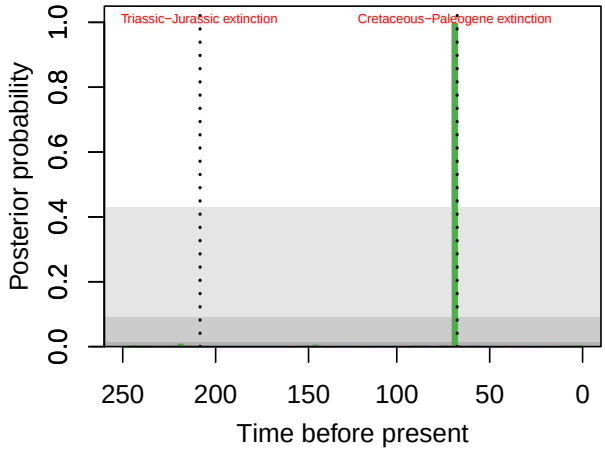


Figure A.5: Support for mass extinctions at all 99 timepoints when assuming that all fossils are tips (treatment).

A.3 Simulated data analyses

We performed a simulation study to explore the power of our crocodylomorph analysis and the false positive rate. To assess power, we simulated trees from the posterior distribution of tree T1 with a prior expected number of 0.5 mass extinctions. We ensured that all simulated datasets experienced a mass extinction at the K-Pg boundary. To do this, we set the mass extinction death probability at the K-Pg to be 0.9 for any posterior sample that had a mass extinction death probability of less than 0.5 (this affects a very small proportion of simulations, as the estimated posterior probability of the K-Pg mass extinction was 0.998).

To assess any tendency for false positives, we fit diversification rates through time for the same dataset but without the possibility of mass extinctions. Disallowing mass extinctions in the real-data inference could lead to inferred rates that produce temporal signatures that look like mass extinctions [23, 136, 95]. Trees simulated using parameter values drawn from the posterior distribution could appear to have mass extinctions and inference of simulated datasets may favor mass extinctions when there were none. Thus, this should provide a worst-case scenario on false positives. To ensure that we had sufficient resolution, we analyzed 250 trees for each scenario, and took the first 200 analyses that passed convergence cutoffs.

In the main text, we focused on the number of inferred mass extinctions per-dataset. In doing so, we used a 2 log Bayes factor threshold of 10 to determine if a mass extinction was detected or not (as previously established by [95]). This cutoff is motivated by examining the distribution of all posterior probabilities in support of mass extinctions pooled across all break times and all simulated analyses. In the "false positive" analyses without mass extinctions, approximately 1% of all (99×200) possible mass extinctions would be inferred to be significant at a Bayes factor cutoff of 2 (Figure A.6), and there are a number of mass extinctions above a cutoff of 6. Thus, to cut down on spurious inference of mass extinctions, we use a cutoff of 10 for determining support for mass extinctions. The rationale behind the rather high significance threshold is multiple testing, because we tested jointly for 99 possible mass extinctions, one per epoch.

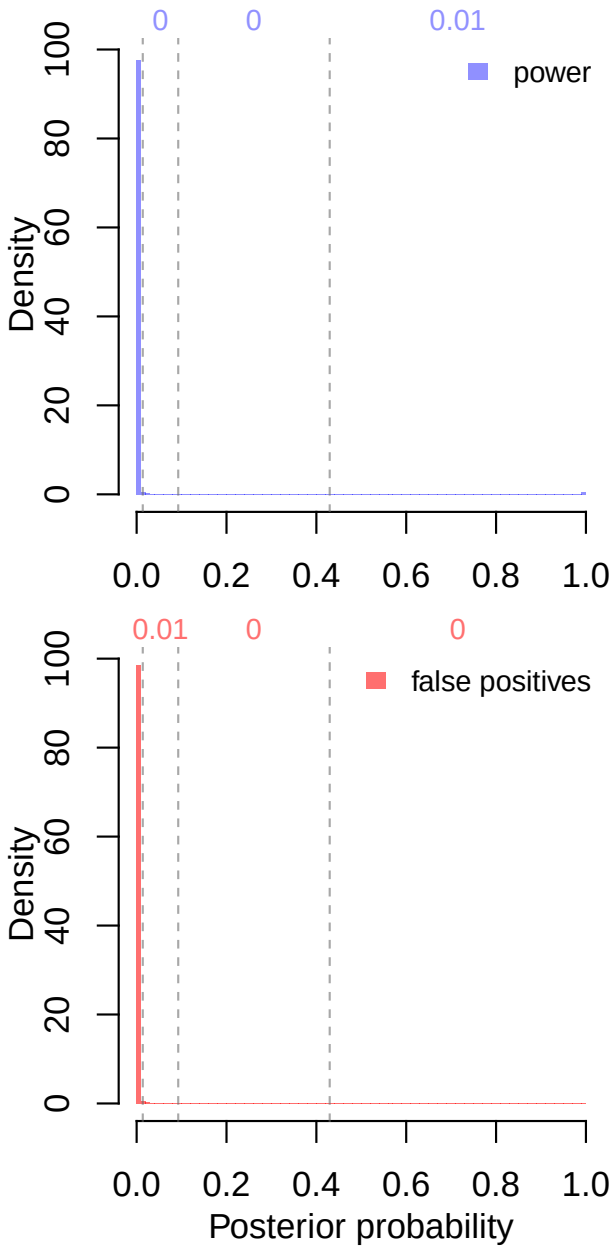


Figure A.6: The posterior probability of a mass extinction in analyses of simulated data, pooled across all 99 times at which mass extinctions were allowed and all 200 simulated datasets. Vertical lines denote 2ln Bayes Factor cutoffs of 2, 6, and 10, which correspond to weak support, support, and strong support for a mass extinction [71]. Numbers in each interval indicate the proportion of all posterior probabilities that fall in this interval, rounded to the nearest 1/100th.

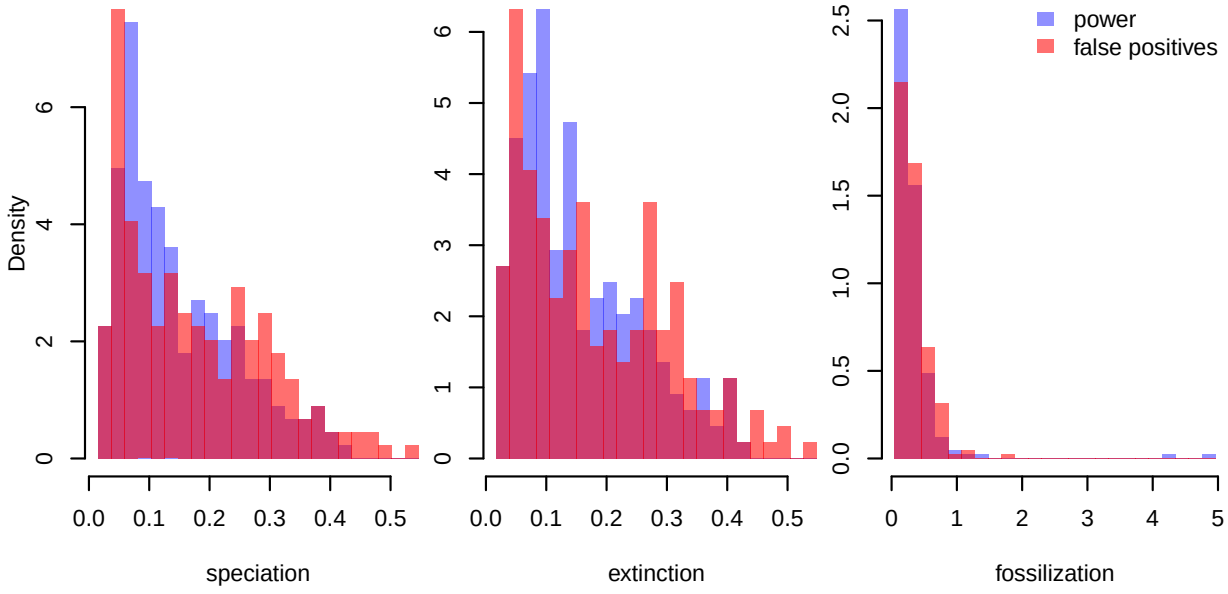


Figure A.7: Accuracy of the estimated continuous parameters in the simulations. Accuracy is measured as the mean relative absolute deviation from the true rate, such that a value of 0.1 means an average absolute relative error of 10%.

To assess the accuracy of our estimated rates through time, we use the mean relative absolute error, $1/n \sum_{i=1}^{n} [(\hat{\theta}_i - \theta_i)/\theta_i]$, where we take the posterior median as the parameter estimate. The distribution of relative errors for speciation and extinction are in general low (Figure A.7). The accuracy for speciation and extinction rate estimates is comparable to what was found in [90] in both their analyses where only speciation varied and their analyses where both speciation and extinction varied. The fossilization rate is apparently much more difficult to estimate, the average error is much higher (Figure A.7). Further, where speciation and extinction rates are generally underestimated, the fossilization rate is generally overestimated. Estimation error is larger when compound parameters like net diversification ($\lambda(t) - \mu(t)$) are considered, suggesting that we are actually estimating the rates parameterized, and not compound parameters.

A.4 Model adequacy

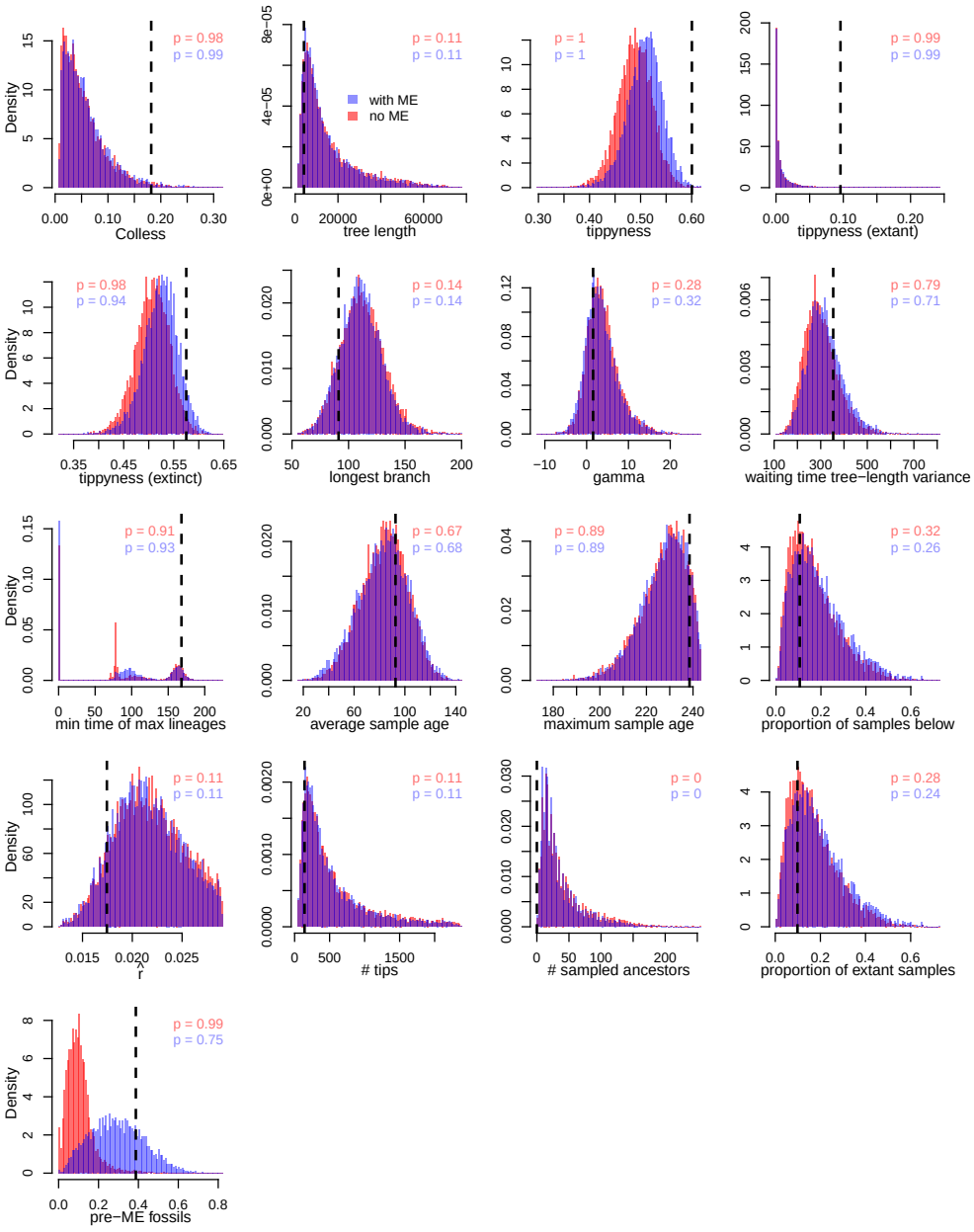


Figure A.8: Posterior predictive distributions (red and blue) for each of 17 summary statistics. The observed value is shown in black. Posterior predictive p-values are rounded to the nearest 1/100th and are represent proportion of posterior predictive values below the observed value.

To assess the adequacy of our model performance, we employ posterior predictive simulations. Specifically, we simulate trees for the same mass extinction priors (0 and 0.5) and dataset (T1) as we use for our false positive and power analyses. For each of the three converged chains, we subsample to 2500 posterior samples, for each of which we simulate one tree, for a total of approximately 7500 trees (in a few cases the simulator failed). As mentioned above, in cases where the K-Pg mass extinction probability was estimated to be less than 0.5, we set the mass extinction probability to 0.9, this affects approximately 0.2% of the simulated datasets with mass extinctions. The first 250 of these trees include the same trees for which we performed our simulated data analyses.

We employ 15 summary measures of phylogenies, many of which are standard in the literature. They are:

1. Colless' (normalized) imbalance statistic, [19]. Larger values mean trees are more imbalanced than expected under lineage-exchangeable models like the one derived in this chapter.

2. Tree length, the sum of all branch lengths in the tree.

3. Tippyness, the proportion of all branch lengths that are edges subtending a (fossil or extant) tip [46, 123].

4. Tippyness (extant), the proportion of all branch lengths that are edges subtending an extant tip. This statistic should be sensitive to misspecification of the random sampling model assumed for sampling events [46, 123].

5. Tippyness (extinct), the proportion of all branch lengths that are edges subtending a fossil tip. Combined with tippyness (extant), this statistic allows one to localize issues with tippyness.

6. Longest branch, the length of the longest branch in the tree [35]. Computed such that sampled ancestors break up branches.

7. The gamma statistic of [115], a measure of the concentration of branch lengths towards the root of the tree.

8. Waiting time tree-length variance, a measure designed to detect heterogeneity through time in the birth-death process. To compute, break the tree into intervals at every birth and sampling event. Let $\tau_i = n_i \Delta_i$ be the total tree length in time interval i, equal to the number of lineages in that interval multiplied by its duration in time. The statistic is then $\mathrm{Var}(\boldsymbol{\tau})$.

9. Minimum time of maximum lineages, the most recent time in the lineage-through-time curve which has the maximum number of lineages, [35]. The minimum ensures uniqueness over multiple modes, though it means that for trees lacking serial samples the value will always be 0.

10. Average sample age, the average age of all samples (including extant tips, fossil tips, and sampled ancestors) [35].

11. Maximum sample age, the oldest age of all samples (including extant tips, fossil tips, and sampled ancestors).

12. Proportion of samples below the youngest branching time. This measure should be sensitive to mis-estimating $\phi(t)$ relative to $\lambda(t) - \mu(t)$ in the recent past.

13. $\hat{r}$, a crude methods-of-moments estimator of the net diversification rate, [89, 90]. This measure should capture whether the number of birth events in the trees are reasonable, relative to its age.

14. The number of tips in the tree, including extant and fossil tips.

15. The number of sampled ancestors in the tree.

16. The proportion of samples which are extant samples. When there is only event-sampling at the present ($\Phi_i = 0$ for $i > 0$), this statistic should be sensitive to how well $\phi(t)$ and Φ_0 are matched.

17. The number of fossil samples occurring in the interval leading up to (the first interval older than) the mass extinction, divided by the maximum number of lineages in any time interval. Simulations suggest mass extinctions lead to a visible band of fossils just before the mass extinction. Dividing by the maximum number of lineages in the LTT curve normalizes this measure across posterior predictive trees which may vary in size quite notably.

Overall, we find that model performance is adequate. Few statistics exhibit very small or very large posterior predictive p-values. Furthermore, for a number of statistics, the mode of the posterior predictive distribution and the observed value appear to align, indicating good fit with respect to those statistics. However, the observed phylogeny has no sampled ancestors, while almost every simulated tree contains at least one sampled ancestor. This is likely at least in part an artifact of the tree building process assuming all samples are tips, which can be modeled by setting the phylodynamic recovery parameter r to 1 (see above).

Colless' imbalance provides evidence for unmodeled among-lineage variation in diversification rates, as the observed tree imbalance is larger than most of the posterior predictive tree imbalances. The tippyness family of measures indicate some issues with the sampling model. Overall, predicted tippyness is lower than the observed tippyness. Looking at tippyness restricted to both extant and fossil tips, we can see that the larger driver here is the length of branches leading to extant tips. This could be explained if the 14 extant crocodylomorphs in the tree represent a diversified sample rather than a random one [65]. The tippyness restricted to fossil tips shows less misspecification, though there is still some discrepancy between the predicted and observed values. The lack of sampled ancestors could play a role here; fossil tips must have a branch subtending them, and this means that a tree with only fossil tips and no sampled ancestors will be longer than one where some fossil samples are sampled ancestors.

A.5 Estimated diversity through time

Once fit to the data, our models can be used to make inferences about the historical diversity through time of a group. These estimated diversity through time curves can be compared to other estimates, *e.g.*from the fossil record, to further validate the results. However, note that we estimate species diversity through time and not the number of genera or families which is common in paleontological studies.

To estimate the diversity through time, for a set of posterior samples of diversification rates we simulate a complete tree, not allowing the tree to go extinct by repeating the simulation until we obtain a tree that survived until the present. Complete trees are necessary because reconstructed trees do not contain the record of all species alive at some time in the past, only those that contribute to the sample. In Figure A.9, we show the inferred number of Crocodylomorph lineages through time using this approach (for the analysis of tree T1 with $\mathbb{E}(n_{\text{ME}}) = 0.5$). Our model predicts a period of rapid growth up through the mid-to-late Jurassic, followed by a slow increase until the end Cretaceous mass extinction (K-Pg), which leaves a massive impact. Following the extinction, there is a period of slow growth into the Eocene, followed by a decline to the present.

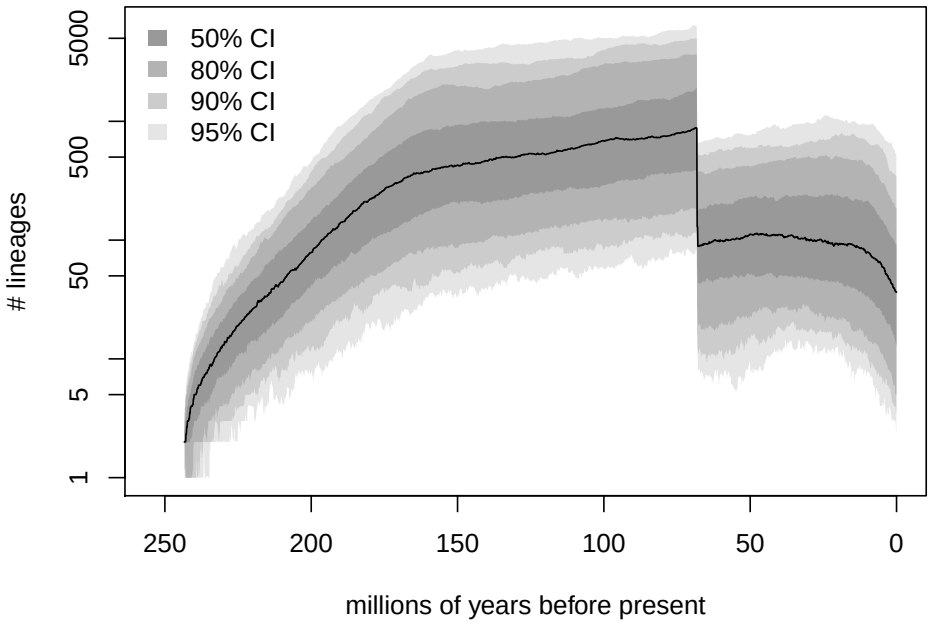


Figure A.9: Posterior predictive distribution of the crocodylomorph diversity through time, median (solid line) and 50% to 95% CIs (shaded areas). For each simulated tree, the number of lineages alive is binned over 1000 intervals and recorded. All quantiles (median and CI) are taken per-interval.

A.6 Predicted fossil counts

The diversity through time data from the previous section can also be used to predict the number of fossils found in different timespans. The estimated fossilization rates are per lineage per million years, and the diversity through time curves provide the numbers of lineages alive. In Figure A.10, we take the diversity through time curves from Figure A.9, multiply in the posterior median fossilization rate (from the same analysis), and sum over the different geological ages. This produces a curve of the predicted number of fossil lineages found in each age. When compared to the number of fossils in the tree from each age, there is strong agreement.

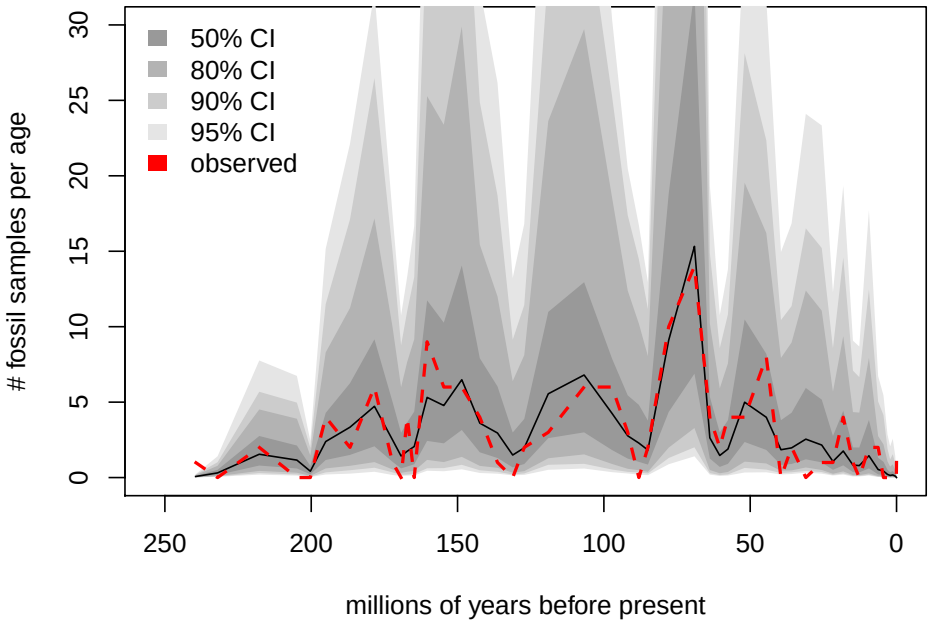


Figure A.10: Model predictions of the number of crocodylomorph fossils across geological eras (black line, grey regions) and observed fossils in the tree (red line). All predicted curves use the posterior median fossilization rate, the black line and 50% to 95% CIs are determined by the median and CIs from the diversity through time curves (Figure A.9). The red curve is from tree T1 on which the analysis is based. The y-axis is truncated to focus on the curve of observed fossil times.

A.7 Fossil tip ages

So far, we have shown that the signature of the K-Pg mass extinction is robust to (i) the choice of phylogeny (ii) the prior on the number of mass extinctions (iii) the exclusion of extant taxa (iv) the (tree inference) assumption that all taxa are tips. We have also shown that the model is largely adequate, and that any inadequacies are shared with models lacking mass extinctions. One factor that we have not addressed is the ages of the fossils. To assess robustness to fossil ages, we simulate 1000 trees based on T1 where we replace the ages with uniform draws from the stratigraphic ranges provided by [162] (rejecting any draws of ages that would produce negative branch lengths). Fossil ages are likely important to identifying mass extinctions: simulated trees with mass extinctions often exhibit a band of fossil tips just prior to a mass extinction. By drawing new ages independently, we produce a sort of worst-case scenario where this signal gets maximally eroded.

Examination of the resampled LTT curves shows that there is still a clear drop (caused by the fossil tips), though there is uncertainty about the timing and magnitude of this drop (Figure A.11). We can also compare summary statistics of these LTT curves to our posterior predictive distributions. Specifically, we can compare the number of fossils in the interval immediately prior to the observed K-Pg mass extinction to the predictive distributions from analyses with and without mass extinctions. For this comparison, we use the analysis with $\mathbb{E}(n_{\text{ME}}) = 0.5$, and we normalize the number of fossils to the peak of the LTT curve for comparability between large trees and small trees (the model for mass extinctions kills a proportion of lineages, rather than a fixed number). The resampled LTT curves show a somewhat smaller drop than in the empirical tree T1, but the drop is more in line with trees simulated with mass extinctions than simulated without (Figure A.12). Further, pooling both the interval immediately prior to the K-Pg with the next oldest recovers essentially the entire drop. Overall, these resampled datasets suggest that the signature of the K-Pg mass extinction is robust to the fossil ages.

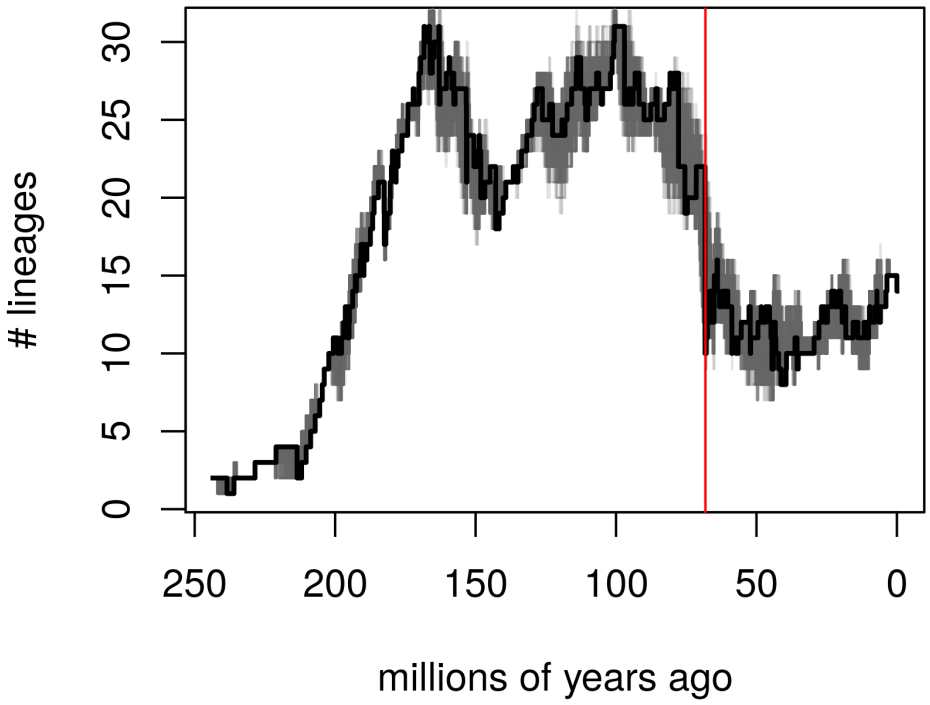


Figure A.11: The LTT curve of T1 from [162] (black), and 1000 LTT curves created by resampling the fossil times uniformly from the stratigraphic ranges. All resampled curves show a large drop around the time of the observed K-Pg mass extinction, suggesting they also contain evidence of the effect of the K-Pg.

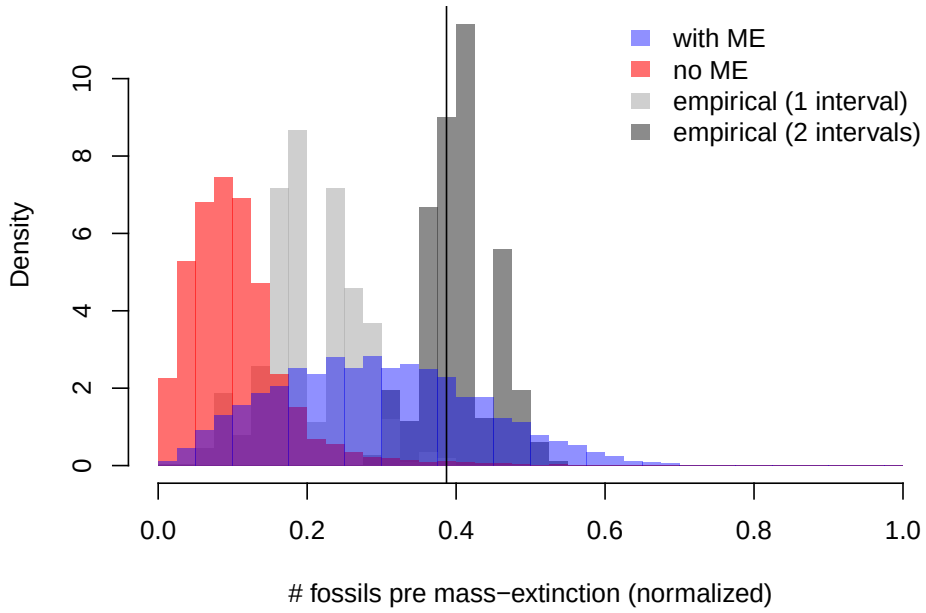


Figure A.12: Distributions of the number of fossil samples in the time shortly before the observed K-Pg mass extinction for resampled versions of tree T1. For comparison between large and small trees, we normalize this number to the peak of the LTT curve. The blue histogram is the posterior predictive distribution based on the analysis with $\mathbb{E}(n_{\text{ME}}) = 0.5$, while the red histogram is an analysis with no mass extinctions. The light grey histogram is the analog of the blue and red histograms, while the dark grey additionally includes fossils in the next oldest interval. The black line is the value in tree T1. Both resampled distributions show much larger numbers of fossil samples than expected without a mass extinction, suggesting that the signal of the K-Pg is robust to fossil times.

A.8 Inferring mass extinctions

In our model setup, we use reversible jump MCMC to ask if there was a mass extinction at some time s_i. If there is a mass extinction, the MCMC also samples from the posterior distribution of the extinction probability, M_{s_i}, equivalent to inferring the survival probability $1 - M_{s_i}$. Without biologically informed priors, single-lineage extinction rates can be confused for consistent pulses of "mass" extinctions affecting all lineages, so we employ realistic and informative priors suggesting that between 74% and 99%. Consequently, we marginalize out the inferred extinction/survival probability and focus on the question of presence/absence (though in Figure A.13 we show the estimate for one tree with $\mathbb{E}[n_{\text{ME}}] = 0.5$). This allows us to use Bayes Factors to measure the support in the data for a mass extinction at time s_i. Given a choice of the prior number of expected non-zero mass extinction events, $\mathbb{E}[n_{\text{ME}}]$, and that there are l times where a mass extinction could occur, the prior probability of a non-zero mass extinction at time s_i is $p_{\text{prior}} = \mathbb{E}[n_{\text{ME}}]/l$. Given an MCMC estimate of the posterior probability, p_{post}, the 2ln Bayes factor for a mass extinction at time s_i is,

$$2 \times \log\left(\frac{p_{\text{post}}}{1 - p_{\text{post}}} \Big/ \frac{p_{\text{prior}}}{1 - p_{\text{prior}}}\right).$$

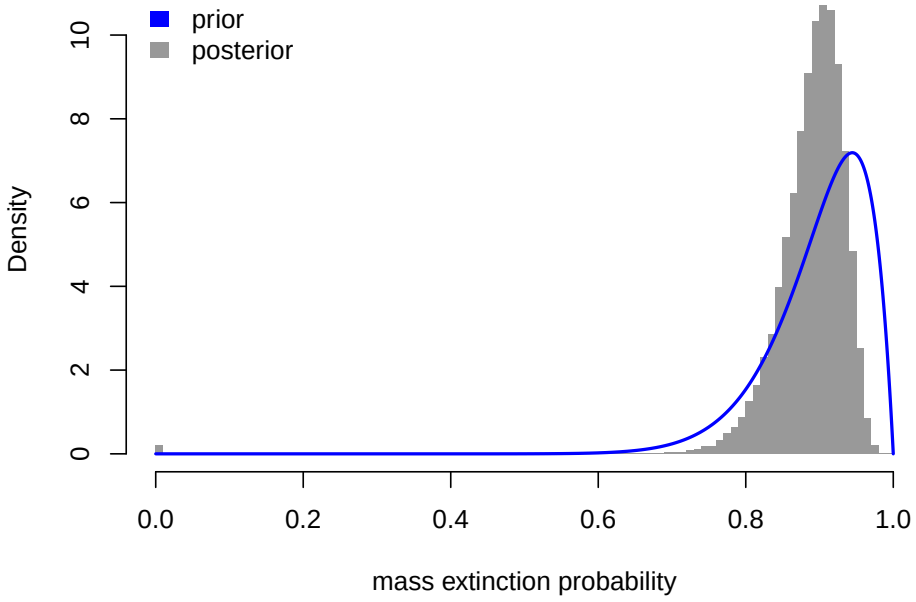


Figure A.13: Comparison of prior (blue density curve) and posterior (grey histogram) distributions on the probability of extinction at the K-Pg. The prior shown is the Beta(18,2) component of the prior, the conditional prior distribution assuming there is a mass exinction. The posterior distribution is the marginal distribution and includes a probability of ≈ 0.002 that there is no mass extinction.

A.9 Interpretation of the terms in the Likelihood of the Generalized Episodic Fossilized Birth-Death Process

In the main text we only provided a brief explanation of our likelihood function. To make the explanation easier to understand, we reproduce the likelihood function again. The probability density of a phylogenetic tree Ψ is

$$
\begin{aligned}
f(\Psi) = \quad & \frac{2^{I+H-||\mathcal{A}||-1}}{(I+H-||\mathcal{A}||)!} && (i)\\
& \times \prod_{t\in\mathcal{N}} \Big[\lambda(t)\Big] && (ii)\\
& \times \prod_{t\in\mathcal{F}} \Big[\phi(t)\big(r(t)+\big(1-r(t)\big)E(t)\big)\Big] && (iii)\\
& \times \prod_{t\in\mathcal{A}} \Big[\phi(t)\big(1-r(t)\big)\Big] && (iv)\\
& \times \prod_{i=1}^{l} \Big[\Lambda_i^{K_i}\big(2\Lambda_i E(s_i)+(1-\Lambda_i)\big)^{L(s_i)-K_i}\Big] && (v)\\
& \times \prod_{i=1}^{l} (1-M_i)^{L(s_i)} && (vi)\\
& \times \prod_{i=0}^{l} \Big[(1-\Phi_i)^{(L(s_i)-I_i)}\Phi_i^{I_i}(1-R_i)^{T_i} \\
& \qquad (R_i+(1-R_i)E(s_i))^{I_i-T_i}\Big] && (vii)\\
& \times \prod_{t\in\mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)}\Big] && (viii)
\end{aligned}
\tag{A.1}
$$

Term (i) is the probability of the topology. There are $I+H-||\mathcal{A}||$ tips (fossil samples without sampled ancestors and extant samples) which have $(I+H-||\mathcal{A}||)!$ labelings. Furthermore, there are $(I+H-||\mathcal{A}||-1)$ internal nodes which have $2^{(I+H-||\mathcal{A}||-1)}$ left-right orientations. Since we do not consider left-right orientations in phylogenetics, the probability of the tree topology is $\frac{2^{I+H-||\mathcal{A}||-1}}{(I+H-||\mathcal{A}||)!}$.

Term (ii) is the probability of the observed serial speciation in the tree (Figure 4g). Each of these happens with a probability density given by the speciation rate at that time.

Term (iii) is the probability of the serially-sampled tips (Figure 4c-d). To be a tip, the sample must have no sampled descendants, which can occur in two ways. The sampling event may be treated, which happens with probability $\phi(t)r(t)$. Alternately, the sampled lineage may not be treated, and the lineage simply has no sampled descendants, which happens with probability $(\phi(t)(1-r(t))E(t))$.

Term (iv) is the probability of the sampled ancestors (Figure 4c). We must sample the ancestor, and then it must go untreated (if it were treated, it would be a sampled tip).

Term (v) is the probability of the observed and unobserved speciation events at tree-wide speciation burst events (Figure 4e). The probability of the observed burst speciation events is $\Lambda_i^{K_i}$. The probability of the lineages without observed burst speciation events is, $(2\Lambda_i E(s_i) + (1-\Lambda_i))^{L(s_i)-K_i}$. Lineages may not have observed burst speciation events for two reasons. A lineage might experience a burst speciation, but one of its children goes unsampled (leaving one continuous lineage in the reconstructed phylogeny), which happens with probability $2\Lambda_i E(s_i)$. Alternately, the lineage may not experience a burst speciation at all, which happens with probability $(1-\Lambda_i)$. In the case that there is no burst speciation at a particular interval time s_i ($\Lambda_i = 0$) then there are no burst speciations ($K_i = 0$), and term (v) is 1.

Term (vi) is the probability of all lineages surviving a tree-wide mass extinctions events (Figure 4f). Each lineage that spans the ith mass extinction survives with probability $(1-M_i)$. We do not assume the possibility of observing any deaths at the time of the mass extinction.

Term (vii) is the probability of all the observed sampling times at given tree-wide sampling events (Figure 4g). This includes the probability of all the sampled lineages, $\Phi_i^{I_i}$, as well as the probability of all the unsampled lineages, $(1-\Phi_i)^{(L(s_i)-I_i)}$. The probability of the sampled ancestors at this time is given by, $(1-r(s_i))^{T_i}$, which is the probability that the sampled ancestors are not treated. The probability of the sampled tips is $(r(s_i)+(1-r(s_i))E(s_i))^{I_i-T_i}$,

which accounts for the possibility that the tip is treated $r(s_i)$, or that it is untreated but leaves no sampled descendants $(1 - r(s_i))E(s_i)$. In the case that there is no sampling event at a particular interval time s_i ($\Phi_i = 0$) then there are no event samples ($I_i = 0$ and $T_i = 0$), and that term in the product collapses to 1.

Term (*viii*) is the probability of the observed branch segments (Figure 5). A branch segment is a portion of a branch that is uninterrupted by an interval time or an event (speciation, extinction, or sampling). The product of all branch segments yields the total probability of all the branches of the tree.

A.10 Different Conditions of the Generalized Episodic Fossilized Birth-Death Process

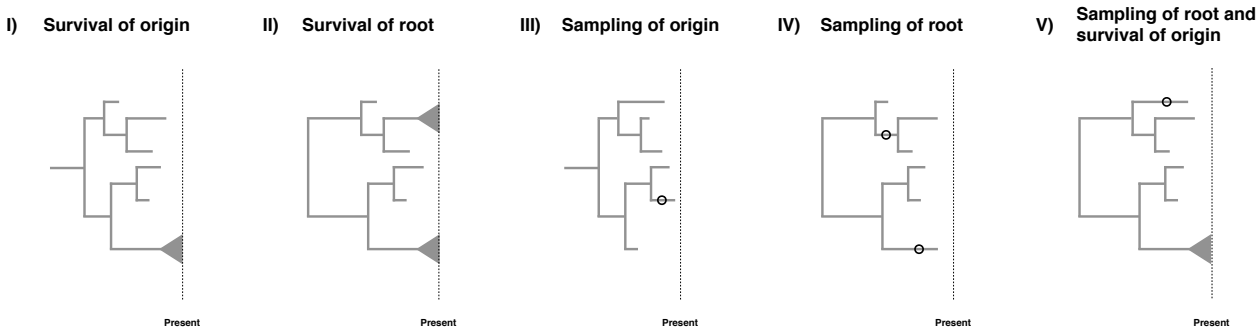


Figure A.14: Five different possible conditions for our generalized fossilized-birth-death process. I) The process survives until the present. II) The process starts at the root and both descendants of the root survive until the present. III) Sampling at least one lineage. IV) The process start at the root and both descendants have at least one lineage sampled. V) The process starts at the root, both descendants have at least one lineage sampled, and the process survives until the present.

Birth-death models are often conditioned on specific events, see [137] and [60] for some discussion on the topic. However, when there are non-contemporaneous samples in the dataset which may be ancestral to other samples, conditioning becomes somewhat complex. The key issues for conditioning are whether it is assumed that the process starts at the root or the origin, and whether the descending lineage(s) is (are) assumed to leave any sampled descendant or specifically to have a descendant sampled at the present day. Consideration of these possibilities leads to five possible conditions, though conditioning is not strictly required.

Survival of the origin We condition the process on survival of one lineage, *i.e.*at least on descendant of the lineage starting at the origin was sampled at the present. This condition represent a case when we have fossils and extant taxa and do not know if the fossils are stem fossils of the entire clade. The condition is obtained by computing $1 - E(t_{or})$ with $\phi(t) = 0$.

Survival of the root We condition the process on survival of both lineages, *i.e.*at least one descendant of each lineage starting at the root was sampled at the present. This is the case for most macroevolutionary analyses without any fossils or if the fossils are known to belong within the crown group of the extant taxa. The condition is obtained by computing $(1 - E(t_{MRCA}))^2$ with $\phi(t) = 0$.

Sampling of origin We condition the process to have at least one sample being a descendant of the origin. This is simply a minimal condition that at least something was observed/sampled. This condition represents the case if we would also consider complete extinct clades. The condition is obtained by computing $1 - E(t_{or})$.

Sampling of the root We condition the process to require that both lineages starting at the root are sampled. In this case, all taxa might be extinct but the root age is known or inferred as a parameter of the model. The condition is obtained by computing $(1 - E(t_{MRCA}))^2$.

Sampling of the root and survival of the origin We condition the process on sampling of both descendant lineages of the root and at least one sample at the present. In this case, we condition on this specific root age but one of the descendant lineages of the root might have gone extinct while the other descendant lineage from the root must have survived. The condition is obtained by computing $(1 - E_{\phi(t)=0}(t_{MRCA}))(1 - E_{\phi(t)\neq 0}(t_{MRCA}))$.

For macroevolutionary analyses of diversification rates, condition (I) is the most adequate if we have both extinct and extant taxa, condition (II) if we have only extant taxa, and condition (III) if we have only extinct taxa. For phylodynamic applications, if it can safely be assumed that there are no sampled ancestors prior to the first observed infection (which will always be true if $r(t) = 1$), condition (IV) may be used, otherwise only condition (III) is applicable. Conditioning on survival as in (I), (II), or (V) requires $\Phi_0 > 0$, and so is primarily of interest in macroevolutionary applications. Of these conditions, (II) is the strictest and

requires prior knowledge that the MRCA of the extant samples is the MRCA of all samples. Condition (V) is less restrictive, requiring only that none of the fossils could be sampled ancestors prior to the first observed speciation event, which would apply if all fossils are within the crown group. We could additionally condition on the number of extant taxa N, as suggest by [51], although there is, as of today and to our knowledge, no solution known to condition on the number of extinct taxa.

A.11 Comparison to the Gavryushkina Model

The model presented by [49] represents the most parameter-rich model prior to the model in our paper, and is in fact a special case of this model. Specifically, the Gavryushkina model is the special case of ours where there are no mass extinction events ($M_i = 0 \ \forall \ i$), no birth bursts ($\Lambda_i = 0 \ \forall \ i$), and there is no distinction in treatment probability between ϕ-sampling and Φ-sampling ($R_i = r_i \ \forall \ i$). However, the presentation of the model in [49] differs somewhat from ours, making a comparison between the two presentations useful.

For clarity, in our presentation of the Gavryushkina model, we keep our parameterization and notation. For readers looking at the original source material, our $\boldsymbol{\phi}$ is their $\boldsymbol{\psi}$ and our $\boldsymbol{\Phi}$ is their $\boldsymbol{\rho}$. Note also that there are some differences due to the choice of the direction of time. In our formulation, time for all terms flows backwards, thus $E_{i-1}(t)$ is an extinction probability for the interval preceding interval i. In the formulation of [49], the preceding extinction probability would be $E_{i+1}(t)$ (or more accurately, $p_{i+1}(t)$).

A.11.1 Terms A and B

Our term $\boldsymbol{A}$ is a generalization of that in the Gavryushkina model, in both cases we have,

$$A_i = \sqrt{(\lambda_i - \mu_i - \phi_i)^2 + 4\lambda_i\phi_i}. \tag{A.2}$$

It can be seen that the term $\boldsymbol{B}$ in the Gavryushkina model is a special case of ours. [49] define,

$$B_i = \frac{(1 - 2(1 - \Phi_i)E_{i-1})\lambda_i + \mu_i + \phi_i}{A_i}, \tag{A.3}$$

and we define instead

$$B_i = \frac{(1 - 2C_i)\lambda_i + \mu_i + \phi_i}{A_i} \tag{A.4}$$

where C_i is defined as

$$\begin{aligned} C_i =& \mathbb{I}_{(\Phi_i>0)}\big((1-\Phi_i)E_{i-1}\big) \\ &+\mathbb{I}_{(\Lambda_i>0)}\big((1-\Lambda_i)E_{i-1}(s_i)+\Lambda_i E_{i-1}^2(s_i)\big) \\ &+\mathbb{I}_{(M_i>0)}\big((1-M_i)E_{i-1}(s_i)+M_i\big) \\ &+\mathbb{I}_{(\Phi_i=0,\Lambda_i=0,M_i=0)}\big(E_{i-1}(s_i)\big). \end{aligned} \tag{A.5}$$

When there are no mass extinctions or birth bursts and $R_i = r_i$, this can be simplified to,

$$\begin{aligned} C_i =& \mathbb{I}_{(\Phi_i>0)}\big((1-\Phi_i)E_{i-1}\big) + \mathbb{I}_{(\Phi_i=0)}\big(E_{i-1}(s_i)\big) \\ =& (1-\Phi_i)E_{i-1}, \end{aligned} \tag{A.6}$$

which is the same definition as in [49].

A.11.2 Extinction Probabilities

The extinction probability terms in [49], $p_i(t)$, are a special case of our $E_i(t)$. Using our s_i to represent the more recent boundary of the ith interval, [49] define,

$$p_i(t) = \frac{\lambda_i + \mu_i + \phi_i - A_i \frac{e^{A_i(t-s_i)}(1+B_i)-(1-B_i)}{e^{A_i(t-s_i)}(1+B_i)+(1-B_i)}}{2\lambda_i}. \tag{A.7}$$

We define

$$E_i(t) = \frac{\lambda_i + \mu_i + \phi_i - A_i \frac{(1+B_i)-e^{-A_i(t-s_i)}(1-B_i)}{(1+B_i)+e^{-A_i(t-s_i)}(1-B_i)}}{2\lambda_i}. \tag{A.8}$$

When there are no mass extinction or birth burst events and $R_i = r_i$, $\boldsymbol{B}$ is the same in both models, and our definition is simply theirs where the last term on the numerator has been multiplied by

$$\frac{e^{A_i(t-s_i)}}{e^{A_i(t-s_i)}}.$$

Thus, $p_i(t)$ in [49] is a special case of the $E_i(t)$ defined in this chapter.

A.11.3 Branch Probabilities

Despite the similarities in both terms $\boldsymbol{A}$ and $\boldsymbol{B}$, and the extinction probabilities, our $D_i(t)$ has no direct equivalent simpler case in the [49] model. We define our $D_i(t)$ such that, for a branch that starts at time t_o ends at time t_y ($t_y < t_o$), the probability of observing that uninterrupted branch is $D(t_o)/D(t_y)$. [49] define a similar quantity, $q_i(t)$,

$$q_i(t) = \frac{4e^{A_i(t-s_i)}}{(e^{A_i(t-s_i)}(1+B_i)+(1-B_i))^2}. \tag{A.9}$$

We define $D_i(t)$ as

$$D_i(t) = D_{i-1}(s_i)\frac{4e^{-A_i(t-s_i)}}{((1+B_i)+e^{-A_i(t-s_i)}(1-B_i))^2}. \tag{A.10}$$

In the simpler case where there are no mass extinctions or birth bursts and $R_i = r_i$, multiplying by

$$\left(\frac{e^{A_i(t-s_i)}}{e^{A_i(t-s_i)}}\right)^2,$$

shows us that

$$q_i(t) = \frac{D_i(t)}{D_{i-1}(s_i)}.$$

In essence, where $D_i(t)$ corresponds to the probability of an unbroken lineage between time t and time 0, $q_i(t)$ track the probability of an unbroken lineage between time t and the nearest younger interval time s_i. This difference is accounted for in [49] by multiplying the likelihood by $q_{i-1}(t)^{L(s_i)-I_i}$ at every time s_i, where $L(s_i) - I_i$ is the number of lineages that are extant at the end of the interval, not counting the lineages sampled at the corresponding tree-wide event-sampling time.

A.12 Arranging terms in the likelihood

We note that our arrangement of terms in the likelihood is not the only possible option. We defined our branch segments such that they do not span multiple intervals and no birth bursts, intensive sampling events, or mass extinctions are possible. Because of this, our $D_i(t)$ reflect only the continuous rates $\lambda(t)$, $\mu(t)$, and $\phi(t)$, and the probabilities of birth bursts, intensive sampling

events, and mass extinctions appear in separate (non-D) terms of the likelihood. We could instead have defined branch segments to only end at observed births and samples, in which case the branch segments would cross interval times and the probabilities of intensive sampling events, and mass extinctions would appear only in $D_i(t)$.

We also note that we can exploit the structure of the phylogeny to simplify the calculation for branch segments, term (vii) in the likelihood function. Along a single lineage, the probabilities of adjacent branch segments will cancel out because t_y for one segment becomes t_o for the next. For segments that begin with bifurcations, the addition of a new lineage means that a single $D(t_o)$ remains in the numerator. For segments that end in tips, there is no next segment, and thus $D(t_y)$ remains in the denominator. If we take $\mathcal{T}$ to be the set of all tip times, we can compute (vii) as,

$$D(t_{or}) \prod_{t \in \mathcal{N}} D(t) \prod_{t \in \mathcal{T}} \frac{1}{D(t)}.$$

A.13 Related models

In Table A.1, we present an overview of related work on birth-death processes. It should be noted that there are essentially two classes of papers in this list. There are papers that are concerned with the theory of computing the probability density of a tree given parameters, which often allow rates to be any time-varying function. There are also papers concerned with *inferring* diversification rates from phylogenies, which generally impose more restrictive assumptions on how rates may vary. Note that we are excluding here the literature on birth-death processes where the rate varies among lineages, which are beyond the scope of this chapter. We also exclude comparison of models for sampling at the present, Φ_0, a more thorough discussion of which is available in [65, 59].

Table A.1: Time-varying birth-death models
Parameters that are absent from a model are marked with a dash (–), and can be assumed to be 0 compared to a model that includes that parameter. Rates through time are classified by whether they are assumed to be constant (const.), piecewise constant or episodic (epis.), or whether they are allowed to be any time-varying function (any). Tree-wide events are either present (any) in a model or they are absent (–), except for tree-wide sampling which may be restricted to a single event at the present (Φ_0). Conditioning includes conditioning on the various survival conditions discussed in Section S6 (I-V), and the number of tips (N). No conditioning listed is equivalent to simply conditioning on the time since the origin or MRCA. As many methods have been re-implemented in multiple software packages, the conditioning column only considers conditions used in likelihood equations in the cited paper. *[135] and [87] consider conditioning on the number of *extant* tips.

Model and citation	$\lambda(t)$	$\mu(t)$	$\phi(t)$	$r(t)$	$\boldsymbol{\Lambda}$	$\boldsymbol{M}$	$\boldsymbol{\Phi}$	$\boldsymbol{R}$	Conditioning
[105]	any	any	–	–	–	–	Φ_0	–	–
[51]	const.	const.	–	–	–	–	–	–	I, II, N
[134]	const.	const.	–	–	–	–	Φ_0	–	I, II, N
[100]	any	any	–	–	–	–	Φ_0	–	II
[59]	any	any	–	–	–	–	Φ_0	–	II, N
[136]	epis.	epis.	–	–	–	any	Φ_0	–	II, N
[60]	any	any	–	–	–	any	Φ_0	–	II, N
[95]	epis.	epis.	–	–	–	any	Φ_0	–	II, N
[135]	const.	const.	const.	0	–	–	–	–	II, III, N*
[139]	const.	const.	const.	const.	–	–	–	–	–
[140]	epis.	epis.	epis.	1.0	–	–	any	–	II
[49]	epis.	epis.	epis.	epis.	–	–	any	–	II
[87]	any	any	any	??	–	any	any	–	I-V, N*
Present paper	epis.	epis.	epis.	epis.	any	any	any	any	I-V
Most general model	any	any	any	any	any	any	any	any	I-V, N

A.14 Special Cases of the Birth-Death-Sampling-Treatment Process

In the following subsection we provide some special cases of our model. This simply shows that our model is a generalization of many previously published birth-death models, and how these models are related to another.

A.14.1 Episodic birth-death process

We get the episodic birth-death process when we specify the parameters as follows:

$$\begin{aligned}\phi(t) &= 0 \\ \Phi_i &= 0 \ \ \forall\ (i > 0) \\ M_i &= 0 \\ \Lambda_i &= 0\end{aligned}$$

This simplifies our equations to

$$C_i = E_{i-1}$$

and

$$\begin{aligned}f(\Psi) = \ & \frac{2^{I-1}}{I!} \\ & \times \prod_{t \in \mathcal{N}} \Big[\lambda(t)\Big] \\ & \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)}\Big]\end{aligned}$$

A.14.2 Episodic birth-death process with mass extinctions

We get the episodic birth-death process with mass-extinctions when we specify the parameters as follows:

$$\begin{aligned}\phi(t) &= 0 \\ \Phi_i &= 0 \ \ \forall\ (i > 0) \\ \Lambda_i &= 0\end{aligned}$$

This simplifies our equations to

$$C_i = (1 - M_i)E_{i-1}(t_i) + M_i$$

and

$$\begin{aligned} f(\Psi) = \ & \frac{2^{I-1}}{I!} \\ & \times \prod_{t \in \mathcal{N}} \Big[\lambda(t)\Big] \\ & \times \prod_{i=1}^{l} (1 - M_i)^{L(s_i)} \\ & \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)}\Big] \end{aligned} \qquad (vi)$$

A.14.3 Episodic fossilized-birth-death process

We get the (purely continuous) episodic fossilized-birth-death process for purely extinct taxa when we specify the parameters as follows:

$$\begin{aligned} r(t) &= 0 \\ \Phi_i &= 0 \\ M_i &= 0 \\ \Lambda_i &= 0 \end{aligned}$$

This simplifies our equations to

$$C_i = E_{i-1}(t_i)$$

and

$$\begin{aligned} f(\Psi) = \quad & \frac{2^{H-||\mathcal{A}||-1}}{(H-||\mathcal{A}||)!} \\ & \times \prod_{t \in \mathcal{F}} \Big[(\phi(t) E(t)) \Big] \\ & \times \prod_{t \in \mathcal{A}} \Big[\phi(t) \Big] \\ & \times \prod_{t \in \mathcal{N}} \Big[\lambda(t) \Big] \\ & \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)} \Big] \end{aligned}$$

A.14.4 Skyline transmission process

We get the skyline transmission-process model [49] by specifying parameters as follows,

$$\Phi_i = 0$$

$$M_i = 0$$

$$\Lambda_i = 0$$

This simplifies our equations to

$$C_i = E_{i-1}(t_i)$$

and

$$
\begin{aligned}
f(\Psi) = \quad & \frac{2^{H-1}}{H!} \\
& \times \prod_{t \in \mathcal{F}} \Big[\phi(t) \big(r(t) + \big(1 - r(t)\big) E(t) \big) \Big] \\
& \times \prod_{t \in \mathcal{A}} \Big[\phi(t)(1 - r(t)) \Big] \\
& \times \prod_{t \in \mathcal{N}} \Big[\lambda(t) \Big] \\
& \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)} \Big]
\end{aligned}
$$

A.14.5 Episodic transmission process with event-samples

We get the episodic sampled ancestor skyline model of [49] as follows,

$$M_i = 0$$

$$\Lambda_i = 0$$

$$R_i = r(s_i)$$

This simplifies our equations to

$$C_i = (1 - \Phi_i) E_{i-1}(t_i)$$

and

$$
\begin{aligned}
f(\Psi) = \quad & \frac{2^{I+H-||\mathcal{A}||-1}}{(I+H-||\mathcal{A}||)!} \\
& \times \prod_{i=0}^{l} \Big[(1-\Phi_i)^{(L(s_i)-I_i)} \Phi_i^{I_i} (1-R_i)^{T_i} (R_i + (1-R_i)E(s_i))^{I_i - T_i} \Big] \\
& \times \prod_{t \in \mathcal{F}} \Big[\phi(t) \big(r(t) + \big(1 - r(t)\big) E(t) \big) \Big] \\
& \times \prod_{t \in \mathcal{A}} \Big[\phi(t)(1 - r(t)) \Big] \\
& \times \prod_{t \in \mathcal{N}} \Big[\lambda(t) \Big] \\
& \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)} \Big]
\end{aligned}
$$

A.14.6 Episodic transmission process with event-samples and perfect treatment

We get the "birth-death skyline" model of [140] as a simplification of the episodic sampled ancestor skyline model by assuming perfect treatment as follows,

$$r(t) = 1$$

$$M_i = 0$$

$$\Lambda_i = 0$$

This simplifies our equations to

$$C_i = (1 - \Phi_i) E_{i-1}(t_i)$$

and

$$
\begin{aligned}
f(\Psi) = \quad & \frac{2^{H+I-1}}{(H+I)!} \\
& \times \prod_{i=0}^{l} \Big[(1-\Phi_i)^{(L(s_i)-I_i)} \Phi_i^{I_i} \Big] \\
& \times \prod_{t \in \mathcal{F}} \Big[\phi(t) \Big] \\
& \times \prod_{t \in \mathcal{N}} \Big[\lambda(t) \Big] \\
& \times \prod_{t \in \mathcal{B}} \Big[\frac{D(t_o)}{D(t_y)} \Big]
\end{aligned}
$$

A.15 Validation of likelihood function of episodic fosslized-birth-death process

The derivation of our likelihood function, *i.e.*the probability density function of a phylogenetic tree, relies heavily on the extinction probability $E(t)$ and the probability of an observed lineage $D(t)$. In the main text we provided our mathematical derivations. Here, we additionally validate the analytical solutions using forward simulations under the generalized fosslized-birth-death process. We started the simulations with one single lineage and simulated forward in time starting at $T = \{0.01, 0.02, \ldots, 0.99, 1.0\}$ time units in the past. We chose $\lambda(t) = 1.0$, $\mu(t) = 0.9$ and $\phi(t) = 0.1$. Additionally, we divided the total time into four epochs, thus, allowing for tree-wide events at $t = \{0.25, 0.5, 0.75\}$ with probabilities $\Lambda = \{0.0, 0.0, 0.3\}$, $\mathbf{M} = \{0.0, 0.5, 0.0\}$ and $\Phi = \{0.2, 0.0, 0.0\}$. We repeated the simulations 100,000 times and recorded how often the process went extinct ($E(t)$) and how often exactly one lineage was observed ($D(t)$). Reassuringly, the probabilities obtained from the simulations and the analytical solutions match exactly (Figure A.15).

The major novelty of our generalized birth-death-sampling process are the tree-wide events for mass extinctions and bursts of births. Thus, our simulations focusing on the three tree-wide events are sufficient as the impact of the continuous rates of speciation, extinction and sampling is validated through comparisons with the special cases in the previous section.

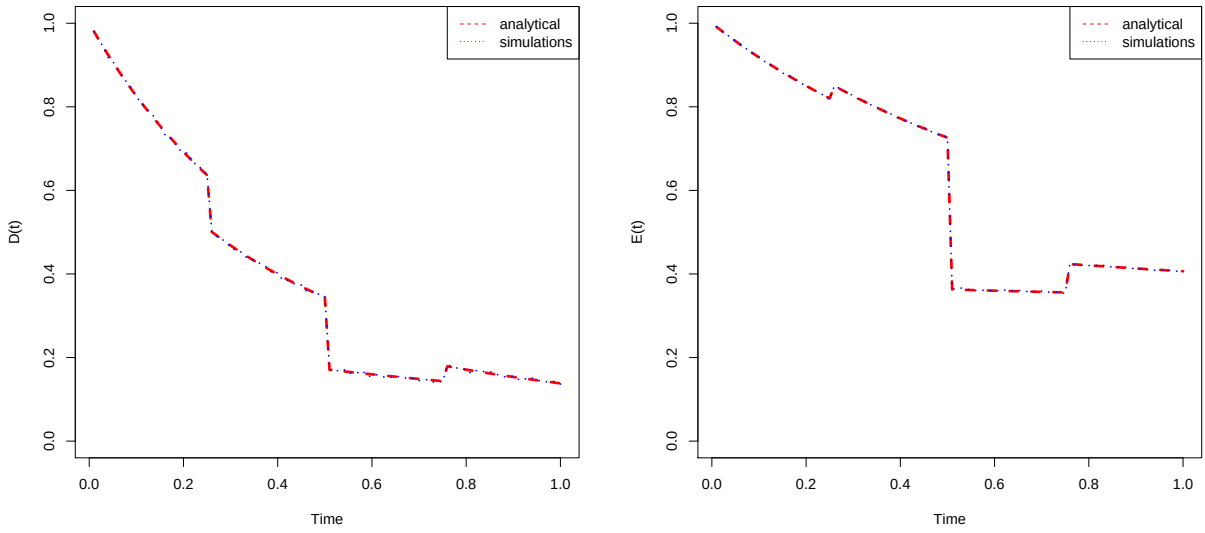


Figure A.15: Comparing the analytical solutions for $E(t)$ and $D(t)$ with probabilities obtained by forward simulating the birth-death-sampling process. The analytical solutions match the expectations obtained through simulations.

A.16 Validation of likelihood function and implementation using simulation based calibration

We performed a simulation based calibration to validate our episodic fossilized-birth-death process. Standard theory of Bayesian inference defines that, if the data are generated under exactly the same model as used for inference, then the true parameter values are included in the credible intervals exactly with the frequency corresponding to the size of the credible interval [66, 22]. For example, the true parameter value should be covered in a 90% credible interval in 90% of the simulation replicates, neither more or less often. A nice feature of simulation based calibration is that the validation only works if all three, the simulation method, the likelihood function and the inference method (*e.g.*the MCMC algorithm) are correctly implemented.

To validate our episodic fossilized-birth-death process, we choose the following approach. We designed a model with four equal-length epochs over a total time of 67.69 time units for the speciation, extinction and fossilization rates. Our assumption is that four epoch are sufficient to capture any potential problem with the per-epoch implementation but still being computationally

manageable to perform thousands of MCMC analyses.

In principle, the choice of prior distribution does not matter. However, in practice, it is beneficial to choose realistic prior distributions so that trees simulated under parameters chosen from the prior distribution are reasonable, *i.e.*are neither too large nor too improbable to survive. Thus, we specified a prior distribution on the net-diversification rate instead of the speciation rate to ensure that the simulated parameter values yield a positive net-diversification rate and hence the probability of the process going extinct is not close to 1.0. We employed a lognormal prior distribution on the net-diversification rate $\lambda_i - \mu_i$ with mean 0.01 and standard deviation 0.58, lognormal prior distribution on the extinction rate μ_i with mean 0.01 and standard deviation 0.58, and a lognormal prior distribution on the fossilization rate ϕ_i with mean 0.04 and standard deviation 0.58. Additionally, we employed a Beta$(20, 2)$ prior distribution on each the mass extinction death probability, the birth probability at a burst event, and the sampling probability at a tree-wide sampling event.

We implemented a forward simulator (which was also used for the posterior predictive distributions) and simulated trees given the parameter values drawn from the prior distribution. We conditioned the simulation on the root age of the extant tree (condition II, survival of the root). Then, we performed a standard MCMC algorithm using the same method as for the empirical analyses except that we used independent per-epoch priors instead of the HSMRF priors. The MCMC simulation was run for 10,000 iterations with 30 moves per iteration. We repeated this procedure 10,000 times to compute the frequency of how often the true parameter values were covered in the credible interval.

Finally, we computed and plotted the coverage frequencies for different credible interval sizes (Figure A.16). The varying credible interval sizes help to validate that the posterior distributions are neither too peaked nor too flat due to heavy and light tailed distributions. Our results (Figure A.16) demonstrate nicely that our forward simulator, our likelihood function, and our MCMC algorithm are all implemented correctly.

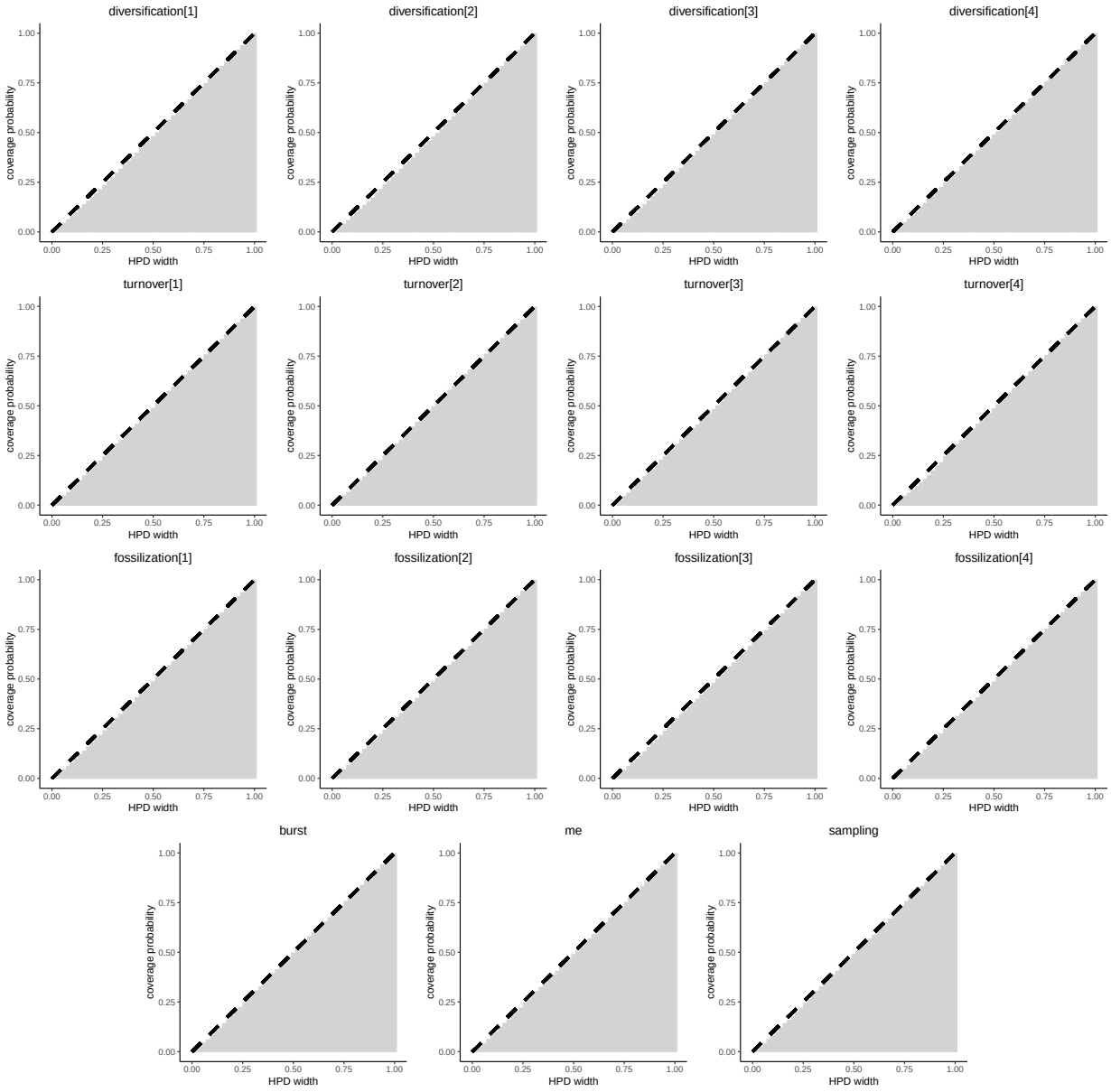


Figure A.16: Validation of our derived likelihood function of the episodic fossilized-birth-death process with tree-wide events of burst of births, mass extinction, and sampling. We performed simulation based calibration and validated that the true parameter values are covered with the expected probability, *i.e.*the size of the credible interval and the frequency of being including have to match. For all parameters in our example we observe a very good match between the expected and simulated coverage frequencies, indicating correct derivation of the theory and implementation of the likelihood function as well as MCMC algorithm.

A.17 Model parameterization

A.17.1 HSMRF

In Listing A.1 we provide the prior model specification for the speciation rates as employed in our analyses. `RevBayes` [63] provides enormous flexibility in specifying how diversification rates vary through time and across lineages. While the model we have employed here has been shown to work well in certain circumstances [90], it remains open to the biologist and future work which type of diversification-rate variation is most prevalent and what model is most robust. Note that the speciation rate at present has two hyperparameters that are determined from a prior analysis of the dataset at hand using a constant-rate fossilized birth-death model.

```
speciation_at_present ~ dnGamma(speciation_rate_hyperprior_alpha ,
                                speciation_rate_hyperprior_beta)


speciation_global_scale ~ dnHalfCauchy(0,1)

for (i in 1:(NUM_INTERVALS-1)) {

  # Variable-scaled variances for hierarchical horseshoe
  sigma_speciation[i] ~ dnHalfCauchy(0,1)

  # non-centralized parameterization of horseshoe
  delta_log_speciation[i] ~ dnNormal( mean=0,
                                      sd=sigma_speciation[i]*
                                      speciation_global_scale*
                                      speciation_global_scale_hyperprior )
}

# Assemble first-order differences and speciation at present
#     into the random field prior for the speciation rate
```

```
speciation := fnassembleContinuousMRF(speciation_at_present,
                                      delta_log_speciation,
                                      initialValueIsLogScale=FALSE,
                                      order=1)
```

Listing A.1: HSMRF on speciation rates.

We employed exactly the same type of model and priors on the extinction and fossilization rates. To keep this excerpt of our model concise, we show only the speciation rates.

A.17.2 Improving MCMC

Applying the HSMRF prior distribution to birth, death, and fossilization rates can make MCMC difficult. We previously developed an MCMC framework for inference consisting of Metropolis-Hastings moves on the initial rate and a mixture of elliptical slice sampling and Gibbs sampling [90]. This elliptical slice and Gibbs mixture works on the parameterization of the HSMRF prior in Listing A.1, with the elliptical slice sampler working on the `delta_log_speciation` while the Gibbs sampler works on the `sigma_speciation` and `speciation_global_scale`. The Gibbs move as previously implemented updates all the `sigma_speciation` in order, then updates `speciation_global_scale`. As the `speciation_global_scale` parameter can be quite difficult to sample, we have implemented a move that is a p, $(1-p)$ mixture of the previous Gibbs update and a Gibbs update solely on `speciation_global_scale`. The conditionals involved in updating `speciation_global_scale` are unchanged, but as the `speciation_global_scale` parameter depends on both the vector `sigma_speciation` and the vector `delta_log_speciation`, more frequent updates to the `speciation_global_scale` parameter allow it to adjust more quickly to changes in `delta_log_speciation` (and vice-versa).

The other update is a simple swap move that operates jointly on the `delta_log_speciation` and the `sigma_speciation`. We outline this move in Listing A.2; in brief, it simply swaps adjacent values of `delta_log_speciation[i]` and `sigma_speciation[i]` over the entire field. The move can migrate any pair

(`delta_log_speciation[i]`, `sigma_speciation[i]`) to any pair (`delta_log_speciation[j]`, `sigma_speciation[j]`), however it does not add any *new* variation to the parameters, and thus the move can only be used to augment MCMC approaches that actually introduce new values into the vectors `delta_log_speciation` and `sigma_speciation`, such as the elliptical slice sampler and Gibbs mixture. The move is symmetric, and so the Hastings ratio is 0. The motivation for the move is as follows. The HSMRF prior enforces that most `delta_log_speciation[i]` are very small, such that the speciation rate contains a number of relatively flat regions interspersed with "jumps" where the rate changes more rapidly. In practice, there is often considerable uncertainty regarding exactly which intervals contain the jumps, and this move allows us to directly explore this uncertainty and move the jump locations around. Simultaneously, this move preserves the large-scale features of the speciation rate: for any pair of indices i, j the total change in the speciation rate at i and at j will remain relatively consistent. The move operates on pairs of (`delta_log_speciation[i]`, `sigma_speciation[i]`) because these are compatible with each other; swapping a large-magnitude `delta_log_speciation[i]` with a small-magnitude `delta_log_speciation[j]` would pair a large-magnitude `delta_log_speciation[i]` with the small `sigma_speciation[j]` and this would lead to rejection.

```
u = randBernoulli(p=0.5)

start = floor(u)
end = start + 2 *
      (floor(length(delta_log_speciation) - start) / 2) - 1)

i = start

while (i < end) {
  tmp_d = delta_log_speciation[i]
  tmp_s = sigma_speciation[i]

  delta_log_speciation[i] = delta_log_speciation[i+1]
```

```
    sigma_speciation[i] = sigma_speciation[i+1]

    delta_log_speciation[i+1] = tmp_d
    sigma_speciation[i+1] = tmp_s

    i = i + 2
}
```

Listing A.2: HSMRF swap move.

Appendix B

APPENDIX TO: LOCALLY ADAPTIVE BAYESIAN BIRTH-DEATH MODEL SUCCESSFULLY DETECTS SLOW

AND RAPID RATE SHIFTS

B.1 Simulating parameters

Table B.1: The simulation values of the birth rates λ_1 and λ_2, and the times of change, t_1 and t_2, for all simulations. For the piecewise linear simulations, between t_1 and t_2, the birth rate is a linear interpolation between λ_1 and λ_2. For the piecewise constant simulations, there is only one time where the rate changes, t_1. For the constant-rate simulations, there are no change times and there is only one birth rate. The piecewise-linear simulations with $t_1 = t_2 = 50$ is also used when understanding the behavior of the models in the piecewise-constant scenarios, as it is nested in both sets of simulations (*). In all cases we round to 4 decimal places, the actual rates are not exactly equal for any pair of simulations.

simulation type	fold change	λ_1	λ_2	t_1	t_2
piecewise linear*	2	0.0727	0.0364	50	50
piecewise linear	2	0.0727	0.0364	37.5	62.5
piecewise linear	2	0.0727	0.0363	25	75
piecewise linear	2	0.0726	0.0363	12.5	87.5
piecewise linear	2	0.0726	0.0363	0	100
piecewise constant	2	0.0570	0.0285	10	N/A
piecewise constant	2	0.0603	0.0301	20	N/A
piecewise constant	2	0.0639	0.0320	30	N/A
piecewise constant	2	0.0681	0.0340	40	N/A
piecewise constant	2	0.0780	0.0390	60	N/A
piecewise linear*	4	0.0877	0.0219	50	50
piecewise linear	4	0.0877	0.0219	37.5	62.5
piecewise linear	4	0.0876	0.0219	25	75
piecewise linear	4	0.0876	0.0219	12.5	87.5
piecewise linear	4	0.0875	0.0219	0	100
piecewise constant	4	0.0586	0.0146	10	N/A
piecewise constant	4	0.0639	0.0160	20	N/A
piecewise constant	4	0.0703	0.0176	30	N/A
piecewise constant	4	0.0781	0.0195	40	N/A
piecewise constant	4	0.0999	0.0250	60	N/A
constant	N/A	0.0540	N/A	XXX	N/A

B.2 Performance on consant-rate datasets

To contextualize the performance of our models when the true scenario was constant-rate, we also inferred constant-rate birth-death processes on these trees. For constant-rate fits, the FC and RMASV performance measures are inapplicable, but the MAD and RP are still useful. Thus we

compared the results of fitting constant-rate models to fitting GMRF-based and HSMRF-based models to constant-rate simulations, and present these in Figure B.1. The GMRF-based model actually appears to perform better than the constant-rate model in terms of MAD, with an approximately 3% decrease in average MAD and an approximately 2% decrease in the 95th percentile of MAD. The HSMRF-based model performs trivially worse than a constant-rate model, with an approximately 1% worse average MAD and a 1% higher 95th percentile. Both models do suffer in precision, with the GMRF-based model's average RP approximately 31% higher than that of the constant-rate model and the HSMRF-based model's average RP 55% worse. As mentioned before, the cost of using a GMRF-based model or an HSMRF-based model instead of a constant-rate model is minor.

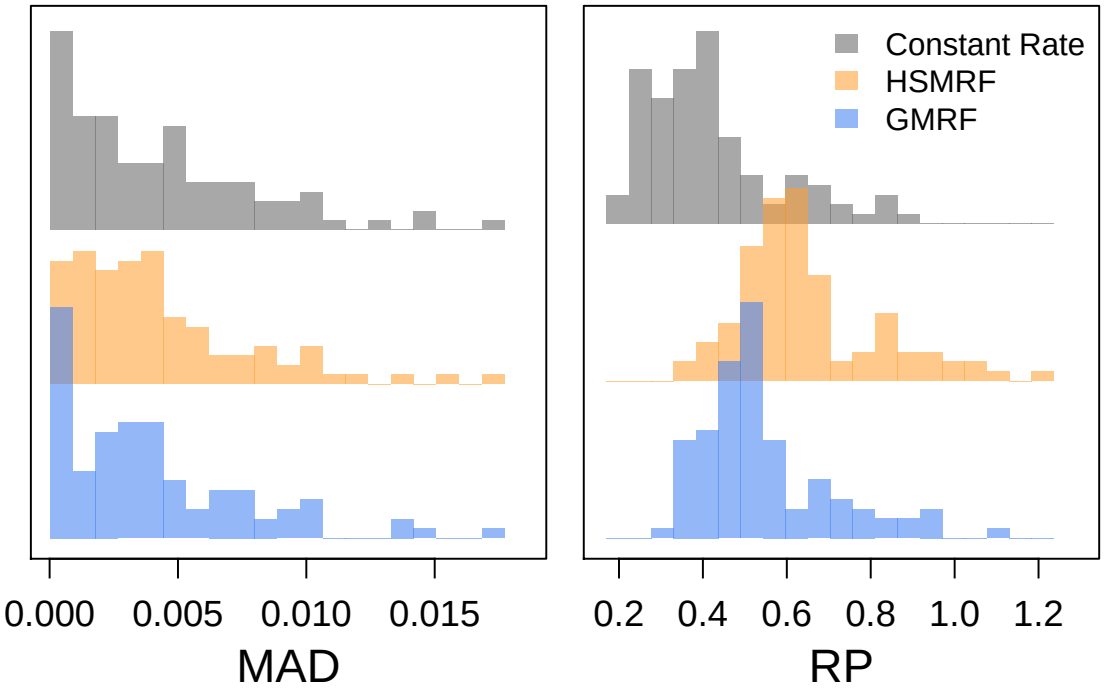


Figure B.1: Performance of the models on simulated constant-rate datasets. MAD measures the error in the estimated trajectory. RP is a measure of precision, the average width of the 90% Credible Interval relative to the birth rate. We compare these measures between the using time-varying models (GMRF in blue, HSMRF in orange) for inference and using the true model (constant-rate, grey).

B.3 Additional examples of performance with time-varying birth rates

In the main text we showed examples of performance on a single constant-rate simulation and four time-varying trajectories. Here we present examples of performance for all of the time-varying birth rate simulations.

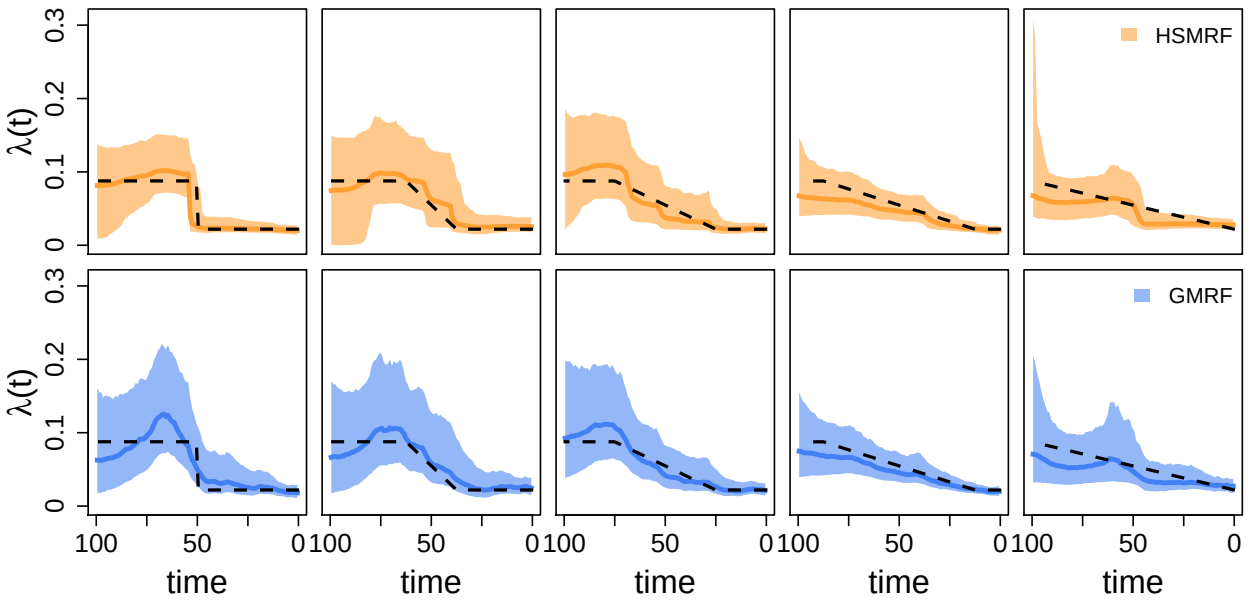


Figure B.2: Inferred birth-rate trajectories from five individual simulations. The dashed line is the true simulating birth rate, the dark colored line is the posterior median trajectory (the median is taken separately for each grid cell), and the shaded region shows the 90% Credible Interval (CI) for the rate. The columns demonstrate the effect of changing the shift duration (the length of the tree over which the birth rate changes), from an instantaneous shift to a constant change model. In each column, we show the simulation with the most average performance measured in terms of the Mean Absolute Deviation of both the GMRF and HSMRF.

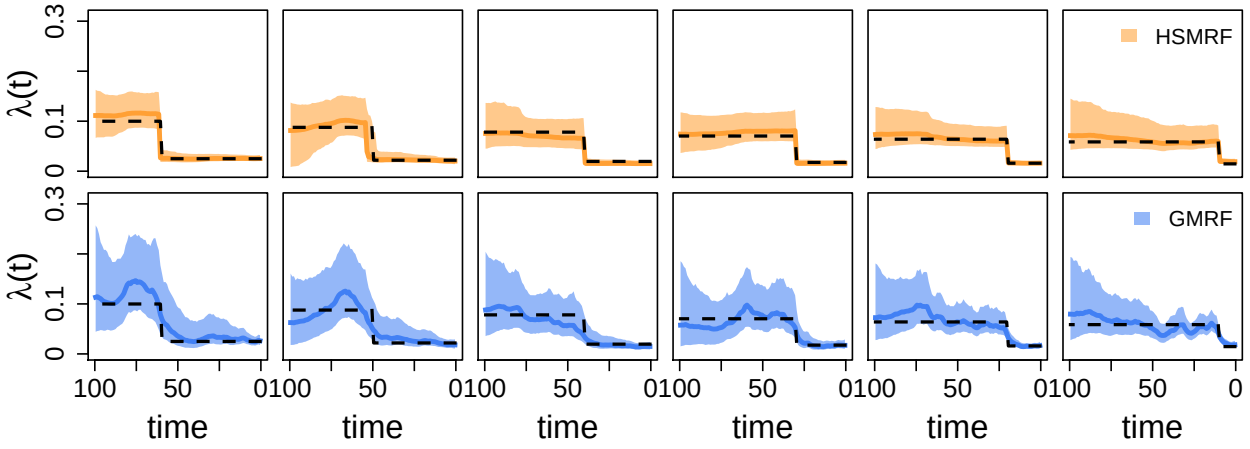


Figure B.3: Inferred birth-rate trajectories from four individual simulations. The dashed line is the true simulating birth rate, the dark colored line is the posterior median trajectory (the median is taken separately for each grid cell), and the shaded region shows the 90% Credible Interval (CI) for the rate. The columns demonstrate the effect of changing the location of the (instantaneous) shift from 60 time units before the present to 10 time units before the present. In each column, we show the simulation with the most average performance measured in terms of the Mean Absolute Deviation of both the GMRF and HSMRF.

B.4 Additional simulation results

In the main text we have focused on the results of simulations that had a four-fold net change. The results for two-fold changes are qualitatively similar, but here for completion we present versions of Figures 3 and 4 for two-fold shifts. We also present a complete table with all simulation parameters.

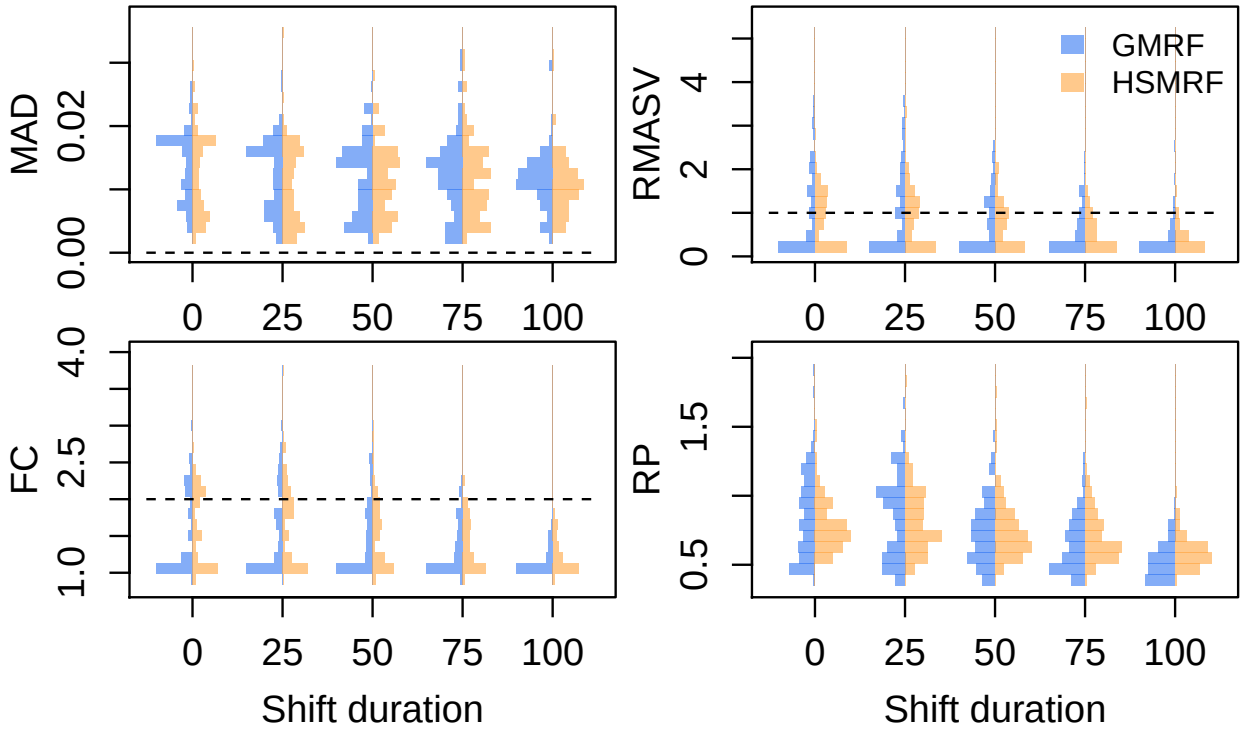


Figure B.4: The effect of changing the duration of the two-fold rate shift, from instantaneous to the entire length of the trajectory. MAD measures the error in the estimated trajectory. RMASV measures the total amount of change relative to the true MASV, horizontal line at 1 for reference. FC measures the fold change from present to past, dotted line at true value for reference. RP is a measure of precision, the average width of the 90% Credible Interval relative to the birth rate.

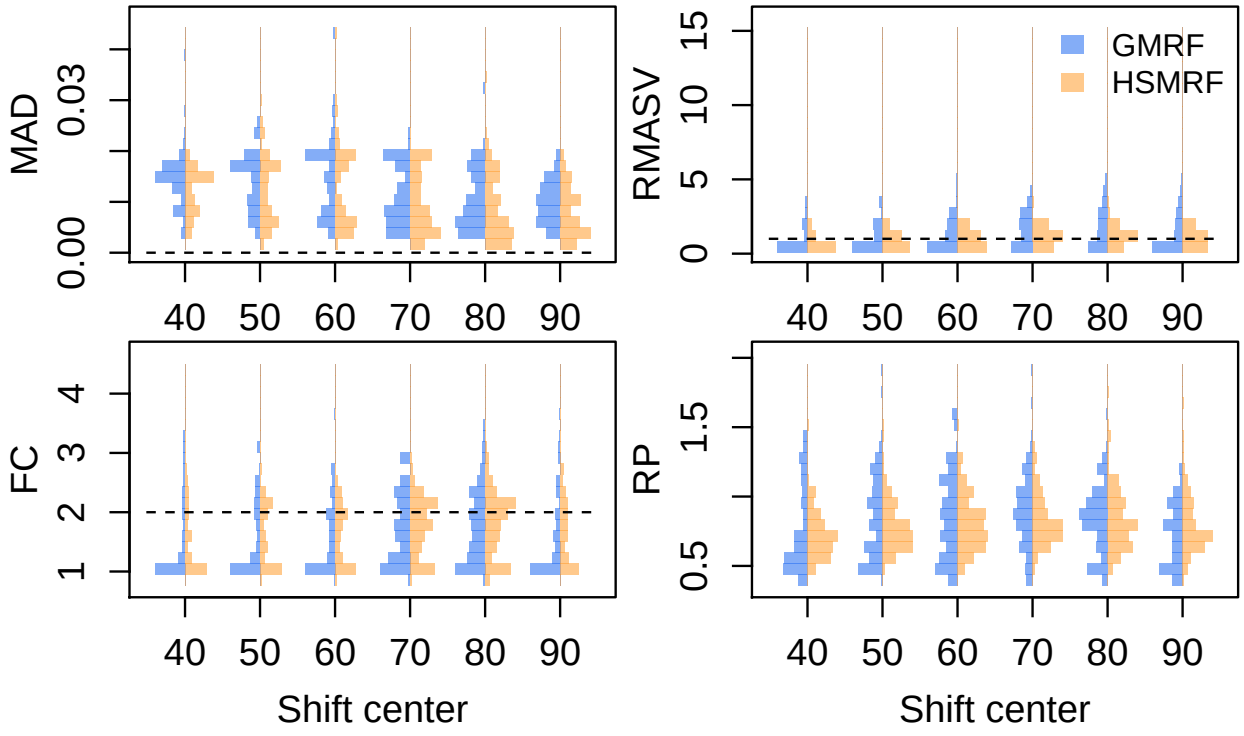


Figure B.5: The effect of changing the location of the two-fold rate shift, from 60 to 10. MAD measures the error in the estimated trajectory. RMASV measures the total amount of change relative to the true MASV, horizontal line at 1 for reference. FC measures the fold change from present to past, dotted line at true value for reference. RP is a measure of precision, the average width of the 90% Credible Interval relative to the birth rate.

We lastly present a Figure for each experiment performed with additional summary measures. The additional summary measures are the minimum Effective Sample Size (minESS, taken across all values in the logfile, including parameters and transformed parameters), the coverage (coverage, percent of posterior birth rate 90% CIs containing the true value), and the Mean Squared Error (MSE).

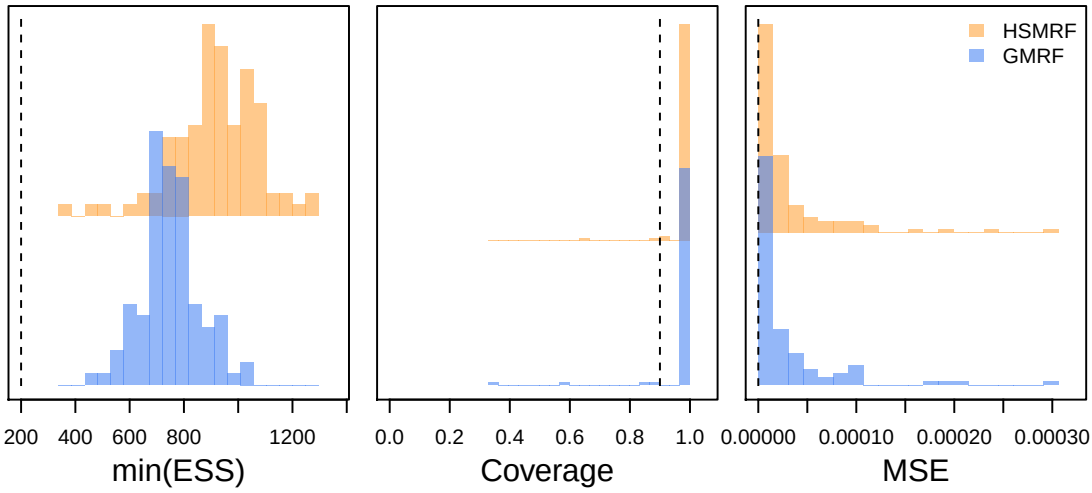


Figure B.6: Performance on the constant-rate simulated datasets. minESS is the mininum ESS of all logged quantities (parameters, prior/likelihood/posterior, and transformed parameters), dashed line at 200 for reference. Coverage is the proportion of 90% CIs of the birth rates that include the true birth rate, dashed line at 0.9 for reference. MSE is the Mean Squared Error of the estimated trajectory, dashed line at 0 for reference.

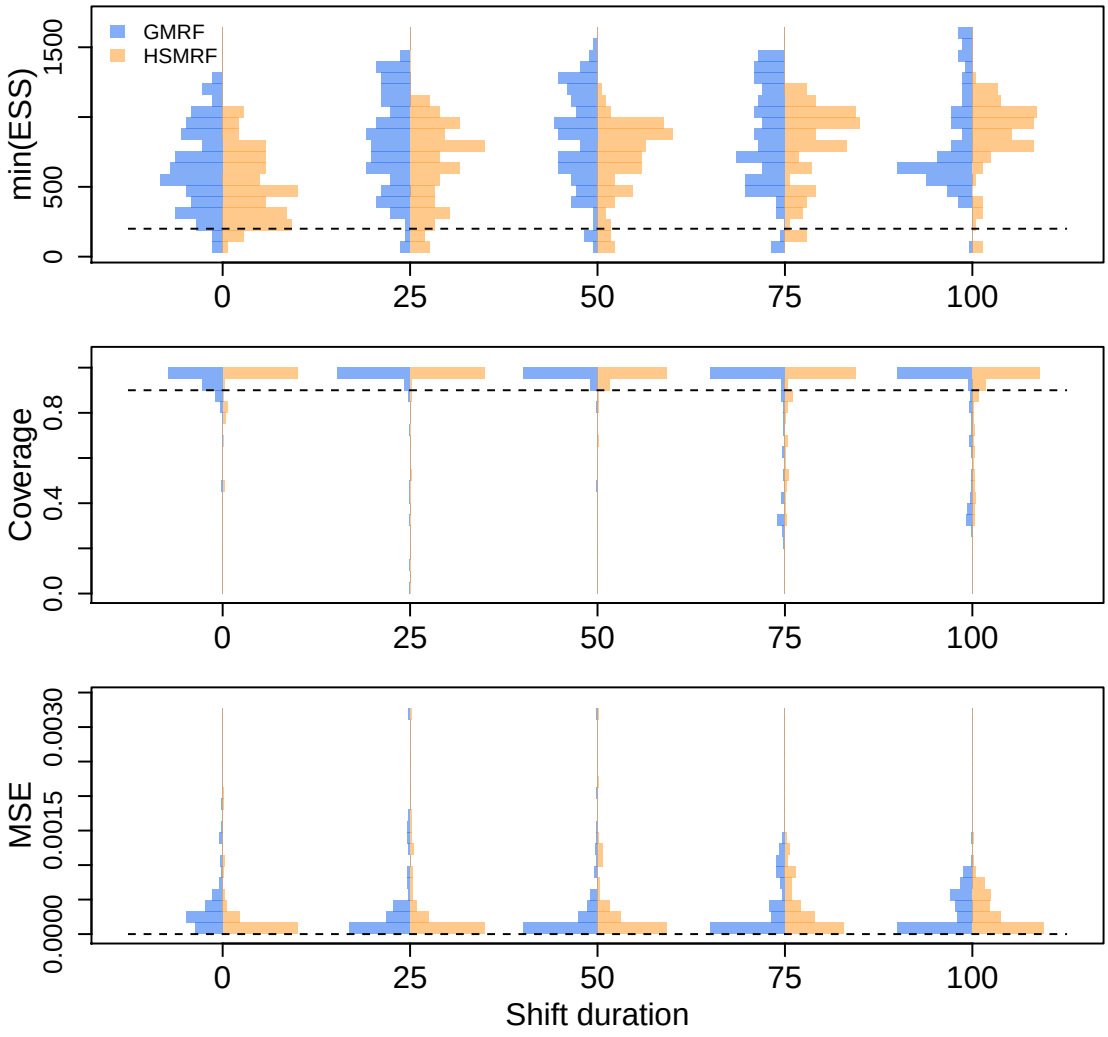


Figure B.7: The effect of changing the duration of the four-fold rate shift, from instantaneous to the entire length of the trajectory. minESS is the mininum ESS of all logged quantities (parameters, prior/likelihood/-posterior, and transformed parameters). Coverage is the percent of 90% CIs of the birth rates that include the true birth rate. MSE is the Mean Squared Error of the estimated trajectory.

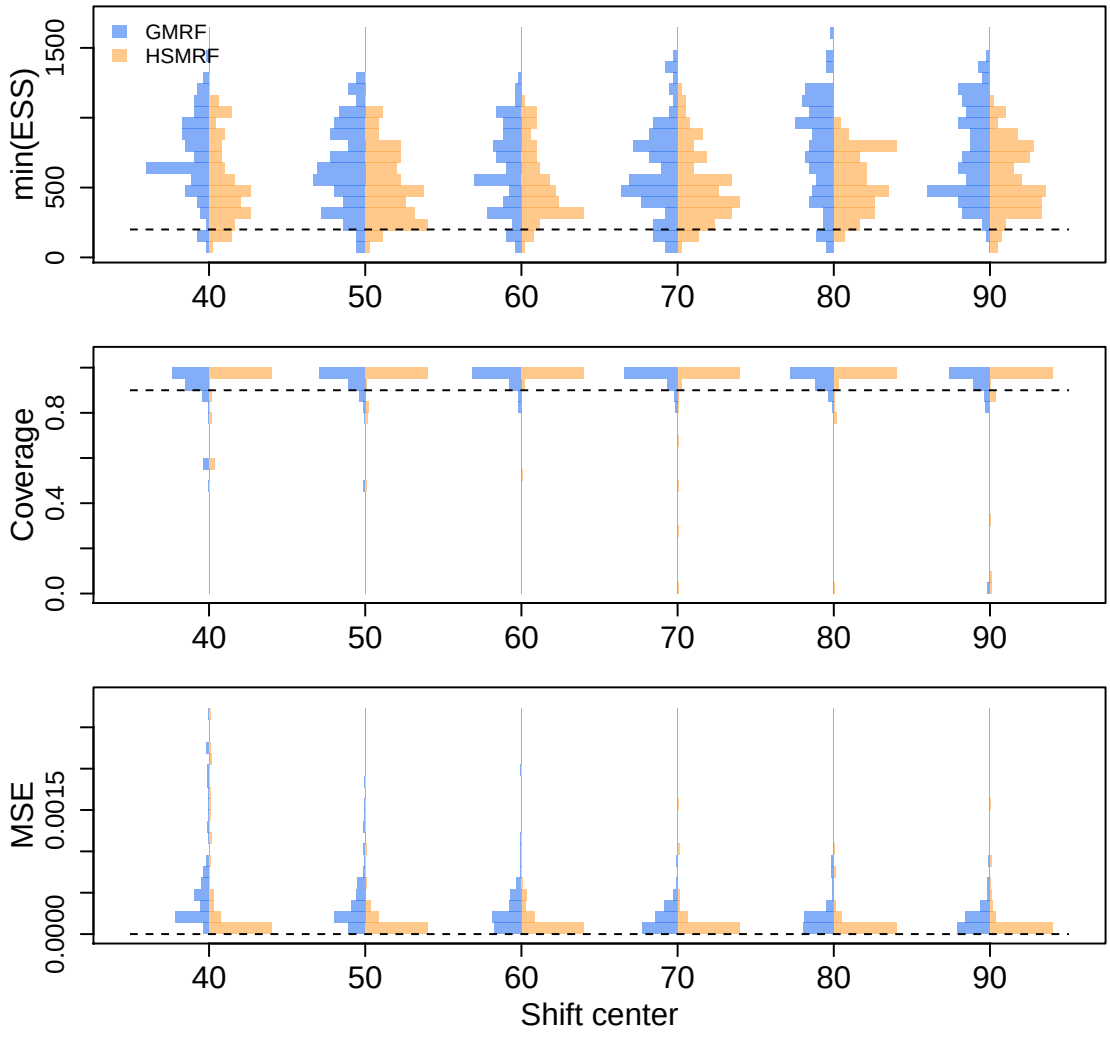


Figure B.8: The effect of changing the location of the four-fold rate shift, from 60 to 10. minESS is the mininum ESS of all logged quantities (parameters, prior/likelihood/posterior, and transformed parameters). Coverage is the percent of 90% CIs of the birth rates that include the true birth rate. MSE is the Mean Squared Error of the estimated trajectory.

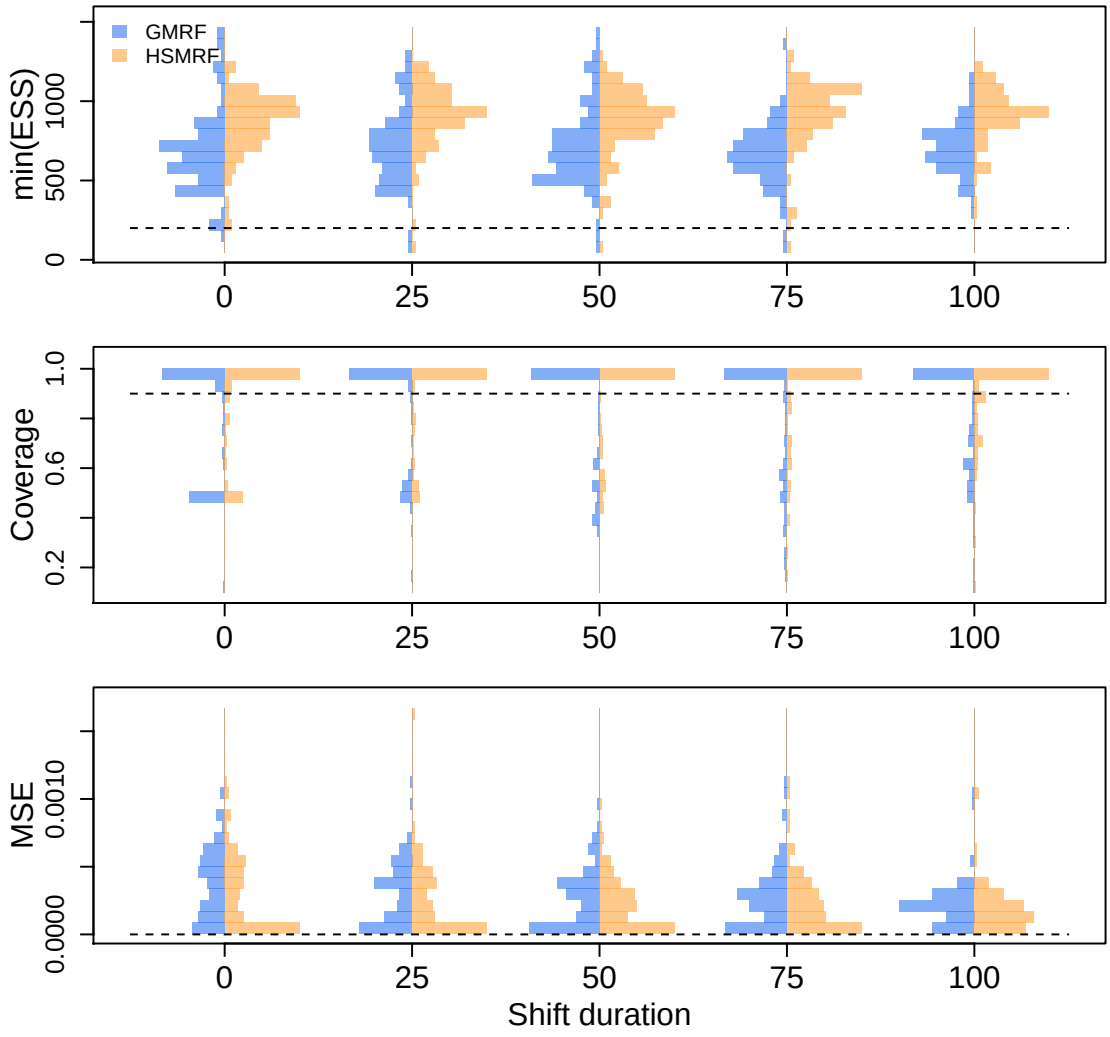


Figure B.9: The effect of changing the duration of the two-fold rate shift, from instantaneous to the entire length of the trajectory. minESS is the mininum ESS of all logged quantities (parameters, prior/likelihood/-posterior, and transformed parameters). Coverage is the percent of 90% CIs of the birth rates that include the true birth rate. MSE is the Mean Squared Error of the estimated trajectory.

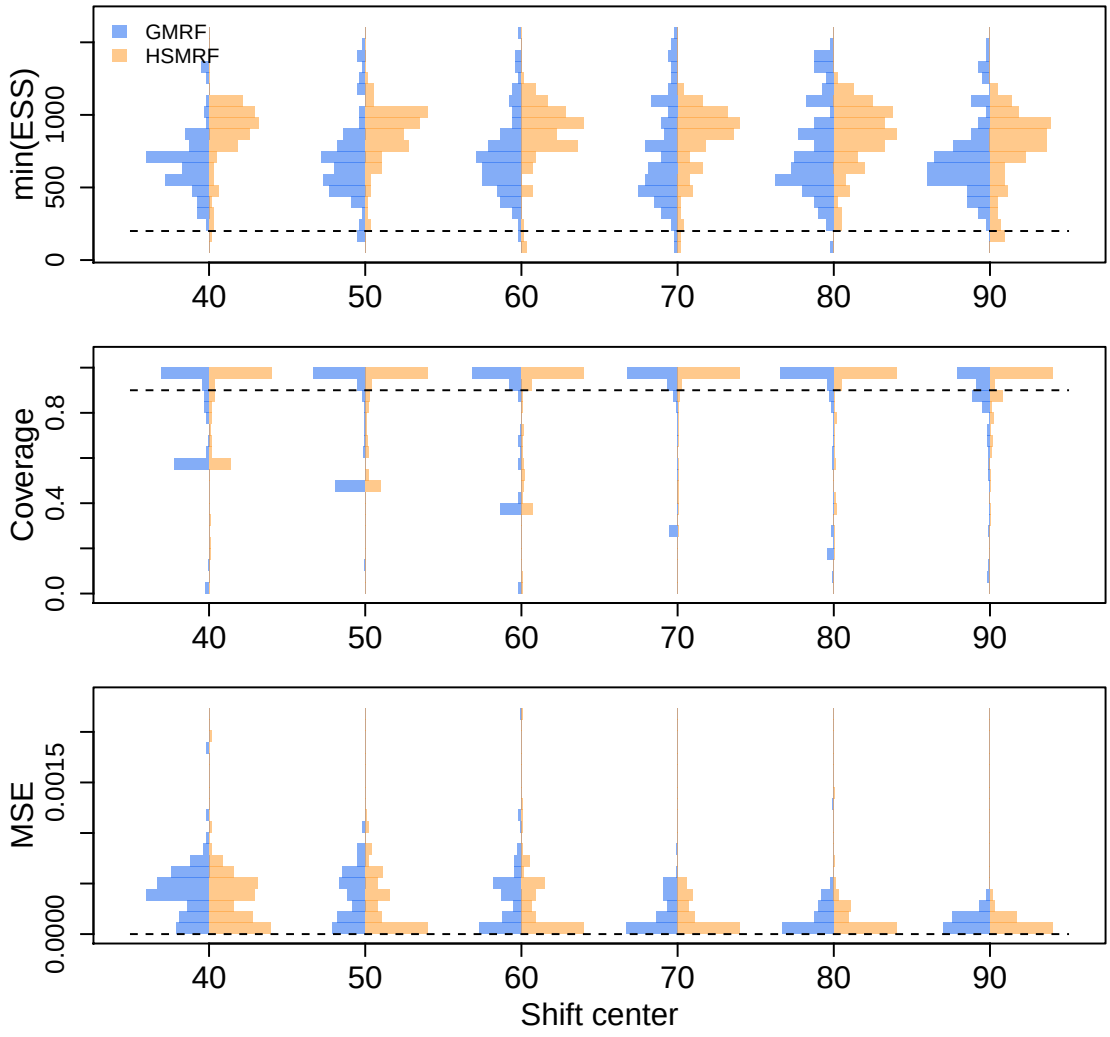


Figure B.10: The effect of changing the location of the two-fold rate shift, from 60 to 10. minESS is the mininum ESS of all logged quantities (parameters, prior/likelihood/posterior, and transformed parameters). Coverage is the percent of 90% CIs of the birth rates that include the true birth rate. MSE is the Mean Squared Error of the estimated trajectory.

B.5 Estimating constant death rates

In the main text we showed examples of performance of estimating the death rate on two time-varying death-rate trajectories. Here we present examples of performance of estimating the death rate for all simulated scenarios where the true death rate was constant.

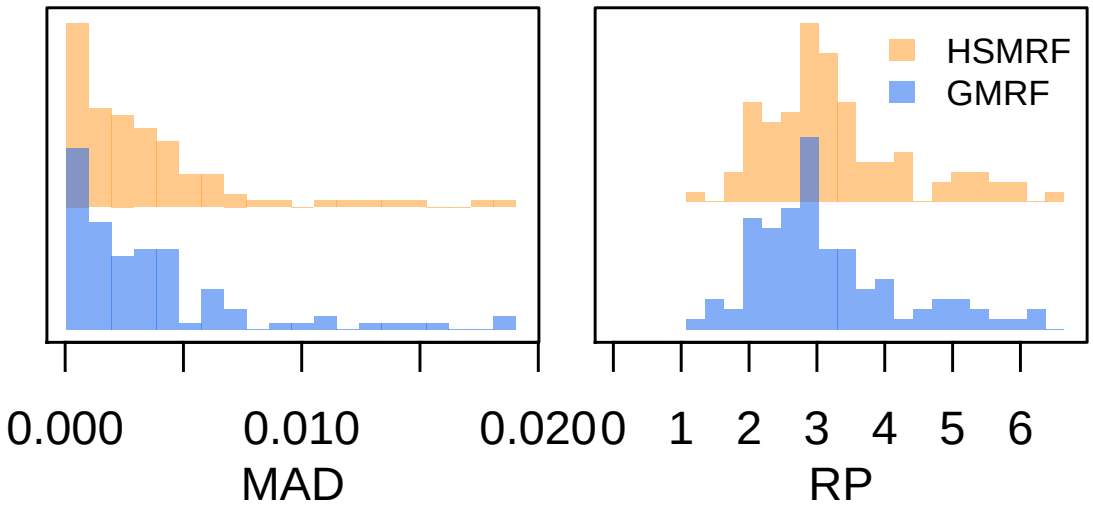


Figure B.11: Performance of estimating the death rate in the constant-rate simulations. MAD measures the error in the estimated death rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the death rate.

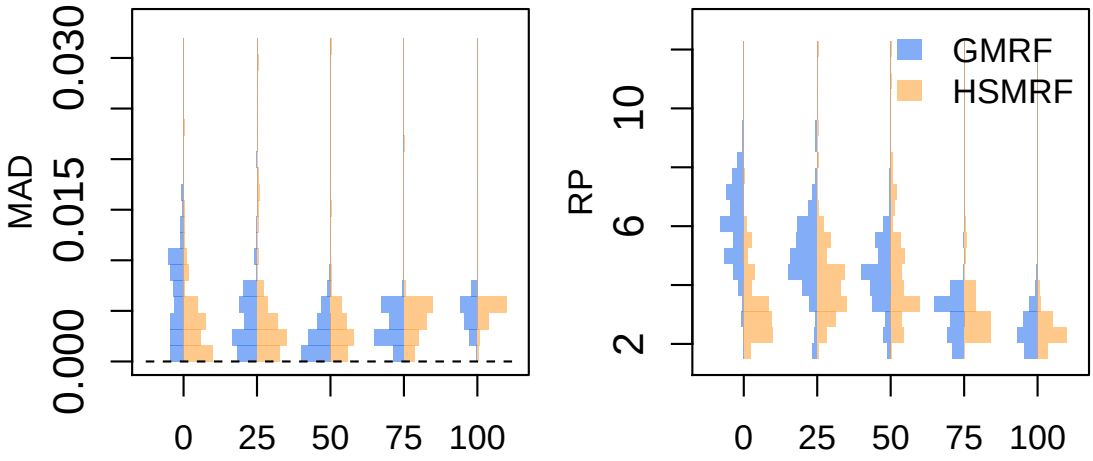


Figure B.12: The effect of changing the duration of the four-fold rate shift, from instantaneous to the entire length of the trajectory on the estimated death rate. MAD measures the error in the estimated death rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the death rate.

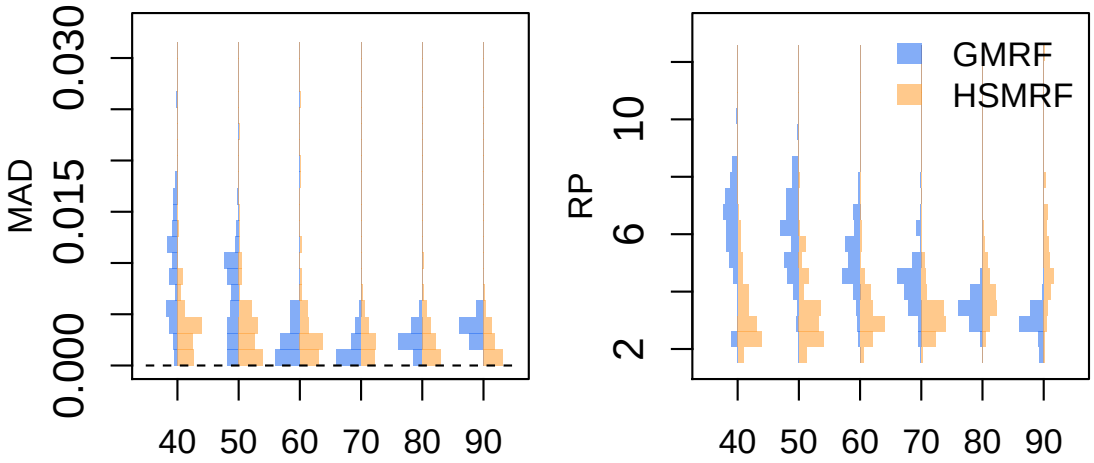


Figure B.13: The effect of changing the location of the four-fold rate shift, from instantaneous to the entire length of the trajectory on the estimated death rate. MAD measures the error in the estimated death rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the death rate.

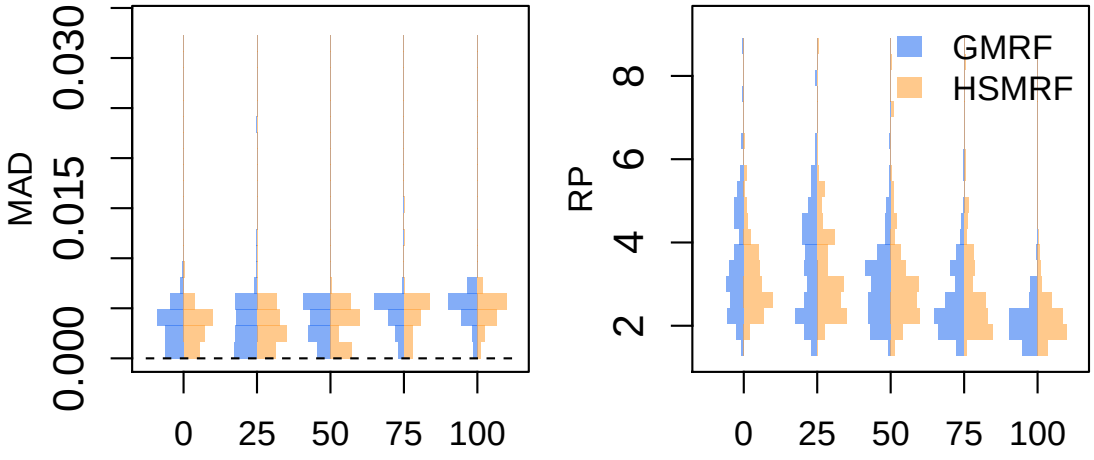


Figure B.14: The effect of changing the duration of the two-fold rate shift, from instantaneous to the entire length of the trajectory on the estimated death rate. MAD measures the error in the estimated death rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the death rate.

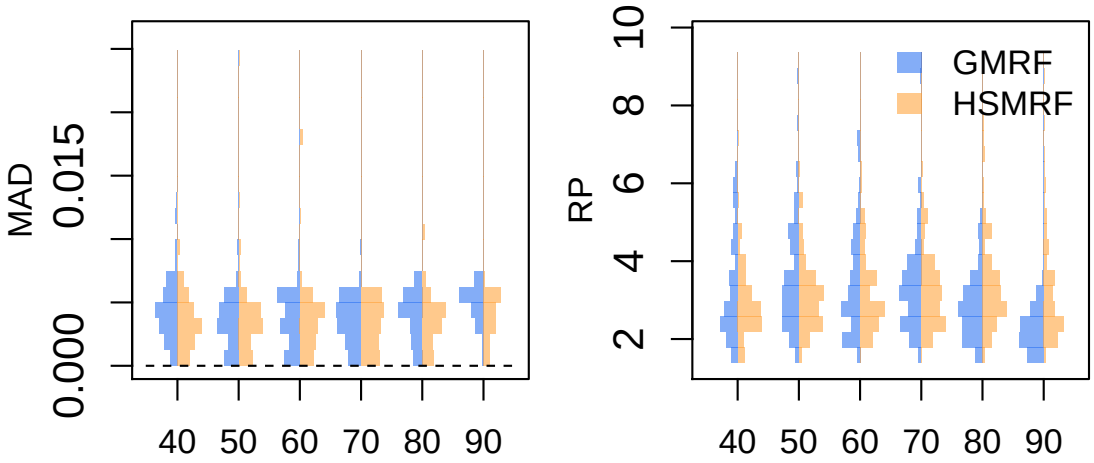


Figure B.15: The effect of changing the location of the two-fold rate shift, from instantaneous to the entire length of the trajectory on the estimated death rate. MAD measures the error in the estimated death rate. RP is a measure of precision, the width of the 90% Credible Interval relative to the death rate.

B.6 Estimating time-varying death rates

In the main text we showed examples of performance when simulating from a model with time-varying death rates, and histograms of the MAD, RP, and RMASV performance measures. Here we present histograms with additional performance measures. We omit coverage, as the true death rate in many intervals is 0, making the coverage of the 90% CI seem artificially low. There is only one set of histograms for the minimum rank-ESS as it is taken across all parameters in the posterior.

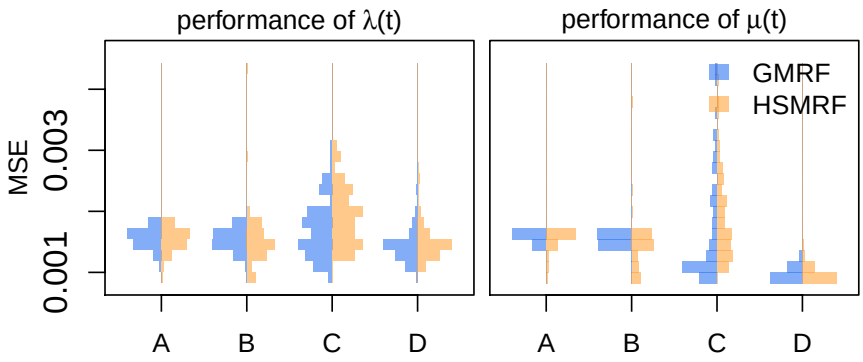


Figure B.16: Performance of the models on simulated datasets with time-varying extinction. MSE is the Mean Squared Error of the estimated trajectory, dashed line at 0 for reference. The column labels A, B, C, and D identify the different combinations of tree simulations and analysis setup. A and B are analyses of trees with isochronous sampling, C and D heterochronous sampling. A and C are analyses where time-varying death rate, $\mu(t)$ is inferred, B and D where a constant death rate, μ, is inferred.

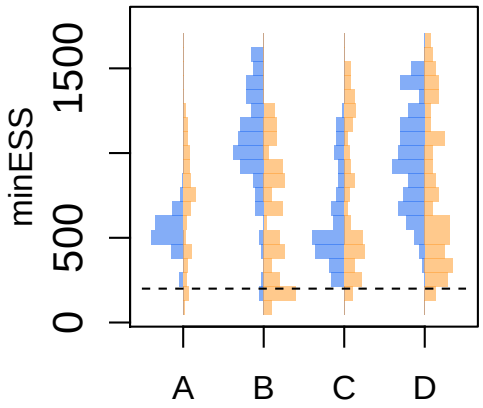


Figure B.17: Performance of the models on simulated datasets with time-varying extinction. minESS is the mininum rank-ESS of all logged quantities (parameters, prior/likelihood/posterior, and transformed parameters), dashed line at 200 for reference. The column labels A, B, C, and D identify the different combinations of tree simulations and analysis setup. A and B are analyses of trees with isochronous sampling, C and D heterochronous sampling. A and C are analyses where time-varying death rate, $\mu(t)$ is inferred, B and D where a constant death rate, μ, is inferred.

B.7 Non-centered parameterizations

Markov random fields can be parameterized by the distribution of the change between neighbors, $\Delta_i := \lambda^*_{i+1} - \lambda^*_i$. In a GMRF, these "increments" of the field, are iid Normal(0,γ) random variables, and we can obtain λ_i from $\lambda_1, \lambda^*_2, \ldots, \lambda^*_{i-1}$ as $\lambda_i = \lambda_1 \exp(\sum_1^{n-1} \Delta_i)$. This approach is called the non-centered parameterization, and simplifies MCMC sampling. The usefulness of non-centralized parameterizations is that they break the dependency between adjacent field parameters by making every move to field parameters multivariate. A proposed change to λ^*_i must fit within the prior imposed by λ^*_{i-1} and must place a reasonable prior on λ^*_{i+1}. However, a change to Δ_i implicitly changes not only λ^*_i, but $\lambda^*_i, \ldots, \lambda^*_n$ as well.

We can circumvent directly placing a half-Cauchy(0,ζ) prior on γ, by placing a halfCauchy(0,1) prior on γ and using $\gamma\zeta$ in its place. This rescaling enables us to use a Gibbs sampler for γ (for more details on this sampler see the next section), which enables us to appropriately explore the tails of the halfCauchy distribution where standard Metropolis Hastings moves fail. Thus, in our non-centered, rescaled parameterization of the GMRF, the increments are given by $\Delta_i \sim \text{Normal}(0, \gamma^2\zeta^2)$. For HSMRFs we expand this re-scaling to the σ_i, and rewrite $\sigma_i \sim \text{halfCauchy}(0, \gamma)$ as $\sigma_i \sim \text{halfCauchy}(0, 1)$ and where we would use σ_i we use $\sigma_i\gamma\zeta$. For our non-centered, rescaled HSMRF, the increments are given by $\Delta_i \sim \text{Normal}(0, \sigma_i^2\gamma^2\zeta^2)$. We show the models as DAGs in Figure B.18 as they are setup in our `RevBayes` analyses for full joint analysis of phylogeny and birth rates.

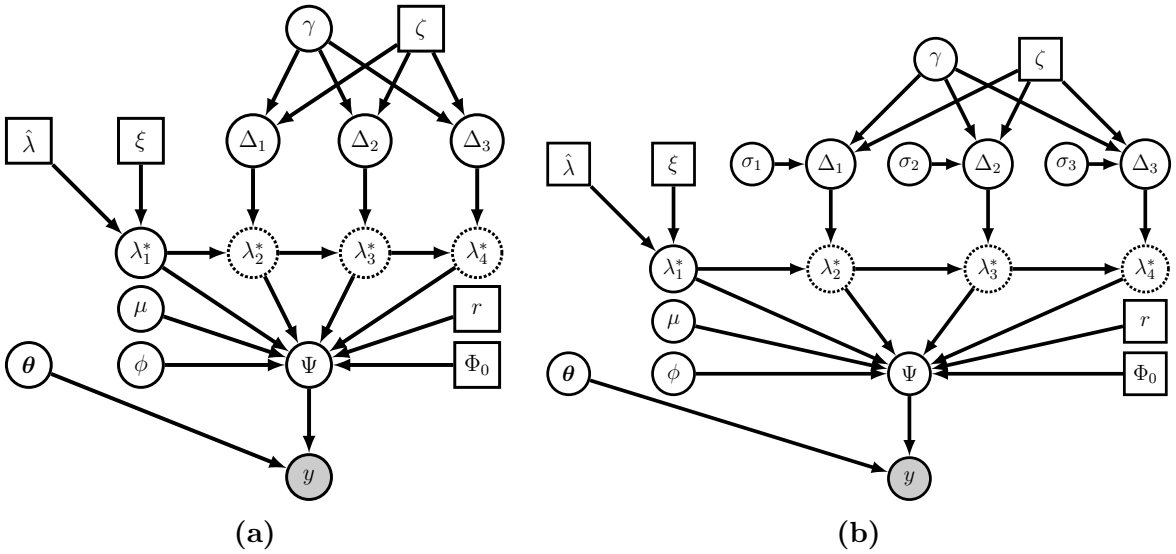


Figure B.18: The parameterization of our models as setup in `RevBayes`, shown as a grid of size 4 for simplicity. In (a) we show our GMRF-based model, and in (b) our HSMRF-based model. The use of $\boldsymbol{\Delta}$ as parameters instead of $\boldsymbol{\lambda}^*_{2:n}$ simplifies MCMC by allowing us to sample independent variables. The rescaling of the $\boldsymbol{\Delta}$ by separating out $\boldsymbol{\sigma}$, γ, and ζ is required for the Gibbs sampler to run, but also serves to minimize the number of layers in the model. In addition to $\boldsymbol{\lambda}$, the tree prior is specified by the serial sampling rate ϕ, death rate μ, sampling probability at the present Φ_0, and conditional probability of death upon sampling (treatment probability) r. We place all substitution and clock model parameters in $\boldsymbol{\theta}$, such that given $\boldsymbol{\theta}$ and the tree Ψ, the likelihood of the data $\boldsymbol{D}$ can be computed. When drawing the model as a DAG, squares represent constant values, closed circles stochastic values, open circles deterministic transformations of other nodes, and shaded circles observed stochastic values (data).

B.8 MCMC procedures

The major drawback to shrinkage priors is that the same fat tailed behavior that makes them adaptive makes inference difficult. We find that sampling is best accomplished under the non-centered, rescaled parameterization. However, while these model reparameterizations make sampling easier, basic Metropolis-Hastings MCMC proposals proved incapable of exploring the tails of the posterior. Our solution is to use an elliptical slice update to all increments of the random fields, $\boldsymbol{\Delta}$, Gibbs updates to all field hyperparameters, $\boldsymbol{\sigma}$ and γ, and standard MH updates to λ^*_1, μ, and any substitution or tree model parameters.

Elliptical slice sampling [102] exploits the geometry of multivariate normal priors. It slices not in a straight line but along an ellipse that should maintain higher posterior probability. Mechanistically, this is accomplished by drawing proposed values from the multivariate normal prior and

searching for an angle of rotation θ (to combine the proposed and current values) that is acceptable under the likelihood. In our case, the $\mathbf{\Delta}$ are all independently distributed, thus we can circumvent the usual need for matrix decomposition when drawing the proposed values, decreasing the time it takes to run the sampler. In Algorithm 1 we provide pseudocode for this special case, for a full explanation of the general case we refer readers to Murray *et al.* [102]. We note that in our implementation, there are two differences from the algorithm as presented in Algorithm 1, tuning and a hard limit on iterations. We allow for tuning by initializing $\theta \sim \text{Uniform}(0, m)$ ($m \leq 2\pi$) and tuning m so as to reduce the total number of iterations required to accept. Larger m correspond to bolder moves, and when the data are particularly informative, most moves will not be accepted until θ is close to 0, thus starting with a smaller m reduces the number of iterations of the loop. We also cap the number of loop iterations at 2000, at which point the sampler will abort or, if the user desires, accept a move of $\theta = 0$ (after 2000 iterations, $\theta \approx 0$).

Gibbs sampling has been proposed for horseshoe distributions in regression contexts for all parameters in the model [91]. Faulkner *et al.* (2018) have expanded this to work on arbitrary HSMRF models and GMRF models [43]. This Gibbs sampling approach nominally requires further reformulating the model, beyond the parameterization discussed above. Specifically, it reparameterizes halfCauchy variables as mixtures of inverseGamma variables by adding yet another layer of auxiliary variables. To make implementing our models more user-friendly (by reducing the number of hierarchical layers in the model), we implement this Gibbs sampler in `RevBayes` to operate directly on the non-centered parameterization we have described thus far. This is permissible because the conditional distributions of the new auxiliary variables only depend on the values of σ_i and γ. Thus they can be drawn at every sampling step conditioned on the current values, and then new values of σ_i and γ can be drawn and returned, leaving the auxiliary variables entirely behind the scenes. For the HSMRF-based model, the Gibbs sampler requires specifying a value of ζ, and the rest of the non-centered HSMRF model is as follows,

$$\begin{aligned} \gamma &\sim \text{halfCauchy}(0, 1), \\ \sigma_i &\sim \text{halfCauchy}(0, 1), \\ \Delta_i \mid (\sigma_i, \gamma, \zeta) &\sim \text{Normal}(0, \sigma_i^2 \gamma^2 \zeta^2). \end{aligned}$$

For the GMRF-based model, the Gibbs sampler requires specifying a value of ζ, and the rest of the non-centered GMRF model is as follows,

$$\gamma \sim \text{halfCauchy}(0, 1),$$
$$\Delta_i \mid (\gamma, \zeta) \sim \text{Normal}(0, \gamma^2\zeta^2).$$

We present pseudocode for implementing the Gibbs samplers in Algorithms 2 and 3, noting that the samplers operate on γ^2 and $\boldsymbol{\sigma}^2$, whereas our parameterization is on γ and $\boldsymbol{\sigma}$. In practice inverseGamma variates are obtained by drawing Gamma random variables and inverting them. Thus in our implementation of the algorithms we draw $\boldsymbol{\psi}^{-1}$, ξ^{-1}, $\boldsymbol{\sigma}^{-2}$, γ^{-2} directly from gamma distributions and invert variables only as needed.

B.9 Diagnosing MCMC convergence

When performing Bayesian analysis of models, performance diagnostics are absolutely necessary. We use the Potential Scale Reduction Factor (PSRF) to determine if two chains have converged [12]. We follow Vehtari *et al.* (2019) in using rank-transformed variables in our PSRF calculations to avoid issues induced by fat-tailed posterior distributions because the original PSRF assumes normally distributed variables [153]. For each variable in our posterior distribution we compute two convergence diagnostics, the PSRF of the rank-transformed variables, $\hat{R}_r$, and the PSRF of the rank-transformed folded variables, $\hat{R}_{rf}$, which can capture differences in variance between chains. To ensure that all our summaries are based on trustworthy MCMC runs, we discard the entire posterior for any analysis where $\max(\hat{R}_r, \hat{R}_{rf}) > 1 + \epsilon$. In the simulations with a constant death rate, a cutoff using $\epsilon = 0.01$ rarely results in more than 2 or 3 analyses discarded per hundred analyses (HSMRF or GMRF analyses of simulated datasets, all constant-rate analysese passed). In the simulations with a time-varying death rate, this cutoff generally results in over 15 analyses being discarded per 100 analyses. Using $\epsilon = 0.02$ results instead in a maximum of 20 analyses discarded, and qualitatively similar results figures. We thus use $\epsilon = 0.01$ for the constant death simulations and $\epsilon = 0.02$ for the variable death simulations. We also calculate the effective sample size (ESS) for all parameters for the two chains combined for each model, and find that almost all parameters have ESS > 200. ESS is generally larger for the time-varying death rate simulations, which were run

twice as long (and thinned twice as much, 400,000 sampled every 400 instead of 200,000 sampled every 200). For the empirical analyses and the analyses of simulations with time-varying death rates, we compute the ESS of the rank-stabilized variables following [153].

B.10 The size of the grid

In all of our simulations and analyses, we have employed a grid with 100 cells (a grid of size 100). Though performance with a grid of size 100 appears to be good, some may wonder if this grid is sufficiently large, or even too large. Using a tree sampled from the posterior distribution for the Pygopodidae HSMRF analysis, we here address that question. We take this tree to be data, and run analyses on it with grid sizes of 10,20,50,100, and 200, which we plot in Figure B.19. We run two independent chains per analaysis, and check rank-based PSRF and rank-ESS for all parameters in both GMRF and HSMRF analyses of all 5 grid sizes. The minimum rank-ESS was 176 and the maximum PSRF was 1.016, indicating adequate convergence for our purposes. While the inferences using smaller grids (10 or 20) look rather different from the others, it is visually hard to distinguish the inferences using larger grids (50, 100, and 200). Thus, 100 is a reasonable grid size, in that it produces inferences similar to larger grids, while being faster to run.

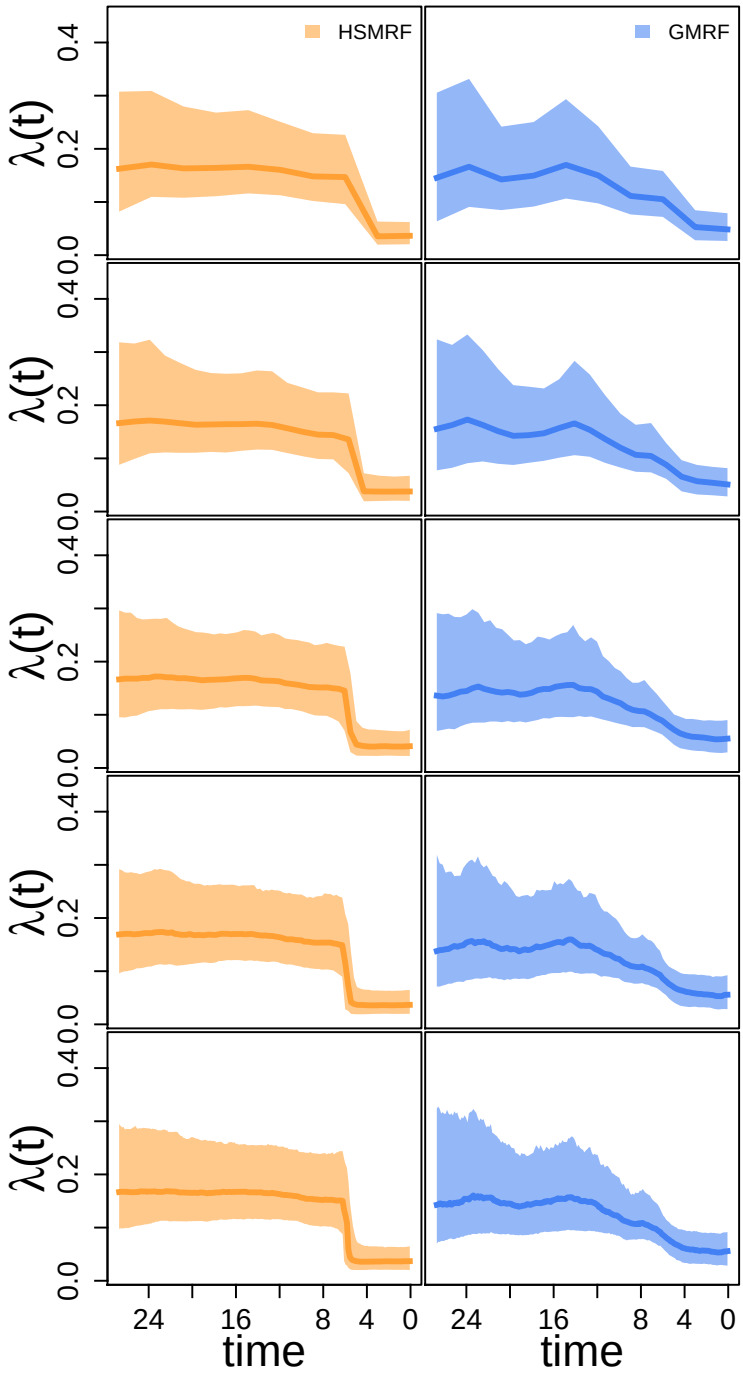


Figure B.19: The effect of the size of the grid on estimated speciation rates in Pygopodidae. The tree is fixed to a tree from the posterior distribution from the Pygopodidae HSMRF analysis. From top to bottom, the grid sizes used for inference are 10, 20, 50, 100, and 200.

B.11 Sensitivity to death-rate prior

In all of our analyses of simulated data, and our empirical analysis of Pygopodidae, we employed an Exponential(10) prior on the death (extinction) rate. However, it is plausible that there is some prior sensitivity to the extinction rate prior. Using a tree sampled from the posterior distribution for the Pygopodidae HSMRF analysis (the same as with the grid size analyses), we here address that question. We take this tree to be data, and run analyses on it with several different death-rate priors. In all cases we hold the birth-rate prior constant and use the same prior we have used everywhere. Three of these are exponential priors, with rates of 1, 10, and 100. Two of these priors are empirical Bayes priors, where we fit a constant-rate birth-death model to the dataset first, and then using the method of moments fit either a Lognormal or a Gamma distribution to the posterior. Following May *et al.* [95], we inflate the variance 10-fold. These empirical Bayes priors have mean 0.011 and variance 0.0013. We run two independent chains per analaysis, and check rank-based PSRF and rank-ESS for all parameters in both GMRF and HSMRF analyses of all 5 analyses. The minimum rank-ESS was 270 and the maximum PSRF was 1.018, indicating adequate convergence for our purposes.

The estimated death rate is notably sensitive to the prior (Figure B.20). The Exponential(1) and Exponential(10) priors produce the most similar posterior distributions, and appear to exert the least influence. The estimated birth rates are much less sensitive to the prior on the death-rate B.21. The credible intervals are wider for the Exponential(1) and Exponential(10) priors than the other three priors. The magnitude of the shift is also somewhat different between the Exponential(1) and Exponential(10) priors (FC of approximately 4.7 for the HSMRF analyses and 2.6 for the GMRF analyses) and the other three priors (FC of approximately 3.9 for the HSMRF analyses and 1.9 for the GMRF analyses). However, qualitatively all trajectories are very similar and show clear evidence of a shift at approximately 6 Ma. Thus, even if the exact estimates of the birth rate change somewhat with the estimated death rate, the pattern remains consistent.

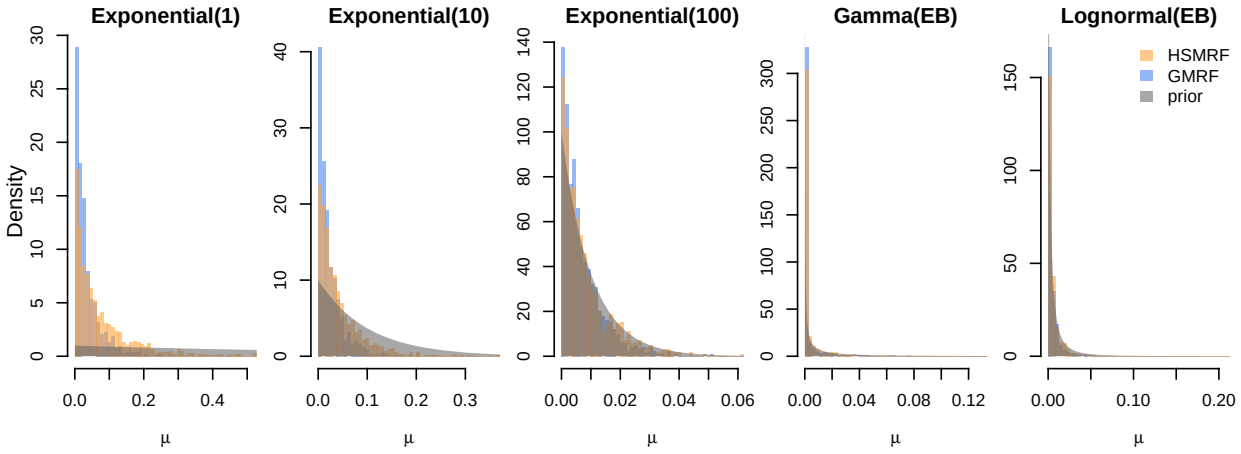


Figure B.20: The effect of the death-rate prior on the estimated death rate. The tree is fixed to a tree from the posterior distribution from the Pygopodidae HSMRF analysis. The plots are not scaled consistently in order to facilitate prior-posterior comparisons for each prior individually.

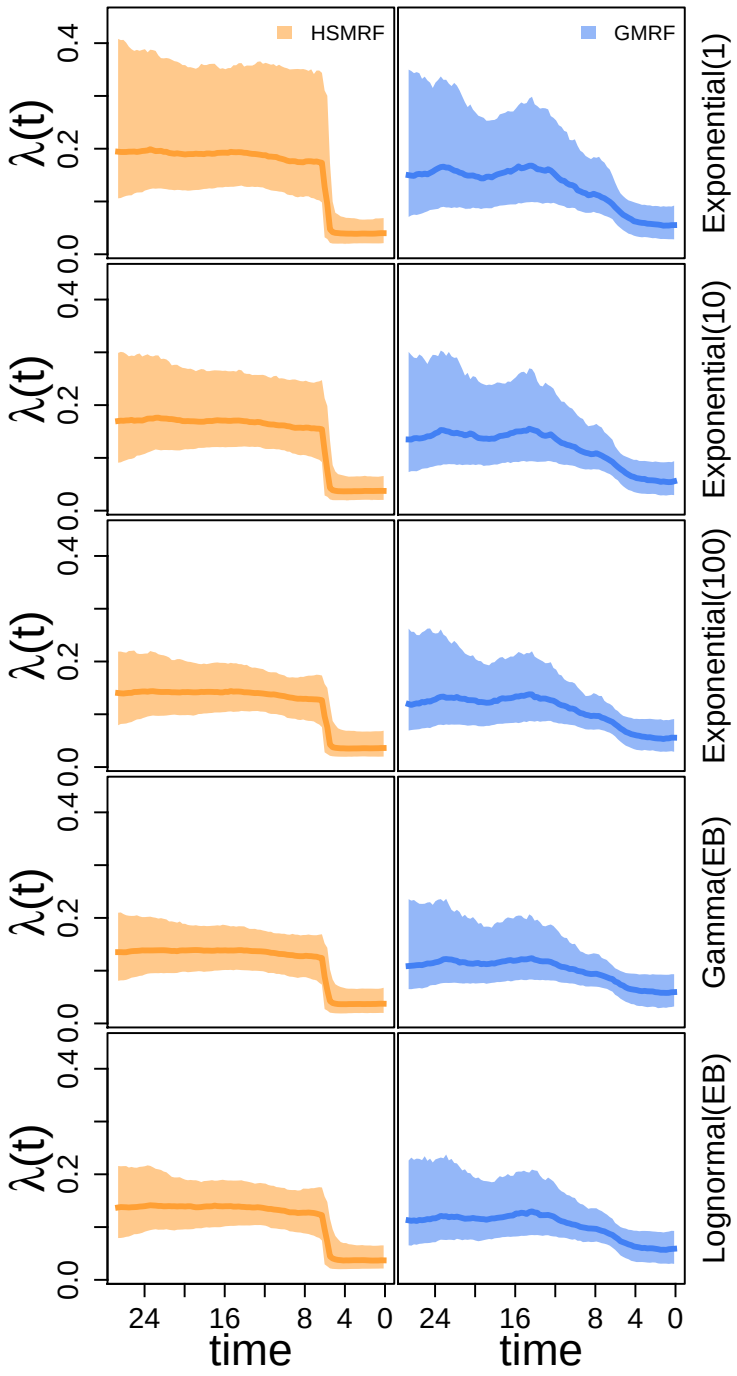


Figure B.21: The effect of the death-rate prior on the estimated birth-rate trajectory. The tree is fixed to a tree from the posterior distribution from the Pygopodidae HSMRF analysis. The label on the right hand side is the prior on the death rate.

B.12 Setting the global shrinkage prior

As discussed in the main text, setting the halfCauchy(0,ζ) prior on γ is a crucial step in framing a random field model. Our strategy is inspired by the way Drummond and Suchard (2010) parameterize their random local clock model, but some adjustments have to be made because we work with completely continuous parameters. Namely, Drummond and Suchard (2010) choose a prior that places 50% probability on a model with 0 shifts, but for our models this implies a non-negligable probability of the process ending at a value over a one-trillion-fold changed from the initial. Because neither random field model is truly binary, even a process that shows no "rate shifts" can exhibit rather substantial variation throughout the history of the process. We therefore set our prior based on the expected number of shifts (the prior mean), and choosing this to be ln(2) aligns our mean with the mean of the prior in Drummond and Suchard (2010) [31].

There are three other classes of approaches to setting ζ that may be useful in different circumstances. In the first class of approaches, one commonly bounds the marginal variances of the GMRF, such that the probability that the marginal variance of the field at t_i exceeds some reference value is α (which may be arbitrarily chosen to be 0.05) [133, 44]. This approach is not immediately amendable to parameterization via biological intuition, however. Alternately, it is possible to employ noisy estimators to bound the total variation of the random field [43]. This approach is applicable to modeling effective population sizes by using the skyline estimator [116] to get fast per-interval maximum likelihood estimates, then using the variances of these estimated effective population sizes to set ζ. While there is no equivalent procedure available for birth-death processes, assuming a pure-birth process would allow the use of a skyline-like estimator to bound the process. However, such skyline-based approaches are only applicable in situations in which a tree already exists for the taxa; if there is no tree there is nothing from which we can estimate skyline birth rates. Therefore, while this approach can be quite useful, we prefer one that can be done before any trees have been estimated.

The third set of approaches considers each of the horseshoe distribution as producing variables that are binary, either effectively 0 or not. The application of horseshoe priors to regression models has been well studied in this framework [111]. The horseshoe introduces shrinkage weights, κ_i on the model parameters β_i, such that at $\kappa_i = 1$ the posterior mean of β_i is 0 and it is effectively not

a model parameter (it is this distribution for which the horseshoe was named). Where analytical maximum likelihood estimators exist, it is possible to construct a prior on the effective number of parameters in the model. Given the complex dependencies among adjacent time-intervals in both birth-death processes and clock models, however, this approach seems out of reach.

B.13 Setting a prior for ϕ

As with our prior on λ_1, it is possible to obtain an empirical estimate of the sampling rate ϕ. Let $N(t)$ be the number of lineages alive at time t (measuring time in time units after the origin, rather than our usual present to past). Assuming a constant rate model, taking $d = \lambda - \mu - \phi r$, and ignoring conditioning on the survival of the tree, the expected number of lineages at time t is $\mathbb{E}(N(t)) = e^{d \times t}$ if we start with a single lineage. If we start with two lineages (as we do if we start at the MRCA), we have instead $\mathbb{E}(N(t)) = 2e^{d \times t}$. The rate of adding samples to the tree at time t is $\phi N(t)$, with expectation $\phi e^{d \times t}$, or $2\phi e^{d \times t}$ if we start at the MRCA. Integrating, we get the expected number of serial samples up to time t being $S(t_{or}) = \phi(e^{d \times t} - 1)/d$, or $S(t_{or}) = 2\phi(e^{d \times t} - 1)/d$ if we start at the MRCA. This means we can use the method of moments to obtain $\hat{\phi} = d \times S(t)/(e^{d \times t} - 1)$, or $\hat{\phi} = d \times S(t)/(2(e^{d \times t} - 1))$ if we start at the MRCA. This of course requires d, but as we have already discussed, estimation of d is also possible by the method of moments. In practice one might wonder how well this approach performs, since we are repeatedly using the data to estimate these parameters. Using our heterochronously sampled simulations, we find that it works surprisingly well (Figure B.22).

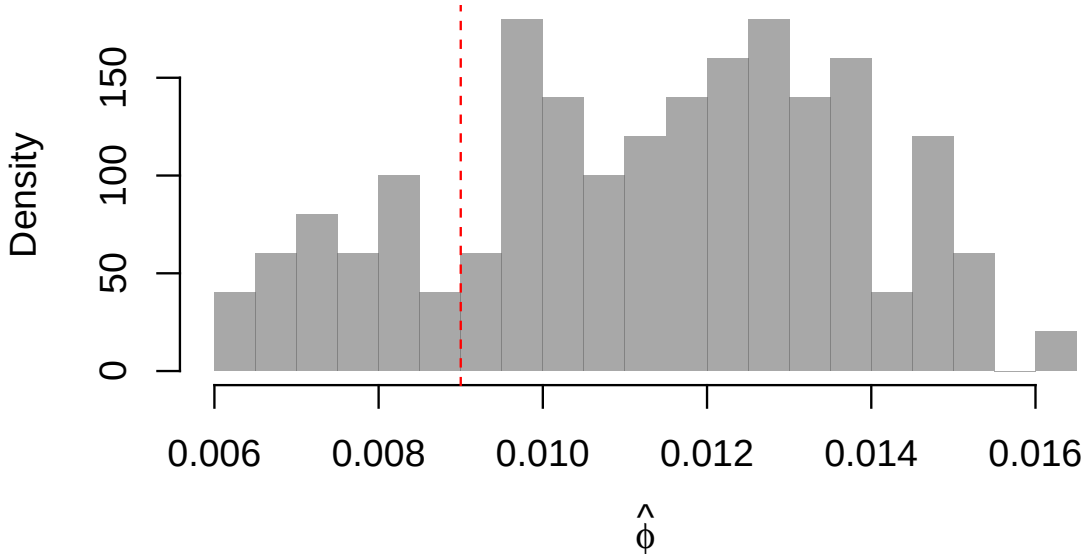


Figure B.22: Performance of our proposed method of moments estimator $\hat{\phi}$ on our simulated heterochronously sampled trees. The true simulating value (0.009) is in red.

B.14 Diversification in Pygopodidae

We obtained the multiple sequence alignment for the entire dataset of Brennan and Oliver (2017) from their Dryad repository [8]. We extracted sequences for Pygopodidae manually in AliView [79], keeping only one representative for each species (the representative with the most sequence data), which left some gap-only sites that we later removed. Since our analyses concern a much smaller number of taxa, representing less variation at any site in the alignment, we used PartitionFinder 1.1.1 [76] to select subsets of the alignment for analysis. In PartitionFinder, we set branch lengths unlinked (such that different partitions should use independent molecular clock models), using BIC for model selection, and set the set of allowable substitution models to (JC, JC+G, JC+I, HKY, HKY+G, HKY+I, GTR, GTR+G, GTR+I), to avoid parameter identifiability issues with +I+G mixture models. The best schemes identified 2 subsets for Pygopodoidae, with the best substitution model being GTR+G for both [145].

In our analyses, for each subset we applied an uncorrelated lognormal clock model (UCLD) [29]. We placed an exponential(rate=3) prior on the log-scale standard deviation and a Normal(ln(0.001),4*0.587405) prior on the log-scale mean (yielding a prior median substitution rate

of 0.001 per site per million years with a 95% prior CI of [1e-6,1]). Following Brennan and Oliver (2017), we applied uniform calibrations to the genera *Apprasia* (uniform(8.5,17)) and *Delma* uniform((14,22.5)) [9]. We used their node calibration for Pygopodidae (uniform(19.5,29.0)) as our root age prior. We ran 4 chains for 500,000 iterations for both the GMRF and HSMRF models (with 218 moves per iteration), downsampling to every 100th sample. This is the equivalent of 109 million generations per chain in a program like BEAST [32], sampled every 21,800.

We used the same rank-based PSRF procedure to assess convergence of the empirical analyses as for the simulated analyses. In convergence diagnostics, we ignore the branch-rate parameters as the recorded parameters in the log-file are not comparable across trees. These diagnostics revealed convergence issues with the substitution model for one of the HSMRF replicate analyses. While all other parameters had acceptable PSRF for the 4 HSMRF replicates, convergence is most properly addressed in an all-or-none framework, so we chose to exclude this entire chain from downstream analyses. Discarding this chain, the maximum rank-PSRF for the HSMRF analyses was 1.0008, and for the GMRF it was 1.0001. We were unable to calculate the effective sample size of the HSMRF chains directly due to some extremely large sampled values for the speciation rate. We thus chose to compute the effective sample size of the rank-transformed (but not folded) parameters for all chains, which should be correlated with (if not equal to) the effective sample size of the untransformed parameters. The rank-ESS of all parameters (pooled across all remaining chains) was above 2272 for all HSMRF parameters and above 2737 for all GMRF parameters.

B.15 HIV Dynamics in Russia and Ukraine

We obtained the multiple sequence alignment for the *env* dataset of Vasylyeva *et al.* from the first author [151]. Following previous analyses, we employed an unpartitioned GTR+G substitution model. We used a UCLD clock model, with a Normal(-9.21034,2.34962) prior on the log-mean (corresponding to a prior 95% CI of [1e-6,1e-2] substitutions per year) and an Exponential(3) prior on the clock log-SD. We used RAxML [142], using the GTRCAT substitution model, and TreeDater [155], with a strict clock, to obtain a starting tree for analysis and an estimate of the age of the tree. We used this estimate (29.1) to set a relatively diffuse Normal(29.1,5.0) prior on the tree age, with a truncation at 18.0, the relative age of the oldest sample. To set our prior on the serial sampling

rate, we employ the empirical-Bayes strategy outlined above, resulting in a prior median sampling rate of 0.13. Back-of-the-envelope calculations of the sampling rate (using either the total number of infections or the numbers of samples out of the numbers of cases) suggest a rate on the order of 10^{-5}. Future work may elaborate whether this discrepancy is common and what, if any, its effect is.

We ran 4 chains for 500,000 iterations for both the GMRF and HSMRF models (with 335 moves per iteration), downsampling to every 100th sample. This is the equivalent of 167.5 million generations per chain in a program like BEAST [32], sampled every 33,500. We performed convergence checks as with Pygopodidae, resulting in 2 GMRF-based model run failing and 2 HSMRF-based model runs failing. Discarding these chains, the maximum rank-PSRF for the HSMRF analyses was 1.003, and for the GMRF it was 1.003. The rank-ESS of all parameters (pooled across all remaining chains) was above 664 for all HSMRF parameters and above 317 for all GMRF parameters.

In Figure B.23, we present posterior distributions of γ for our analyses. As there are rapid changes evident, one might expect that γ, especially for the GMRF-based model, would need to be large to explain these changes. This is in fact what is seen: for both models, the posterior distributions of γ are pulled up from the prior, and for the GMRF-based model the change is quite large. This also helps to explain the differences in the inferred $R_e(t)$ in the mid-1990s: since for the GMRF-based model γ is so large, it infers substantial variability in this time, while the HSMRF-based model with a smaller γ smooths out this variability.

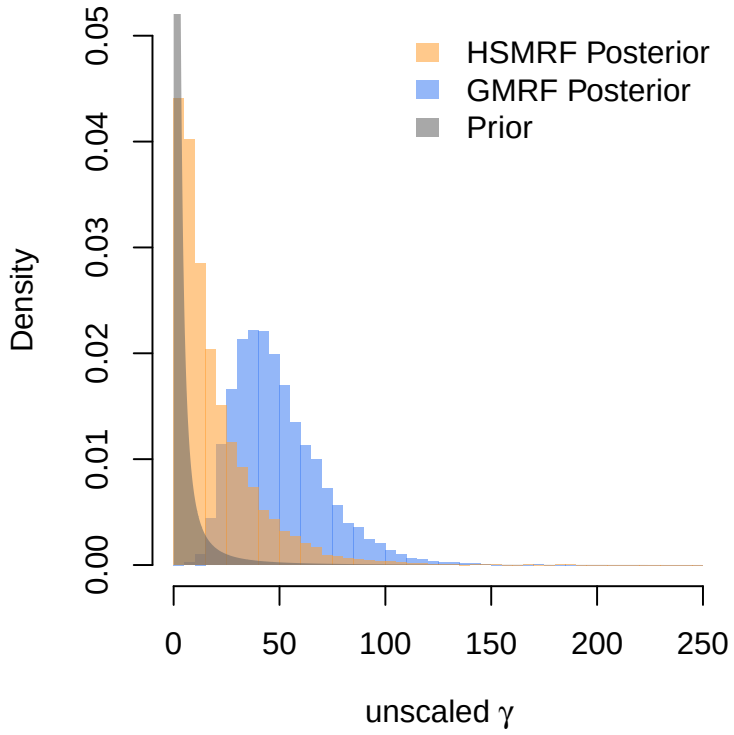


Figure B.23: Comparison of the global scale parameter γ for the HIV analysis. The grey curve is the halfCauchy(0,1) prior used on the unscaled γ, the orange histogram is the posterior distribution inferred by the HSMRF-based model, and the blue histogram the posterior distribution inferred by the GMRF-based model. The GMRF-based model requires a very large γ in order to accomodate the periods of rapid change inferred, while the HSMRF-based model does not require such a large value.

B.16 Comparison to BEAST

For time-varying birth-death process models, the current state-of-the-art is the birth-death skyline model as implemented in the BEAST 2 [7] package BDSKY [140]. For serially-sampled datasets, this package implements a model in terms of the compound parameters $\boldsymbol{R_e} = \boldsymbol{\lambda}/(\boldsymbol{\mu}+\boldsymbol{\phi r})$, $\boldsymbol{\delta} = \boldsymbol{\mu}+\boldsymbol{\phi r}$, and $\boldsymbol{s} = (\boldsymbol{\phi r})/(\boldsymbol{\mu}+\boldsymbol{\phi r})$, though in all cases it is assumed that $r = 1$. These parameters are all assumed to be iid in some number of intervals that the user defines. To compare our results to the state-of-the-art, we used BEAST 2.6 [7] to perform joint inference of phylogeny and phylodynamic parameters. Comparability cannot be perfect, since some parameters of the model are different, however, we attempted to keep priors similar where possible and biologically motivated otherwise. Due to some difficulty with clock model convergence in preliminary analyses, we placed a tighter

prior on the clock rate informed by our estimated clock rate from the `RevBayes` analyses. Our prior, a Lognormal(-4.9,0.587), corresponds to the clock rate being within approximately one order of magnitude of the clock rate estimated in our `RevBayes` analyses We assumed a single rate of becoming noninfectious, δ, and a single sampling proportion, s, analogous to our assumptions that μ and ϕ are constant. We placed a Lognormal(-2.272,0.073) prior on δ and a Beta(1,19) prior on s, reflecting our knowledge about the rate of becoming noninfectious (absent treatment) and the fact that we have a small sample out of a much larger epidemic. We placed Lognormal(0.0,0.821) priors on R_{ei}, reflecting that the effective reproductive number should likely be lower than 5, and the possibility that R_e may dip below 1.0. We employ a grid with 15 intervals, with break-points specified to occur at 1990,1991,...,2002,2003. Given that there are few infections inferred to have occurred post-2003 or prior to 1990 based on our analyses using the GMRF-based and HSMRF-based models, our most recent grid cell begins in 2011 and ends in 2003, while our oldest grid cell contains all times prior to 1990. Convergence diagnostics showed a maximum rank-based PSRF of 1.001 and a minimum rank-based ESS of 4932.

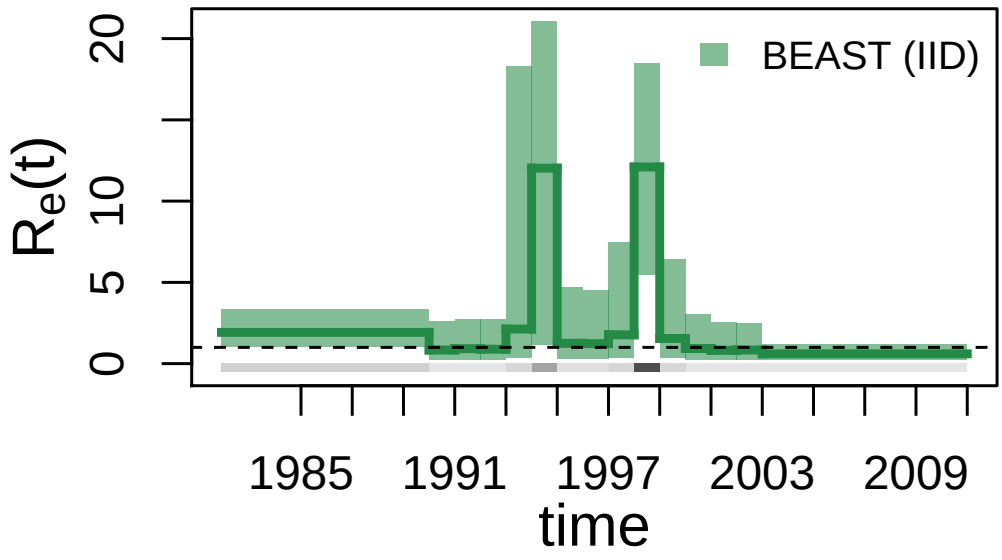


Figure B.24: Analyses of the HIV dataset with BEAST. Plotted are posterior median trajectories (dark lines) and 90% credible intervals (shaded regions). Time is plotted as calendar time. A line at $R_e = 1$ is provided for convenience, as below this threshold the epidemic cannot be sustained. In grey is a heatmap of the inferred divergence times.

The results of our BEAST analysis bear many similarities to the results from our GMRF-based model. Qualitatively, the pattern is similar, two peaks in $R_e(t)$ at approximately 1994 and 1998, each proceeded by a decrease, a low rate at the present, and a higher rate in the past. The 2ln Bayes Factor for a decrease at the end of the 1990s is 6.36 (strong support), while the 2lnBF in favor of an increase at the beginning of the decade is 6.09 (strong support). However, where the GMRF shows clear signs of an elevated rate for the entirety of the 1990s, the BEAST analysis shows a very strong decrease in the middle of the decade. The 2lnBF for the decrease after 1995 is 4.62 (positive support) and for the increase before 1998 is 5.79 (positive support). The inferred rates in the peak periods are implausibly high, with the estimated $R_e(t)$ at the peaks being much higher than in our HSMRF-based and GMRF-based model analyses at approximately 12.

The results from BEAST form one end of the spectrum of temporal smoothing, with the GMRF-based model in the middle and the HSMRF-based model on the opposite side. When all intervals are iid, there is no smoothing and thus no pooling of information among adjacent (or close) intervals, and any apparent change in the birth rate will be picked up. When intervals are temporally autocorrelated, these apparent changes are weighted against a prior belief that the rate is likely similar over small periods of time, requiring more evidence before a shift is inferred. The HSMRF-based model infers $R_e(t) > 1$ prior to 1991, where both the less-smoothed GMRF-based model and unsmoothed BEAST analysis suggest $R_e(t) < 1$ prior to the first increase. The HSMRF-based model infers a high $R_e(t)$ throughout most of the 1990s, while the GMRF-based model infers two peaks of $R_e(t)$ with a notable decrease between them, and the BEAST analysis exaggerates this peak (and shows a posterior probability of 0.3 to 0.4 that $R_e(t) < 1$ in this range). In the case of the HIV results, smoothing and sharing information across intervals produces analyses more concordant with other lines of epidemiological research (which do not show these decreases).

B.17 Marginal likelihoods and Metropolis Coupled MCMC

As our MCMC sampler is somewhat atypical in phylogenetics, and may thus be unfamiliar to readers, we now examine how the sampler interacts with common variants of MCMC in phylogenetics. Calculating marginal likelihoods via path sampling [80] or stepping stone [40] both require running a number of MCMC chains, each of which raises the likelihood of the model to an increasing power

$0 \leq \beta \leq 1$. Metropolis-coupling employs multiple MCMC chains where all but one (the so-called cold chain) have their target posteriors raised to a power $0 \leq \beta < 1$. While these are not issues for the standard Metropolis-Hastings moves in most phylogenetic inference software, they can pose problems for other types of samplers such as Gibbs samplers. Employing a Gibbs Sampler requires that the conditional distributions of the parameters being sampled remain unchanged and cannot be raised to any power. This means that the distribution on all parameters above and those directly below $\boldsymbol{\sigma}$ and γ in the model DAG (see Figure B.18) cannot be raised to any power $\beta \neq 1$. The elliptical slice sampler requires that the prior on $\boldsymbol{\Delta}$ be multivariate normal, so nothing above $\boldsymbol{\Delta}$ in the model DAG can be raised to a power. Taken together, these imply that anything below $\boldsymbol{\Delta}$ can be raised to a power, and thus marginal likelihoods may be computed, but that these samplers are incompatible with Metropolis-coupled MCMC. For both Gibbs samplers and for the elliptical slice sampler, `RevBayes` will throw an error if the user attempts to raise any distributions to unallowed powers.

B.18 Implementation details

In `RevBayes`, a distinction is made between the EBD, which has no serial sampling, and the episodic birth-death sampling treatment process, which allows for serial sampling and conditional death upon sampling (treatment). That is, the EBD is the model of Stadler (2011) and Höhna (2015), whereas the EBDSTP is the model of Gavryushkina *et al.* (2014) [136, 60, 49]. When there is no serial sampling in the model ($\phi(t) = 0$) and there are tips only at the present ($\boldsymbol{t_\Phi} = t_\Phi = 0$), the models are eactly equivalent. In these cases, we employ the simpler EBD model in our Rev scripts, though we note that changing back is a simple matter of replacing calls to `dnEBDP` with calls to `dnBDSTP` in `RevBayes`.

Algorithm 1 Elliptical Slice Sampler

Input: $\boldsymbol{\Delta}^t$ the value of $\boldsymbol{\Delta}$ at the current MCMC step, $\boldsymbol{\sigma}^t$ the standard deviations of $\boldsymbol{\Delta}^t$, function $f(x)$ that returns the (unnormalized) log-probability density for the downstream DAG for $\boldsymbol{\Delta} = \boldsymbol{x}$, n the size of $\boldsymbol{\Delta}$.
Initialization:
draw $u \sim \text{Uniform}(0, 1)$
set $y \leftarrow f(\boldsymbol{\Delta}^t) + \ln(u)$
for i in 1:n **do**
  draw $\nu_i \sim \text{Normal}(0, (\sigma_i^t)^2)$
end for
draw $\theta \sim \text{Uniform}(0, 2\pi)$
set $L \leftarrow \theta - 2\pi$
set $R \leftarrow \theta$
for i in 1:n **do**
  set $x_i \leftarrow \Delta_i^t \cos(\theta) + \nu_i \sin(\theta)$
end for
Run sampler:
while $f(x) \leq y$ **do**
  if $\theta > 0$ **then**
    $R \leftarrow \theta$
  else
    $L \leftarrow \theta$
  end if
  draw $\theta \sim \text{Uniform}(L, R)$
  for i in 1:n **do**
    set $x_i \leftarrow \Delta_i^t \cos(\theta) + \nu_i \sin(\theta)$
  end for
end while
set $\boldsymbol{\Delta}^{t+1} = \boldsymbol{x}$

Algorithm 2 Gibbs sampler for HSMRF

Input: $\boldsymbol{\Delta}^t$ the value of $\boldsymbol{\Delta}$ at the current MCMC step, $\boldsymbol{\sigma}^t$ the standard deviations of $\boldsymbol{\Delta}^t$, γ^t the global scale parameter, ζ the global scale hyperparameter, n the size of $\boldsymbol{\Delta}$.
for i in 1:n **do**
  draw $\psi_i \sim \text{InverseGamma}(1, 1 + 1/\sigma_i^2)$
  draw $(\sigma_i^{t+1})^2 \sim \text{InverseGamma}(1, \psi_i^{-1} + (\Delta_i^t)^2/(2(\gamma^t)^2\zeta^2))$
end for
draw $\xi \sim \text{InverseGamma}(1, 1 + 1/(\gamma^t)^2)$
draw $(\gamma^{t+1})^2 \sim \text{InverseGamma}(1, \xi^{-1} + 1/(2\zeta^2) \sum_{i=1}^n \Delta_i^2/\sigma_i^2)$

Algorithm 3 Gibbs sampler for GMRF

Input: $\boldsymbol{\Delta}^t$ the value of $\boldsymbol{\Delta}$ at the current MCMC step, γ^t the global scale parameter, ζ the global scale hyperparameter, n the size of $\boldsymbol{\Delta}$.
draw $\xi \sim \text{InverseGamma}(1, 1 + 1/(\gamma^t)^2)$
draw $(\gamma^{t+1})^2 \sim \text{InverseGamma}(1, \xi^{-1} + 1/(2\zeta^2) \sum_{i=1}^{n} \Delta_i^2)$

Appendix C

APPENDIX TO: HOW TRUSTWORTHY IS YOUR TREE? BAYESIAN PHYLOGENETIC EFFECTIVE SAMPLE SIZE THROUGH THE LENS OF MONTE CARLO ERROR

C.1 Visualizing convergence of a single chain

To explore the uncertainty of estimates from a single MCMC chain through time, we employ a block-bootstrap approach in which we resample from the MCMC sample [113, 144]. This approach requires a vector of subsample sizes, $n_1, \ldots, n_s$. For a given subsample length n_i, we define the batch size to be $b = \lfloor \sqrt{n_i} \rfloor$ and the number of batches $a = \lfloor n_i / b \rfloor$. The summary tree is computed for the first ab samples of the real chain, and then for some number of bootstrap replicates r we re-estimate the summary tree. We use block-bootstrapping to preserve autocorrelation in the samples. Thus, for each of the r bootstrap replicates (at a given n_i), we draw a starting indices uniformly on $1, \ldots, n - b + 1$, and concatenate the resulting a blocks of length b into a bootstrap replicate chain. Then we compute the median RF distance from the real-chain-subsample summary tree to the r bootstrap replicates, as well as the 5th and 95th percentiles. As the longer subsamples of the real chain include the shorter subsamples (all real-chain subsamples start at the first sample), this procedure allows us to track how the summary tree converges over the course of the MCMC run. We can similarly track split or topology probabilities over the course of the run, in which case we use the ASDSF and the Euclidean distance, respectively, to compare the real chain estimates to the bootstrap estimates.

In Figure C.1, we explore the convergence behavior of chain 1 of the Paroedura dataset using three summary measures. These measures are the ASDSF between bootstrap and real-chain split probabilities, the Euclidean distance between bootstrap and real-chain estimates of the vector of split probabilities, and the RF distance between bootstrap and real-chain summary trees. While both the split probabilities and tree probabilities appear to converge relatively well (the ASDSF quickly declines below the usual field-standard for good convergence of 0.01), there is still consid-

erable Monte Carlo variability evident in summary trees. This pattern holds across all datasets and almost all chains (Supplemental Figures XX-YY), indicating that classical ASDSF cutoffs for convergence of chains are not guarantees of the convergence of summary trees from those chains. We note, however, that this can only ever help determine whether estimates from a single run have stabilized. To diagnose issues such as convergence to a local mode, practitioners must run multiple chains. We note that this is suggested as standard practice [81] and is a widely available option, including in BEAST [143] and RevBayes [63], and is the default in MrBayes [125].

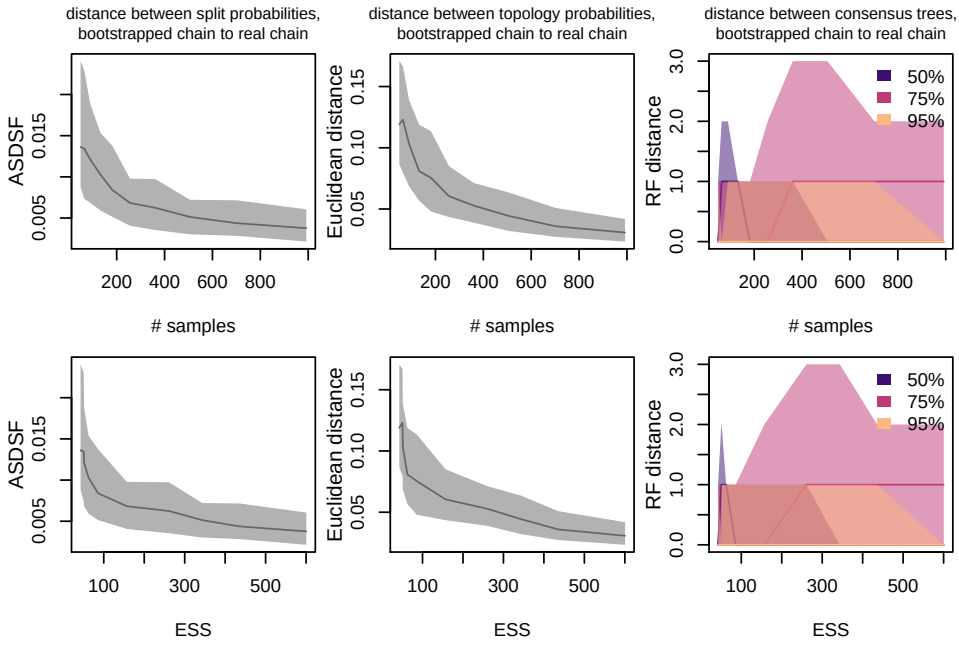


Figure C.1: Monte Carlo error visualized over the length of one chain of the Paroedura dataset from [127]. The top and bottom rows are equivalent except that the x-axis is scaled to the absolute number of MCMC samples (top), and the split-frequency ESS (bottom). The left column plots the ASDSF between bootstrap replicate estimates of the split probabilities and the split probabilities estimated from the first n_i samples of the chain. The central column plots the Euclidean distance between bootstrap replicate estimates of the (vector of) tree probabilities and the (vector of) tree probabilities estimated from the first n_i samples of the chain. The right column plots the RF distance between bootstrap replicate estimates of consensus trees and the consensus trees estimated from the first n_i samples of the chain. The different colors show consensus trees constructed with different minimum inclusion probabilities of splits, such that the purple curve shows the classical MRC tree, and the yellow curve shows a consensus tree containing only splits with 95% probability. In all cases, the dark lines are the median and the shaded region is the central 90% range.

C.2 More efficient simulated phylogenetic MCMC

Recall that our simulated phylogenetic MCMC is based on real-data phylogenetic posterior distributions, potentially truncated. This consists of a vector of trees, $\boldsymbol{\tau}$, and an associated probability mass function, $\widehat{\Pr}(\boldsymbol{\tau})$ (we use the hat as a reminder that this target is based, indirectly, on real data). We use NNIs to move between tree topologies, by uniformly drawing a tree from the set of neighbors, $N(\Psi)$. Then we accept or reject according to the estimated topology probability $\widehat{\Pr}(\Psi)$ (any tree not in the real-data posterior has probability 0). If we redefine $N(\Psi)$ to instead be the NNI neighbors of Ψ with positive probability (*e.g.* $\Psi \in \boldsymbol{\tau}$, which also requires far less storage), we can instead simulate the proposal in two steps. First, draw $u \sim \text{Uniform}(0, 1)$, and if $u < |N(\Psi)|/|N|$ (where $|N|$ is the number of NNI neighbors of any tree in the posterior), we draw our proposed tree Ψ^* uniformly at random from $N(\Psi)$ and set $A = \min(1, \Pr(\Psi^*)/\Pr(\Psi))$. Otherwise, we have drawn a tree outside the set of supported neighbors of Ψ ($\Psi^* \notin \boldsymbol{\tau}$) and we do not need to specify which tree, as in this case it has probability 0 and so $A = 0$ and we will always reject the proposal. Then we accept or reject the move with probability A and proceed normally. This approach requires us only to know what trees in the support of the posterior are neighbors, which for real phylogenetic posterior distributions is a much smaller set than the set of all NNI neighbors.

C.3 Explicit definitions and derivations of tree ESS measures

In the following sections, we present more thorough derivations of the frechetCorrelationESS and approximateESS use in the main text, and derivations for 6 other potential tree ESS methods. The ten total methods fall into the same three categories as the main text and are (using $*$ to denote those appearing in the main text):

- ESS measures based on Frechét-like generalizations of Equation 4.5 to trees

  $*$The Frechét Correlation ESS (frechetCorrelationESS)

  The split frequency ESS (splitFrequencyESS)

- ESS measures based on projecting the tree to a single dimension and computing the ESS of that using standard univariate approaches

The folded rank-medoid ESS (foldedRankMedioidESS)

∗The median pseudo-ESS (medianPseudoESS)

∗The minimum pseudo-ESS (minPseudoESS)

The total distance ESS (totalDistanceESS)

The classical multidimensional scaling ESS (CMDSESS)

- Ad-hoc ESS measures

∗The approximate ESS (splitFrequencyESS)

The unsmoothed bootstrap jump-distance ESS (jumpDistanceBootstrapUnsmoothedESS)

The (smoothed) bootstrap jump-distance ESS (jumpDistanceBootstrapESS)

C.3.1 Calculating the ESS by generalizing previous definitions

In this section, we provide more in-depth derivations of our two ESS approaches that generalize Equation 4.5 using concepts borrowed from the notions of Frechét mean and Frechét variance. For a continuous random variable X, the sample mean, $\bar{x}$, minimizes the sum of squared deviations to all sampled points. The Frechét mean generalizes this concept to other metric spaces and higher dimensions by keeping the idea of minimizing the sum of squared distances. The Frechét mean of a set of samples is,

$$\bar{x} = \text{argmin}_{\bar{x}} \int_{x \in X} \sum_{i=1}^{n} (x_i - \bar{x})^2 \mathrm{d}\bar{x},$$

where $d(\cdot, \cdot)$ is a distance metric. The Frechét mean may not be unique, in which case the collection of values that minimize the sum of squared distances are known as Karcher means. Where the variance is the average squared deviation from the mean, the Frechét variance is the average squared distance from the Frechét mean. In the case where X is continuous and one-dimensional and $d(\cdot, \cdot)$ is the Euclidean distance, the Frechét mean is the mean and the Frechét variance is the variance.

These definitions take some adaptation to the setting considered here. If we were considering phylogenetic trees with branch lengths, we could use the definitions directly because BHV space is a complete metric space. For the discrete tree topology space, one can think of an "RF space" where topologies are encoded as a binary vector. For a tree with n_{taxa} tips, there are $2^{n_{\text{taxa}}} - n_{\text{taxa}}$ possible

non-trivial splits. Thus we can represent a tree as a vector of length $2^{n_{\text{taxa}}} - n_{\text{taxa}}$ which has a one entry exactly when the corresponding split is present in the tree. There are $n_{\text{taxa}} - 3$ non-trivial splits in a fully resolved tree, thus the maximum sum of entries in such a vector representation is $n_{\text{taxa}} - 3$. The Hamming distance (or equivalently the Manhattan distance) between two trees in RF space is the classical RF distance. This also means that we only need to consider coordinates in RF space which are non-zero in at least one tree in the set. As we use RF distances in this chapter, all this work can be seen to live in RF space. As this does not define a complete metric space, the following generalizations are best thought of as Frechét-like.

The frechetCorrelationESS

In this section, we will explore how to generalize the sum-of-correlations ESS of Equation 4.8 to trees. To do so, we first review several key identities, including relationships between pairwise distances and both covariance and variance. For two real-valued variables, X and Y, we can express the expected squared Euclidean distance as a function of the variances, the difference in means, and the covariance. For convenience, we will write $\Delta^2 = (X - Y)^2$. Then, taking advantage of the fact that $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, we get,

$$\begin{aligned}
\mathbb{E}[\Delta^2] &= \mathbb{E}[(X - Y)^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[XY] + E[Y^2] \\
&= \text{Var}(X) + E[X]^2 + \text{Var}(Y) + E[Y]^2 - 2(\text{Cov}(X, Y) + E[X]E[Y]) \\
&= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y) + (\mathbb{E}[X] - \mathbb{E}[Y])^2 \\
&\geq \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)
\end{aligned}$$

Where the last line follows because $(\mathbb{E}[X] - \mathbb{E}[Y])^2 > 0$. The last two lines of this equation block rearrange to:

$$\text{Cov}(X, Y) = \frac{1}{2}(\text{Var}(X) + \text{Var}(Y) - \mathbb{E}[\Delta^2] + (\mathbb{E}[X] - \mathbb{E}[Y])^2). \tag{C.1}$$

If $\mathbb{E}[X] \approx \mathbb{E}[Y]$, then we have the approximate equality,

$$\mathrm{Cov}(X, Y) \approx \frac{1}{2}(\mathrm{Var}(X) + \mathrm{Var}(Y) - \mathbb{E}[\Delta^2]) \tag{C.2}$$

It is worth noting that the sum of pairwise distances for a sample of a random variable can be used to estimate its variance.

$$\widehat{\mathrm{Var}}(X) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{2\binom{n}{2}}\sum_{j>i}(x_i - x_j)^2. \tag{C.3}$$

To show this, first we need that,

$$\sum_i \sum_j (x_i - x_j)^2 = 2n \sum_i (x_i - \bar{x})^2. \tag{C.4}$$

This can be shown as follows,

$$\begin{aligned}
\sum_i \sum_j (x_i - x_j)^2 &= \sum_i \sum_j ((x_i - \bar{x}) - (x_j - \bar{x}))^2 \\
&= \sum_i \sum_j (x_i - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x}) + (x_j - \bar{x})^2 \\
&= n\sum_i (x_i - \bar{x})^2 + n\sum_j (x_j - \bar{x})^2 - 2\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2\sum_i \sum_j (x_i x_j - x_i\bar{x} - x_j\bar{x} + \bar{x}^2) \\
&= 2n\sum_i (x_i - \bar{x})^2 - 2(n^2\bar{x}^2 - n^2\bar{x}^2 - n^2\bar{x}^2 + n^2\bar{x}^2) \\
&= 2n\sum_i (x_i - \bar{x})^2.
\end{aligned}$$

Having shown that Equation C.4 is true, from it we can get,

$$\frac{1}{2n^2}\sum_i \sum_j (x_i - x_j)^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2. \tag{C.5}$$

Lastly, note that

$$\frac{1}{2n^2}\sum_i\sum_j(x_i - x_j)^2$$

is simply half the average squared distance between all pairs (i, j). We can also get this average squared distance by looking at only pairs $(i > j)$ as,

$$\frac{1}{2\binom{n}{2}}\sum_{i>j}(x_i - x_j)^2.$$

Then we simply substitute this into Equation C.5 and we get Equation C.3. (We note this relationship can also be derived analogously to Equation C.1 by letting X and Y be IID.) Letting $d(\cdot, \cdot)$ be a distance measure, we can write a Frechét-like generalization of Equation C.3 as,

$$\widehat{\mathrm{Var}}(X) = \frac{1}{2\binom{n}{2}}\sum_{j>i} d(x_i, x_j)^2 \tag{C.6}$$

In the same way that mean and variance can be generalized using the Frechét mean and variance, Equation C.1 allows us to generalize covariance to a Frechét covariance. This is accomplished by defining $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ to be the Frechét variances, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ to be the Frechét means, and redefining $\Delta^2 = d(X, Y)^2$. Note that $\mathbb{E}[\Delta^2]$ is simply the average distance between X and Y,

$$\mathbb{E}[\Delta^2] = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i)^2. \tag{C.7}$$

Thus, we can get a single-dimensional summary of the dependency of two random variables, and compute a single ESS measure for a high-dimensional object. However, as discussed in the main text, since we are not working in BHV space, we term this a Frechét-like covariance. Equation C.2 is particularly useful in this circumstance because it avoids the need to compute the topological mean of a set of trees, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, which may not be unique. Equations C.6, and C.7 are also useful, and they allow us to compute everything we need for Equation C.2 from the sample distance matrix.

To compute the ESS for trees using Equations 4.8 and C.2, we specifically need to be able to compute the Frechét-like autocorrelation ρ_s of the chain at time lag t, and thus X and Y are actually

X_t and X_{t+s}. If the chain is stationary, then the mean does not change over time, and we should expect that $\mathbb{E}[X_t] \approx \mathbb{E}[X_{t+s}]$, and the use of Equation C.2 rather than Equation C.1 is justified. If instead of trees we had a time series of a Euclidean variable $\boldsymbol{X}$, the estimated autocorrelation is the sample Pearson correlation coefficient between samples at the given time lag,

$$\hat{\rho}_s = \frac{\widehat{\mathrm{Cov}}(\boldsymbol{X}_t, \boldsymbol{X}_{t+s})}{\sqrt{\widehat{\mathrm{Var}}(\boldsymbol{X}_t)\widehat{\mathrm{Var}}(\boldsymbol{X}_{t+s})}}. \tag{C.8}$$

For trees, instead we use Frechét-like variances and Equation C.2 to get an approximation to the Frechét-like covariance, and plug this into Equation C.8,

$$\hat{\rho}_s = \frac{\frac{1}{2}(\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t) + \widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s}) - \hat{\mathbb{E}}[\Delta^2])}{\sqrt{\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t)\widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s})}} \tag{C.9}$$

Once we obtain our estimates $\hat{\rho}_s$, we use Equation 4.8 to estimate the ESS. We call this approach the Frechét-like correlation ESS, or frechetCorrelationESS. In this set up, we must estimate $\mathrm{Var}(\boldsymbol{\tau}_t)$, $\mathrm{Var}(\boldsymbol{\tau}_{t+s})$, and $\mathbb{E}[\Delta^2]$, which we compute using Equations C.6 and C.7. Let n be the total number of tree samples, $\boldsymbol{\tau}$ be the vector of tree samples, and let us define $\boldsymbol{\tau}_t, \boldsymbol{\tau}_{t+s}$ to be the pair of vectors of trees separated by time lag s. Then, we can compute the terms as,

$$\widehat{\mathrm{Var}}(\boldsymbol{\tau}_t) = \frac{1}{2\binom{n}{2}} \sum_{i=1}^{n-s} \sum_{j=i+1}^{n-s} d(\tau_i, \tau_j)^2,$$

$$\widehat{\mathrm{Var}}(\boldsymbol{\tau}_{t+s}) = \frac{1}{2\binom{n}{2}} \sum_{i=s+1}^{n} \sum_{j=i+1}^{n} d(\tau_i, \tau_j)^2,$$

$$\hat{\mathbb{E}}[\Delta^2] = \frac{1}{n-s} \sum_{i=1}^{n-s} d(\tau_i, \tau_{i+t})^2.$$

Given that most distances will appear in several calculations, it is most efficient to pre-compute the sample distance matrix $\boldsymbol{D}$ with $D_{ij} = d(\tau_i, \tau_j)$.

In practice, while the above definition could theoretically permit ESS $> n$, we enforce frechetCorrelationESS $\leq n$.

The splitFrequencyESS

Our next generalization approach we term the split frequency ESS, or splitFrequencyESS. This is a generalization of the univariate [152] estimator of the effective sample size, which we will call the batch means ESS, which we will now review. The batch means ESS is based on the relationship,

$$\widehat{\text{ESS}} = n\frac{\hat{\sigma}_\pi^2}{\hat{\lambda}_L^2}, \tag{C.10}$$

where $\hat{\lambda}_L^2$ is an estimate of the limiting variance (σ_{lim}^2) and $\hat{\sigma}_\pi^2$ is the estimate of the posterior variance computed from the samples. Let $\boldsymbol{X}$ be the vector of MCMC samples, B be a batch size (define a to be the according number of batches), and $\boldsymbol{Y}$ the vector of batch means, with Y_i the mean in the ith batch (subset of the chain). Then, [152] define,

$$\hat{\lambda}_B^2 = B/(a-1)\sum_{i=1}^{a}(Y_i - \bar{X})^2.$$

To use the batch means approach in practice, the batch size must scale with n. Following [152], we use a batch size $b = \lfloor n^{1/2} \rfloor$. Then, the estimate of the limiting variance from the batch-means approach is given by,

$$\hat{\lambda}_L^2 = 2\hat{\lambda}_b^2 - \hat{\lambda}_{b/3}^2,$$

where $\hat{\lambda}_{b/3}^2$ is computed using a batch size $\lfloor b/3 \rfloor$ [Equation 5, 152].

To apply the batch means ESS to trees, we represent trees as vectors of splits. We now walk through this generalization. If the posterior distribution contains S non-trivial splits, $s_1, \ldots, s_S$, then we transform the vector of trees into a matrix, where in each row we represent that tree as its vector of coordinates in RF-space $\boldsymbol{X}$. Namely,

$$X_{ij} = \begin{cases} 1 & \text{if } s_j \in \tau_i, \\ 0 & \text{otherwise,} \end{cases}$$

As our distance metric we take the Euclidean distance, so the Frechét mean is the arithmetic mean, $\bar{\boldsymbol{X}}$. We choose a batch size b that scales with n (we use $\lfloor n^{1/2} \rfloor$). Again, $a = \lfloor n/b \rfloor$ is the

number of batches, and $\boldsymbol{Y}$ the matrix of batch means, with $\boldsymbol{Y}_i$ the vector of means in the ith batch (subset of the chain). For a fixed split j,

$$Y_{ij} = \frac{1}{b} \sum_{k=(i-1)b-1}^{ib} X_{kj}.$$

We use a Frechét-based generalization for $\hat{\lambda}_b^2$, namely,

$$\hat{\lambda}_b^2 = \frac{b}{a-1} \sum_{i=1}^{a} d(\boldsymbol{Y}_i, \bar{\boldsymbol{X}})^2.$$

Similarly, We use a Frechét-based generalization for $\hat{\sigma}_\pi^2$,

$$\hat{\sigma}_\pi^2 = \frac{1}{n} \sum_{i=1}^{n} d(\boldsymbol{X}_i, \bar{\boldsymbol{X}})^2.$$

Given these modified $\hat{\lambda}_b^2$ and $\hat{\sigma}_\pi^2$, we use Equation C.10 to compute the ESS. We note that all the batch means,$\boldsymbol{Y}_i$, are in fact the split frequencies (or estimates split probabilities) in those batches, and the global mean, $\bar{\boldsymbol{X}}$ are the marginal split frequencies across the entire posterior distribution. We thus call this the split frequency ESS, or splitFrequencyESS.

Approaches to calculating the ESS by projecting the tree to a single dimension

All dimension-reduction approaches entail first transforming the trees into a 1-D representation, then taking the ESS of that. We use the The R package `coda` [112] implementation of the ESS, and before we discuss our approaches we first outline how it works.

The ESS computation in `coda`

The R package `coda` [112], commonly used for MCMC diagnostics, fits an autoregressive model to the MCMC samples to estimate the ESS. Specifically, the `coda` estimate of the ESS, which we will call the power spectrum ESS, is,

$$\text{ESS} = n \frac{\hat{\sigma}_\pi^2}{\widehat{\Gamma(0)}}, \tag{C.11}$$

where $\widehat{\Gamma(0)}$ is an estimate of the power spectrum at frequency 0 [see 58, for details], and $\hat{\sigma}_{\pi}^2$ is the estimate of the posterior variance computed from the samples. This follows from Equation 4.5 and the fact that the standard error of the mean of a covariance-stationary process is $\Gamma(0)/n$ [58]. The power spectrum at 0, $\Gamma(0)$, can be linked to the autoregressive parameters by,

$$\Gamma(0) = \frac{\sigma_e^2}{\left(1 - \sum_{i=1}^{p} \phi_i\right)^2},$$

where σ_e^2 is the error variance (the variance unexplained by the autoregressive model, also called the noise variance), also called the noise variance [156]. In practice, `coda` estimates $\widehat{\Gamma(0)}$ using an autoregressive process of unknown order p. With an estimated order, $\hat{p}$, a fitted set of autoregression coefficients $\phi_1, \ldots, \phi_{\hat{p}}$, and an estimated error variance $\hat{\sigma}_e^2$, the estimate is,

$$\widehat{\Gamma(0)} = \frac{\hat{\sigma}_e^2}{\left(1 - \sum_{i=1}^{\hat{p}} \hat{\phi}_i\right)^2}.$$

The foldedRankMedioidESS

Vehtari *et al.* introduce two new approaches for computing ESS measures, one of which, the folded rank-transformed ESS, can be co-opted for phylogenies relatively painlessly [153]. For a real-valued parameter x, this ESS is computed for the transformed variable z, where there are a few layers of transformation.

$$\begin{aligned}
\zeta &= |x - \text{median}(x)| \\
r &= \text{rank}(\zeta) \\
z &= \Phi^{-1}\left(\frac{r - 3/8}{n - 1/4}\right)
\end{aligned}$$

The first step is to "fold" the variable, and track the absolute deviations from the median. Then a rank tranformation is applied, which stabilizes for any extreme deviations. Lastly, a Normal inverse-CDF is applied (with an offset). [153] then take the folded rank-transformed ESS to be the ESS of z using Equation 4.8. In the case there is not a unique medioid tree, we compute the ESS using all possible reference trees and take the minimum.

To use this approach for trees, we make a few generalizations, and we call the resulting ESS the foldedRankMedioidESS. First, we replace the sample median with the medioid, which is a generalization of the median to higher dimensions. Specifically, the medioid tree is the (sampled) tree with the minimum sum of distances to all other sampled trees, $\text{medioid}(\boldsymbol{\tau}) = \underset{\Psi \in \boldsymbol{\tau}}{\text{argmin}} \sum_i d(\Psi, \tau_i)$. Then, we replace the absolute divergence with the distance (in one dimension, these are equivalent). The foldedRankMedioidESS is computed for the transformed variable z, where we obtain z through the following tranformations.

$$\begin{aligned}\zeta &= d(\tau, \text{medioid}(\tau)) \\ r &= \text{rank}(\zeta) \\ z &= \Phi^{-1}\left(\frac{r - 3/8}{n - 1/4}\right)\end{aligned}$$

The totalDistanceESS

As an alternative to picking a specific reference tree, as in the foldedRankMedioidESS, medianPseudoESS, or minPseudoESS, we also consider an ESS based on the sum of distances between each tree and all the other trees. In this setup, we compute the ESS of the transformed variable y, defined by $y_i = \sum_{j=1}^{n} d(\tau_i, \tau_j)$. We call this the total distance ESS, or totalDistanceESS.

The CMDSESS

We also consider multidimensional scaling of the (squared) distance matrix $\boldsymbol{D}^2$ to compute an ESS. Specifically, we use classical multidimensional scaling. This approach seeks to find a matrix $\boldsymbol{Y}$ which minimizes a loss function called the strain between $\boldsymbol{Y}$ and the $\boldsymbol{B}$, a doubly centered version of $\boldsymbol{D}$. As our new variable we take the first column of the new matrix, $\boldsymbol{Y}_1$ We call this the classical multidimensional scaling ESS, or CMDSESS.

C.3.2 Ad-hoc approaches to computing the ESS

If we define s_0 to be the time lag at which samples from our MCMC become independent of each other, then we could somewhat conservatively estimate the ESS as n/s_0. This approach can be seen

as a naive implementation of the idea that the effective sample size is the hypothetical number of independent samples contained within the n MCMC samples. This is not tied to any mathematical definition of the ESS, and is not without problems. For one, the approach is expected to be overly conservative, as it effectively discards all samples in between an estimated autocorrelation time, whereas classical ESS approaches keep fractions of all samples. Additionally, in this approach ESS can take on only n distinct values because it is guaranteed that s_0 is an integer between 1 and n (inclusive). The approximate ESS of [77] can be seen as one approach to overcoming these limitations, as it requires estimating s_0 then uses identities about expected distances. In the following sections, we first explore a thorough derivation of the approximate ESS, then we consider alternatives for estimating s_0 and simply using this to estimate the ESS directly, $\widehat{\text{ESS}} = n/\hat{s}_0$.

The approximate ESS

[77] define a measure they call the approximate ESS, which is computed in three steps. First, s_0 is estimated using the curve of jump distances as described below. Then, by restricting to samples at time lags larger than s_0, the expected squared distance between any pair of trees drawn IID from the posterior, $\mathbb{E}[\Delta^2_{\text{iid}}]$, is estimated. Lastly, the approximate ESS is computed by finding the equivalent number of IID samples such that the expected squared distance matches the observed average squared distance from the MCMC samples. We first define the key equation, then walk through the steps required to compute this measure. Let $\mathbb{E}[\Delta^2_{\text{iid}}]$ be the expected squared distance between any two trees drawn iid from the posterior (as in Equations C.6 or C.7). [77] define the approximate ESS using the equation,

$$\frac{\text{ESS}(\text{ESS}-1)}{\text{ESS}^2}\hat{\mathbb{E}}[\Delta^2_{\text{iid}}] = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d(\tau_i,\tau_j)^2}{n^2} \tag{C.12}$$

Note that our presentation of this equation is somewhat different from the way it is presented in [77], in that both sides of this equation are larger than in [77] by a factor of 4, and that [77] use averages of distances at time lags on the RHS. These differences do not affect the values calculated, but we hope that this representation makes understanding the basis for using it to estimate ESS somewhat easier. As the RHS is simply the observed average squared distance, we can denote it

$\overline{\Delta^2}$ and solve Equation C.12 for ESS giving,

$$\frac{2\,\mathrm{ESS}(\mathrm{ESS}-1)}{\mathrm{ESS}^2}\hat{\mathbb{E}}[\Delta^2_{\mathrm{iid}}] = \overline{\Delta^2}$$

$$\mathrm{ESS} = \frac{1}{1-\overline{\Delta^2}/\hat{\mathbb{E}}[\Delta^2_{\mathrm{iid}}]}$$

The RHS of the equation is simply the average value of the matrix of all squared pairwise distances in the MCMC samples. To see this, note that the average value in the observed $n \times n$ matrix of squared distances is,

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d(\tau_i,\tau_j)^2}{n^2} = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d(\tau_i,\tau_j)^2}{n^2}$$

This follows because the upper and lower triangular portions of the matrix are symmetric (that is, $d(\tau_i,\tau_j) = d(\tau_j,\tau_i)$), which allows us to sum only the lower triangular portion and doubling the value, and because $d(\tau_i,\tau_i) = 0$, allowing us to ignore the diagonal. Thus, we can consider the RHS the *observed* average squared distance, while the LHS will prove to be an expectation of the average squared distance, assuming all trees were drawn independently. The idea, then, is that we can solve for the effective sample size by equating the observed average squared distance from dependent MCMC samples with the expected average squared distance from IID samples.

The LHS of Equation C.12 can be derived as follows. For ESS trees drawn iid from the posterior, there are $\binom{\mathrm{ESS}}{2}$ unique comparisons between τ_i and τ_j $(i \neq j)$, each of which has expectation $\mathbb{E}[\Delta^2_{\mathrm{iid}}]$. In the full ESS $\times$ ESS matrix of squared distances, each of these comparisons shows up twice, and the comparisons on the diagonal are all 0. Thus, the sum of this hypothetical matrix is $2\binom{\mathrm{ESS}}{2}\mathbb{E}[\Delta^2_{\mathrm{iid}}]$, and the expected average value of this hypothetical matrix is,

$$\frac{2\binom{\mathrm{ESS}}{2}}{\mathrm{ESS}^2}\mathbb{E}[\Delta^2_{\mathrm{iid}}] = \frac{\mathrm{ESS}(\mathrm{ESS}-1)}{\mathrm{ESS}^2}\mathbb{E}[\Delta^2_{\mathrm{iid}}].$$

Equation C.12 uses $\hat{\mathbb{E}}[\Delta^2]$ instead of $\mathbb{E}[\Delta^2]$ as $\mathbb{E}[\Delta^2]$ is an unknown quantity which must be estimated.

[77] use the curve of jump distances to estimate s_0 as described in the next paragraph, and define their estimate of $\mathbb{E}[\Delta^2_{\mathrm{iid}}]$ by taking the average squared distance of all samples with time lags

of at least s_0. In other words,

$$\hat{\mathbb{E}}[\Delta^2_{\text{iid}}] = \frac{1}{\binom{n-s_0+1}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(\tau_i, \tau_j)^2 \mathbb{I}(|i-j| \geq s_0)$$

In the regime where it is estimated that $s_0 > n$, [77] suggest using the largest observed squared distance instead,

$$\hat{\mathbb{E}}[\Delta^2_{\text{iid}}] = \max_{i,j}(d(\tau_i, \tau_j)^2).$$

[77] estimate s_0 by fitting a curve $J(s)$, which is essentially a smoothed version of the observed curve of average squared distances at increasing time lags s. If we define J_s to be the average squared distance among all MCMC samples at time lag s, then $J_s = \mathbb{E}[d(\boldsymbol{\tau}_t, \boldsymbol{\tau}_{t+s})^2]$ and $J(s)$ is a smoothed version of J_s. The functional form assumed is $J(t) = \beta_0(1 - e^{-t/\beta_1})$, which starts at 0 for $s = 0$ and has an asymptote at β_0 as $s \to \infty$. $J(s)$ is fit to J_s via least squares. Then, some threshold α is chosen, and s_0 is defined to be the first time lag s for which J_s is within $\alpha \times 100\%$ of its asymptote,

$$s_0 = \operatorname*{argmin}_s \ J_s \geq \beta_0(1-\alpha).$$

Then s_0 is used to estimate $\hat{\mathbb{E}}[\Delta^2_{\text{iid}}]$ as defined above.

As discussed in the main text, while our approximateESS shares all these steps, it produces more conservative estimates of the ESS than observed in [77]. This suggests that the curve-fitting approach underestimates s_0.

The jumpDistanceBootstrapUnsmoothedESS and the jumpDistanceBootstrapESS

We now define two new approaches to computing the ESS based on estimating s_0. In both approaches, we start with a similarity or dissimilarity measure for trees at time lag s, $g(s)$, which we then smooth into a monotonically increasing function $G(s)$. We do this by defining $G(s) = \max(g(s), g(t-1))$ for dissimilarity measures and $G(s) = \max(-g(s), -g(t-1))$ for similarity measures. In essence, regardless of $g(s)$, $G(s)$ is a distance or dissimilarity measure. We also consider a smoother version of $G(s)$, which we call $G^*(s)$, which we define below. We do not search for an asymptote of either curve directly, as [77] do for the approximate ESS. Rather, we seek the

point at which the dissimilarity of trees at time lag s is indistinguishable from the dissimilarity of a pair of trees drawn independently from the posterior distribution.

Let $\Pr(G(1) \mid \text{iid})$ be the distribution of $G(1)$ given a set of iid samples from the posterior. Given a probability α, we define a threshold ϵ to be the $(1-\alpha)$th percentile of $\Pr(G(1) \mid \text{iid})$. We estimate $\hat{\epsilon}$ using bootstrap resampling of the posterior samples, which breaks the autocorrelation but preserves the fact that the samples are from the posterior distribution. Given a choice of α and an estimate $\hat{\epsilon}$, s_0 is the first time lag s for which $G(s) > \epsilon$. If we use median... then we call the estimator in this chapter, we define $G(s)$ to be the median RF distance between trees at time lag s, and call this resulting estimate the unsmoothed jump-distance bootstrap ESS, jumpDistanceBootstrapUnsmoothedESS, though we note that any choice of $G(s)$ could rightfully be called a bootstrap ESS. In practice, we set $\alpha = 0.05$, such that s_0 is the time lag at which the tree-to-tree dissimilarity is at least as big as the 5th percentile of the tree-to-tree dissimilarity for trees drawn identically and independently from the posterior distribution.

To circumvent the fact that the jumpDistanceBootstrapUnsmoothedESS can only take on values in $n/1, n/2, n/3, \ldots, n/n$, we also consider using smoothing. Specifically, we use linear interpolation to smooth $G(s)$ into $G^*(s)$. If $\boldsymbol{s}^{\text{step}}$ is the vector of times at which $G(s)$ changes, we can define a piecewise linear function $G^*(s)$ as,

$$G^*(s) = G(s_i^{\text{step}}) + \frac{s - s_i^{\text{step}}}{s_{i+1}^{\text{step}} - s_i^{\text{step}}}(G(s_{i+1}^{\text{step}}) - G(s_i^{\text{step}})), \tag{C.13}$$

where s is in the ith interval ($s_i^{\text{step}} \leq s < s_{i+1}^{\text{step}}$). Defining s_0 to be the time s such that $G^*(s) >= \epsilon$ allows us to assign fractional s_0, and have a continuous estimator. We call the resulting estimator the (smoothed) jump distance bootstrap ESS, jumpDistanceBootstrapESS.

There are a few constraints that must be imposed to complete this approach. In the pathological case where all trees are the same tree, we set $\widehat{\text{ESS}} = 1$, as clearly if we have only sampled one topology we have an effective sample size of 1. The unsmoothed approach would not have a defined answer, as there is s for which $G(s) > \epsilon = 0$, and the smoothed approach would yield $\widehat{\text{ESS}} = n$ as $G(1) = \epsilon = 0$. It is also possible that there is no observed s for which $G^*(s) = \hat{\epsilon}$ and that $G^*(s) < \hat{\epsilon}$ for all s, in which case we enforce a minimum ESS of 1. Further, while $g(0)$ is defined, and we could infer $s_0 < 1$ we enforce a maximum ESS of n.

There is evidence that the curve-fitting approach of [77] for underestimates s_0. It is not completely clear how well these jump-distance bootstrap approaches work to estimate s_0, but it is possible that combining these with Equation C.12 will result in a better version of the approximate ESS. We leave this to future work.

C.4 Performance of the ESS measures below ESS $= 500$

As an alternative to binning ESS performance using a cutoff of 500, we also consider a laxer cutoff of 250. From Figures C.2,C.3, and C.4, it is evident a cutoff of 250 is not sufficient. In the $250 \leq$ ESS < 500 regime, it would appear most methods are generally conservative and underestimate the ESS. However, there are splits and tree topologies where the error in the estimated probability can be quite large compared to the ESS ≥ 500 regime.

C.5 Performance of all additional tree ESS measures

Across all 10 ESS methods (4 main text ESS methods and the 6 introduced in the supplement), performance is mostly similar to the main-text results. The two *ad-hoc* "jump distance" approaches that only use s_0 to estimate the ESS drastically underestimate the ESS. The splitFrequencyESS performs about as well as the frechetCorrelationESS, with slightly worse performance in the ESS < 500 regime and slightly better performance in the ESS ≥ 500 regime. The dimension-reduction approaches all perform relatively similarly. The foldedRankMedioidESS generally performs equivalently to medianPseudoESS, the CMDSESS is slightly more conservative, and the performance of the totalDistanceESS is a bit more variable. In Figures XX-YY, we plot all 10 methods for all 3 MCMCSE measures. For simplicity, and since the results are similar, we present only the RMCE. Overall, a combination of the minPseudoESS and either the frechetCorrelationESS, splitFrequencyESS, foldedRankMedioidESS, or medianPseudoESS should cover both the ESS < 500 and ESS ≥ 500 regimes in practice.

C.5.1 Scalability

Computing most of the tree ESS measures described requires computing the entire $n \times n$ distance matrix, which is computationally costly and scales with the square of the number of trees. Many of

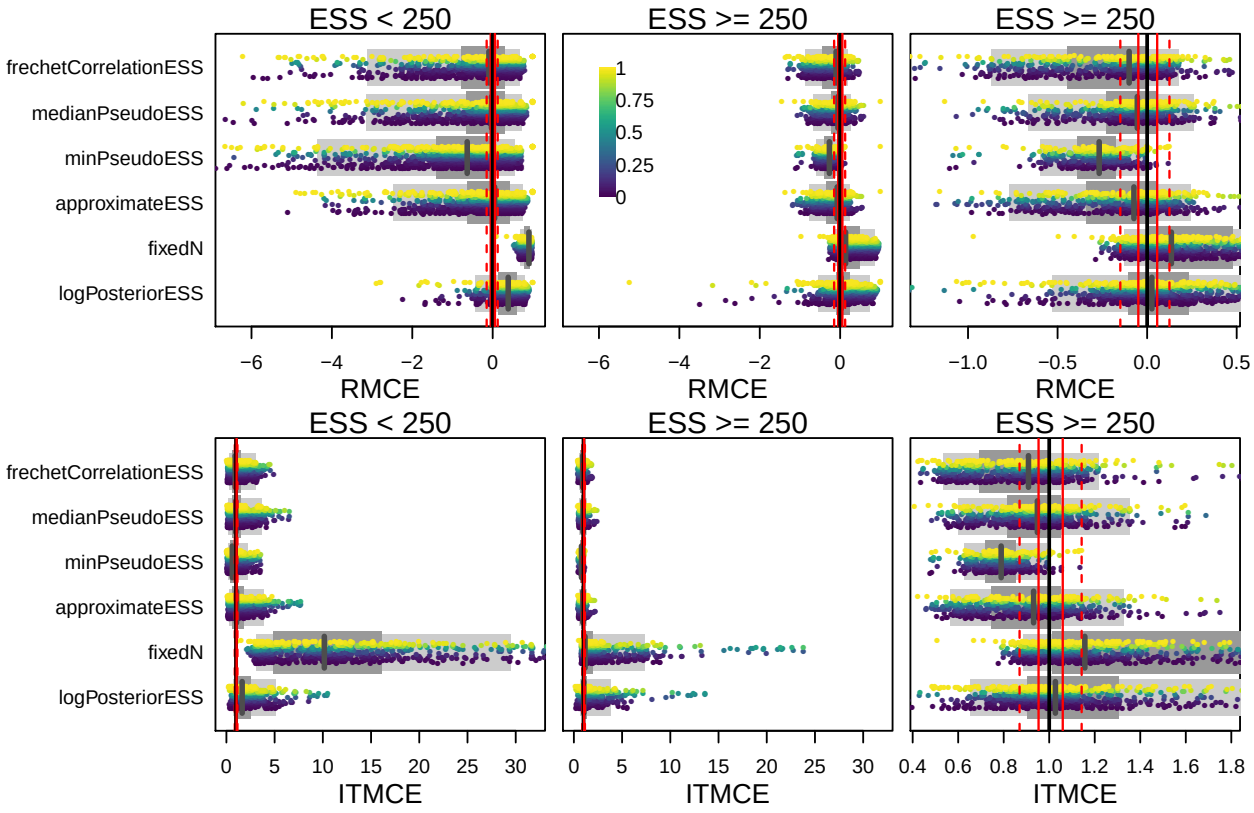


Figure C.2: The RMCE $((\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{ESS}})/\widehat{\text{SE}}_{\text{MCMC}})$ and ITMCE $(1/(1-\text{RMCE}))$ for split probabilities for all topological ESS measures and all 45 DS by run-length combinations. Splits are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). The two right panels are the same except for the scale of the x-axis. The divide between the left and right panels is based on the estimated average ESS of each of the 45 simulations, such that all splits from a simulation with average frechetCorrelationESS of 100 would show up in the left panel, while all splits from a simulation with an average Frechét correlation ESS of 600 would show up in the right two panels. As fixedN always assumes ESS = 1000, for this row we split by the number of MCMC iterations run, with the left panel including 10^3 and 10^4, and the right panel 10^5, 10^6, and 10^7. The thinner light grey bar below the points shows the 95% quantile range, the thicker dark grey bar the 50% quantile range, and the grey line is the median. Ideal performance is RMCE = 0 and ITMCE = 1 (perfect estimation of the Monte Carlo SE). As references we have plotted a solid black line for perfect performance, while the dashed (solid) red lines represent the 95% quantile range (50% quantile range) from the univariate Normal(0,1) experiment. The best performance that might reasonably be expected of a tree ESS measure would match the Normal(0,1) experiment, and thus have the grey line on the solid black line, the the thinner light grey bar align with the dashed red lines, and the thicker dark grey bar align with the solid red lines. RMCE < 0 (ITMCE < 1) implies underestimating the ESS, while RMCE > 0 (ITMCE > 1) implies overestimating the ESS, thus the log-posterior ESS and assuming ESS $= n$ tend to overestimate the ESS for splits, often substantially, while most tree ESS measures are much closer to the truth.

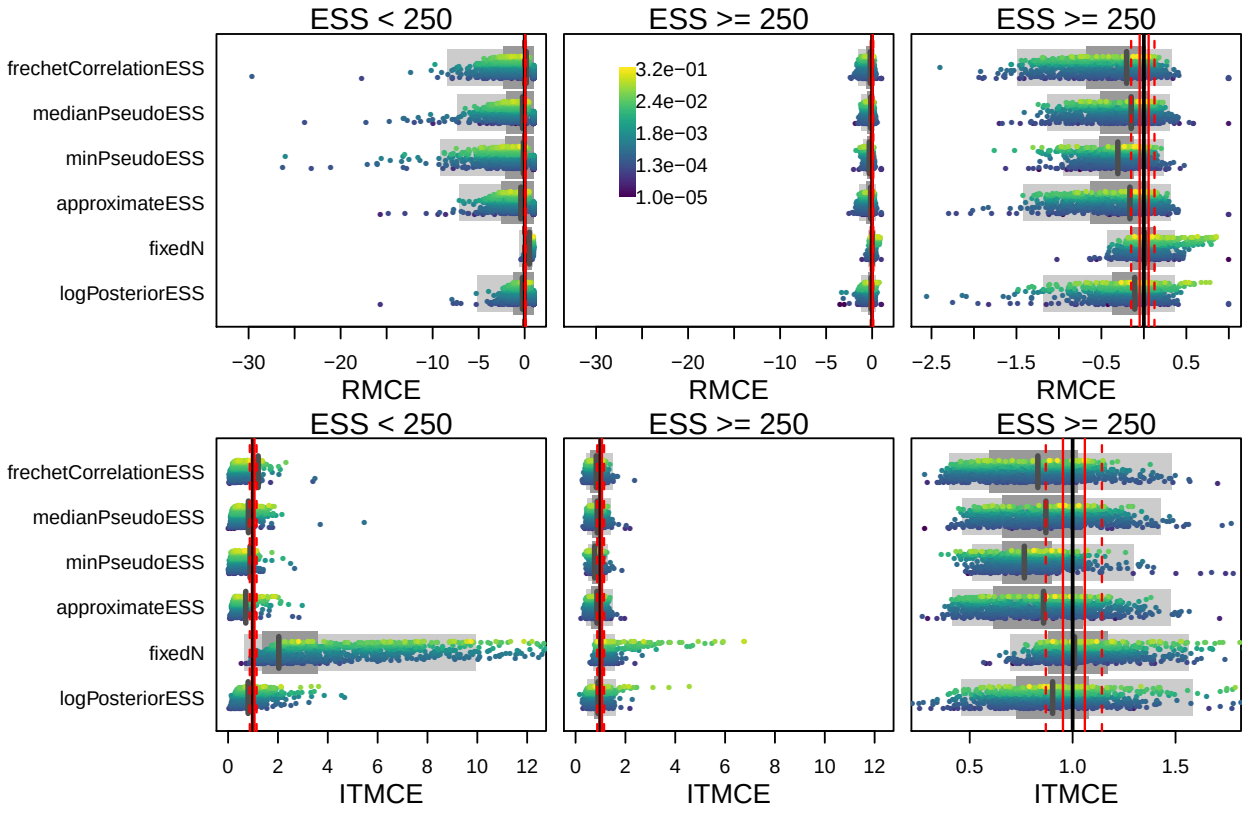


Figure C.3: The RMCE $((\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{ESS}})/\widehat{\text{SE}}_{\text{MCMC}})$ and ITMCE $(1/(1 - \text{RMCE}))$ for topology probabilities for all topological ESS measures and all 45 DS by run length combinations. Tree topologies are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). As there are too many distinct topology probabilities (nearly 100,000 across all 45 simulations), we plot only 1000 per row, preferentially keeping the highest-probability trees as these are the ones that contribute most to summary trees. For more explanation, see Figure C.2 caption.

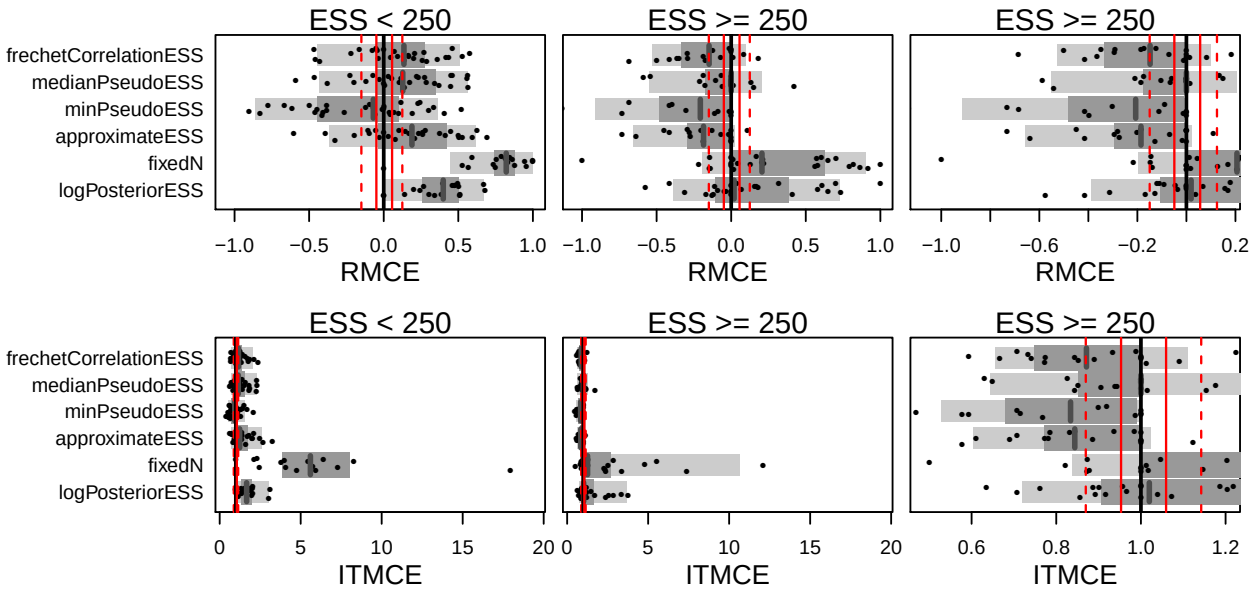


Figure C.4: The RMCE ($(\widehat{\text{SE}}_{\text{MCMC}} - \widehat{\text{SE}}_{\text{ESS}})/\widehat{\text{SE}}_{\text{MCMC}}$) and ITMCE ($1/(1 - \text{RMCE})$) for the majority-rule consensus (MRC) tree for all topological ESS measures and all 45 DS by run length combinations. The standard error for the MRC tree is a Frechét-like Monte Carlo SE, rather than a classical Euclidean Monte Carlo SE. For more explanation, see Figure C.2 caption.

the described methods can be altered to accommodate subsampling, and the RWTY implementation implements this for both the approximate ESS and medianPseudoESS [159]. Future work will be needed to determine whether any methods perform adequately with subsampling, and which methods provide an adequate runtime for either very large samples of trees or samples of very large trees.

C.6 Additional empirical results

For completion, we now present split-split plots with confidence intervals for the other 5 datasets of [127]. The confidence intervals here are computed using the Jeffreys interval [16], which appears to work well in practice, though in `treess` we implement several alternatives.

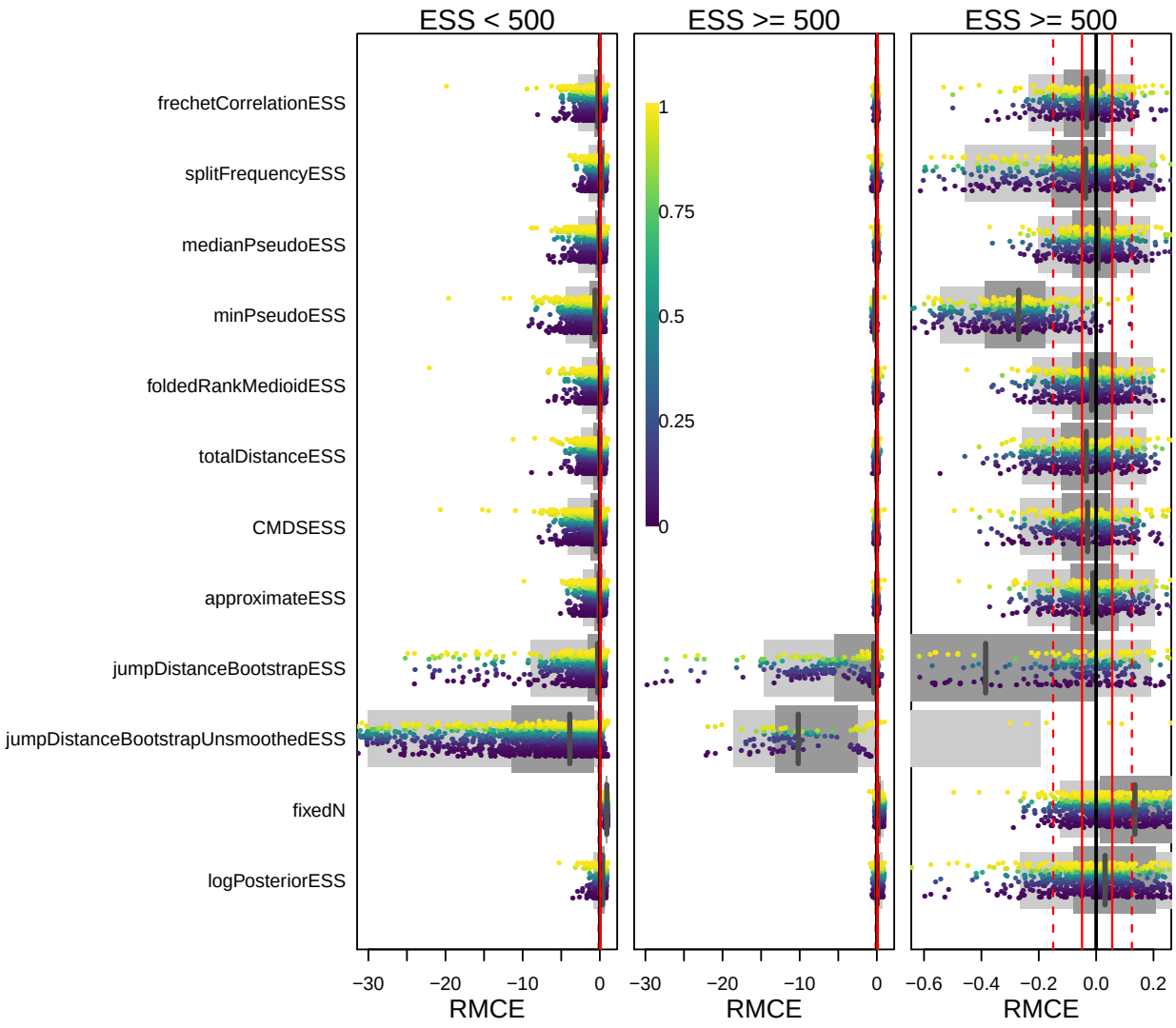


Figure C.5: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{ESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for split probabilities for all topological ESS measures and all 45 DS by run-length combinations. Splits are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). For more explanation, see Figure C.2 caption.

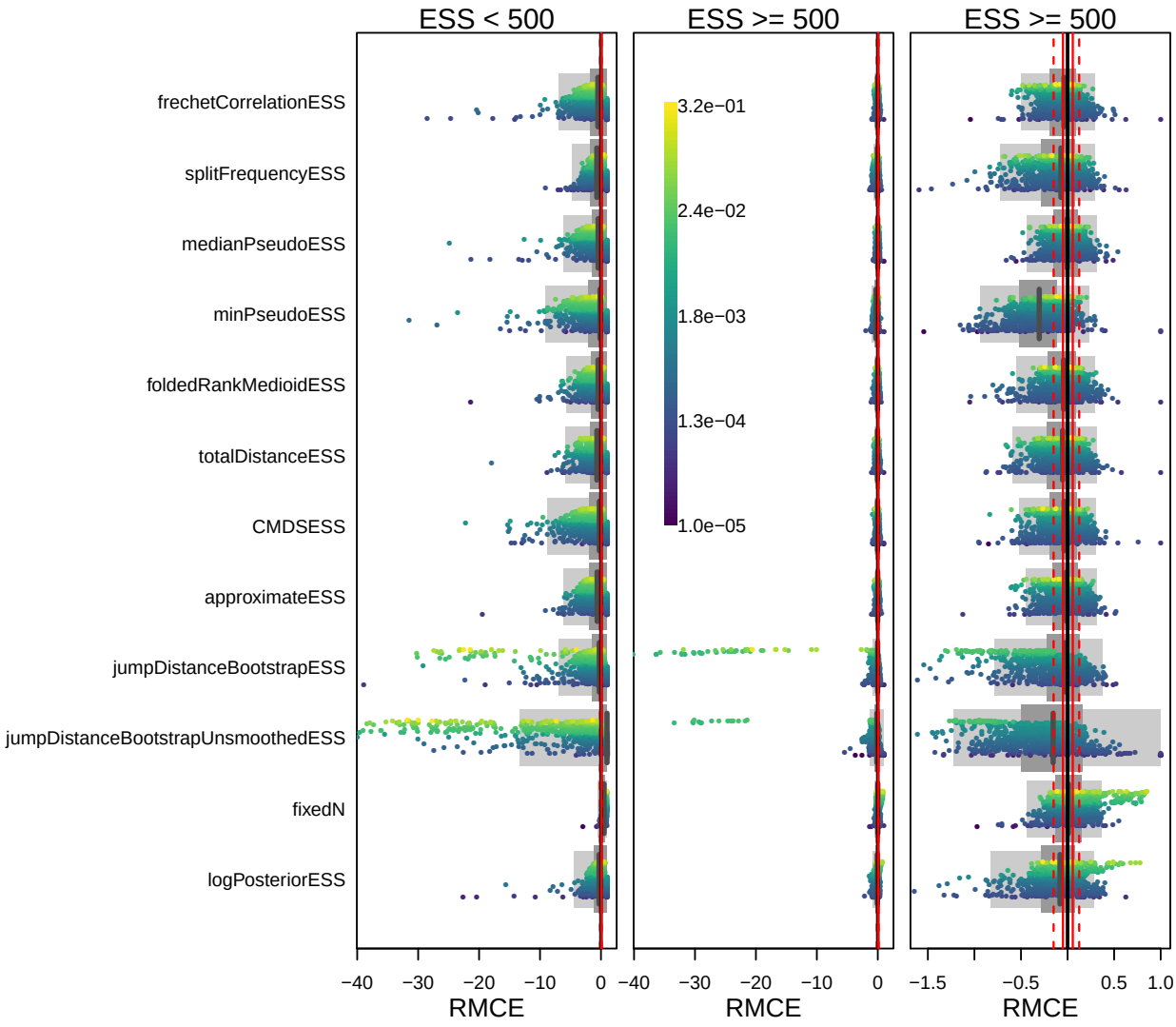


Figure C.6: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{ESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for topology probabilities for all topological ESS measures and all 45 DS by run length combinations. Tree topologies are aggregated across all 45 simulated conditions, and colored by their estimated probabilities (see scale bar in top middle panel). As there are too many distinct topology probabilities (nearly 100,000 across all 45 simulations), we plot only 1000 per row, preferentially keeping the highest-probability trees as these are the ones that contribute most to summary trees. For more explanation, see Figure C.2 caption.

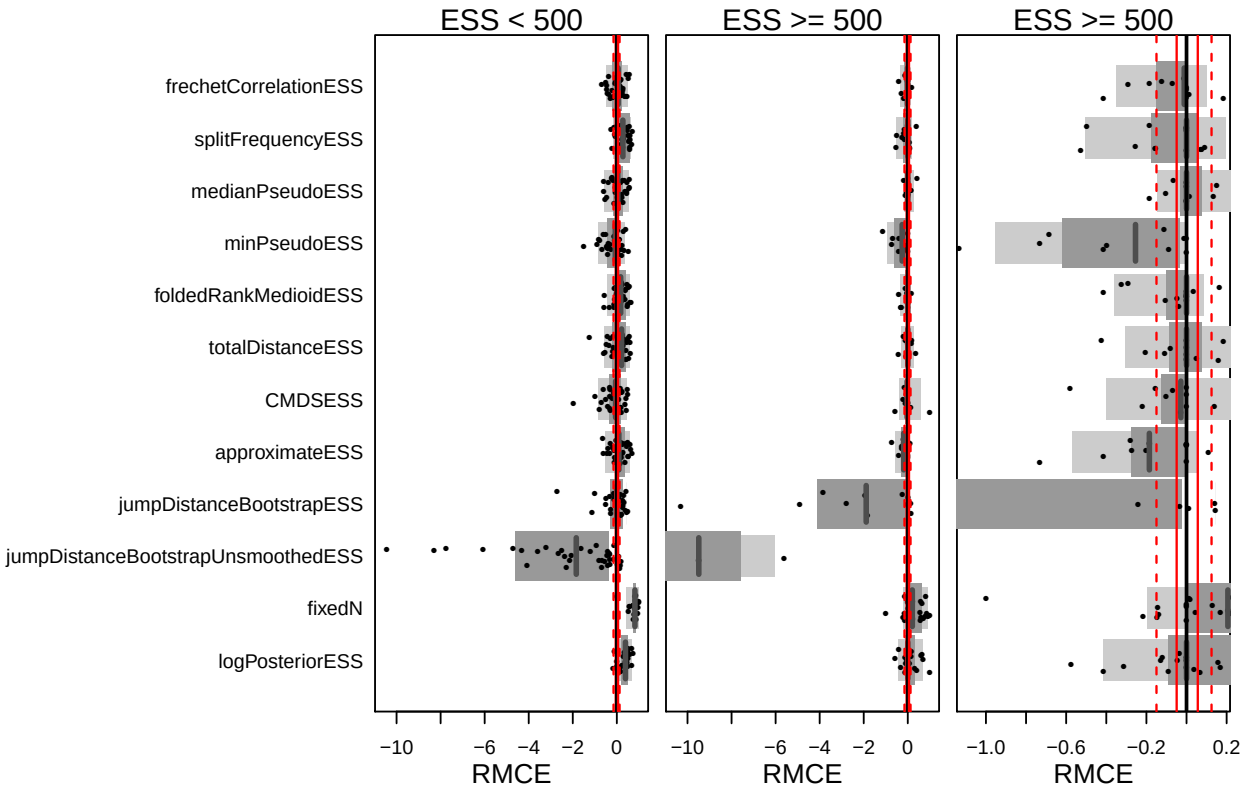


Figure C.7: The RMCE $((\widehat{\mathrm{SE}}_{\mathrm{MCMC}} - \widehat{\mathrm{SE}}_{\mathrm{ESS}})/\widehat{\mathrm{SE}}_{\mathrm{MCMC}})$ for the majority-rule consensus (MRC) tree for all topological ESS measures and all 45 DS by run length combinations. The standard error for the MRC tree is a Frechét-like Monte Carlo SE, rather than a classical Euclidean Monte Carlo SE. For more explanation, see Figure C.2 caption.

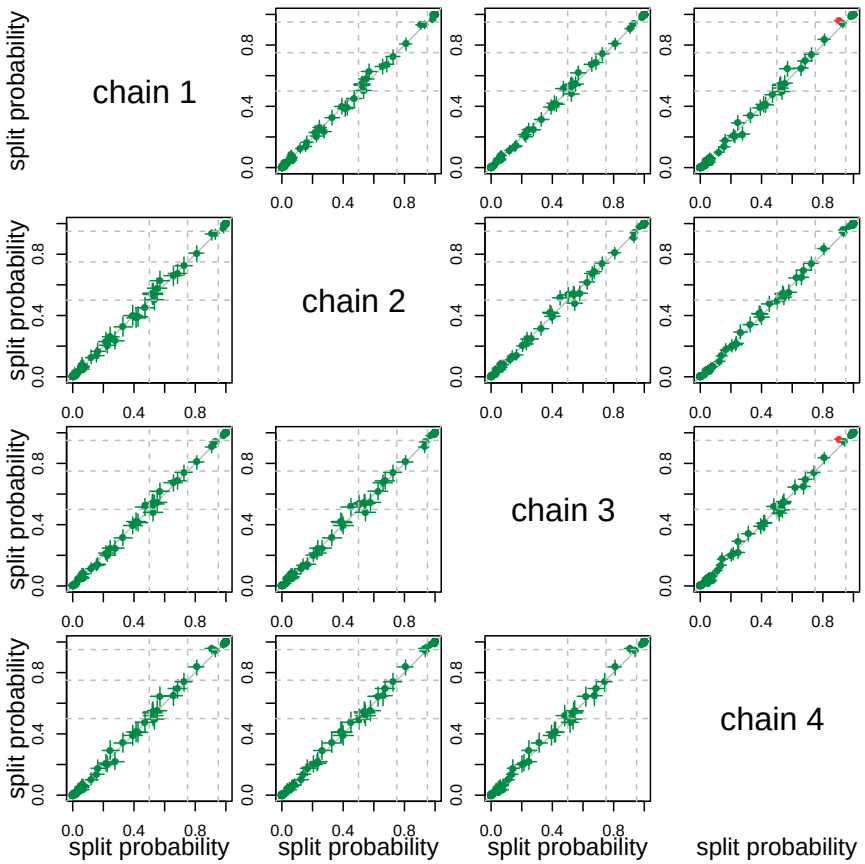


Figure C.8: Split probabilities computed for all chains of the Cophyline dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. The upper diagonal uses the frechetCorrelationESS to compute confidence intervals, while the lower diagonal uses the minPseudoESS, which is generally smaller and thus leads to larger confidence intervals. The confidence intervals are colored by whether or not the intervals for chains i and j overlap (green for overlap, red for no overlap). Non-overlapping confidence intervals are more likely to be indicative of convergence issues between chains, such that longer runs may still result in different estimated split probabilities. Overlapping confidence intervals suggest that longer runs will likely lead to identical split probabilities. Narrower confidence intervals from larger tree ESS estimates will flag more splits as problematic (as in chains 1 and 4). Dashed grey lines indicate posterior probabilities of 0.5 (threshold for inclusion in the MRC tree), 0.75 (moderate support for a split), and 0.95 (strong support for a split).

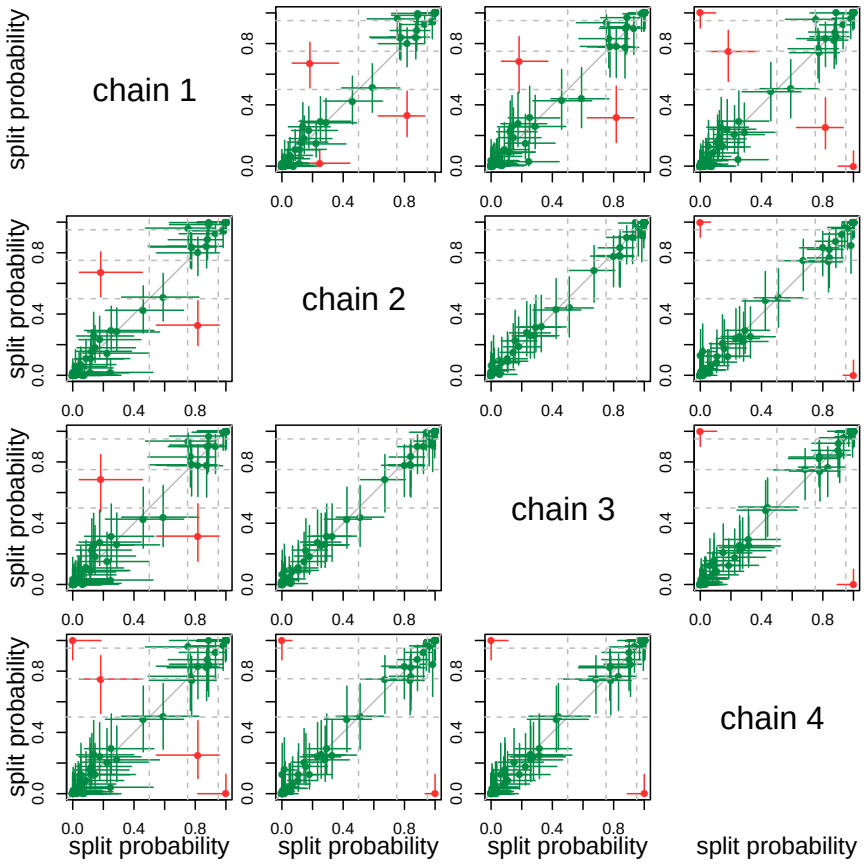


Figure C.9: Split probabilities computed for all chains of the Gephyromantis dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure C.8 caption.

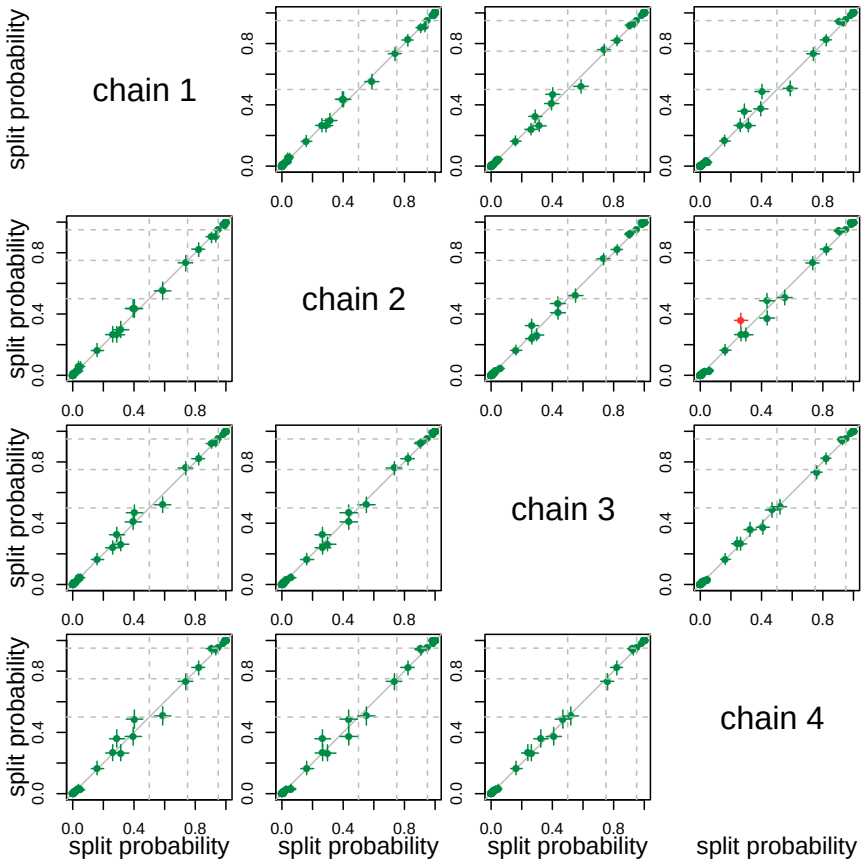


Figure C.10: Split probabilities computed for all chains of the Heterixalus dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure C.8 caption.

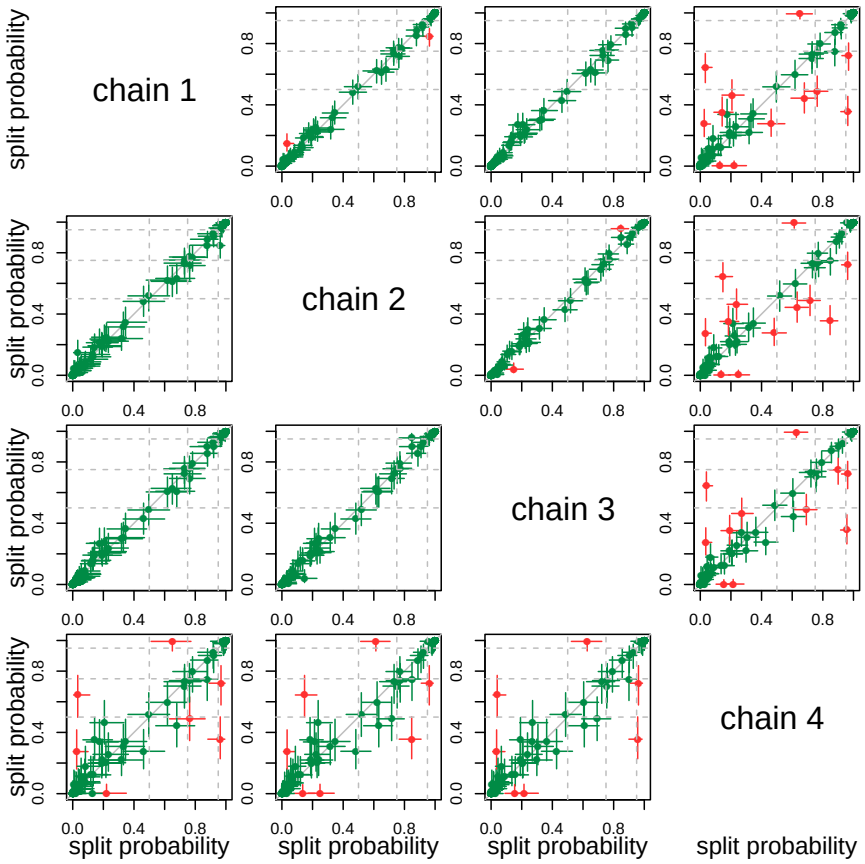


Figure C.11: Split probabilities computed for all chains of the Phelsuma dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure C.8 caption.

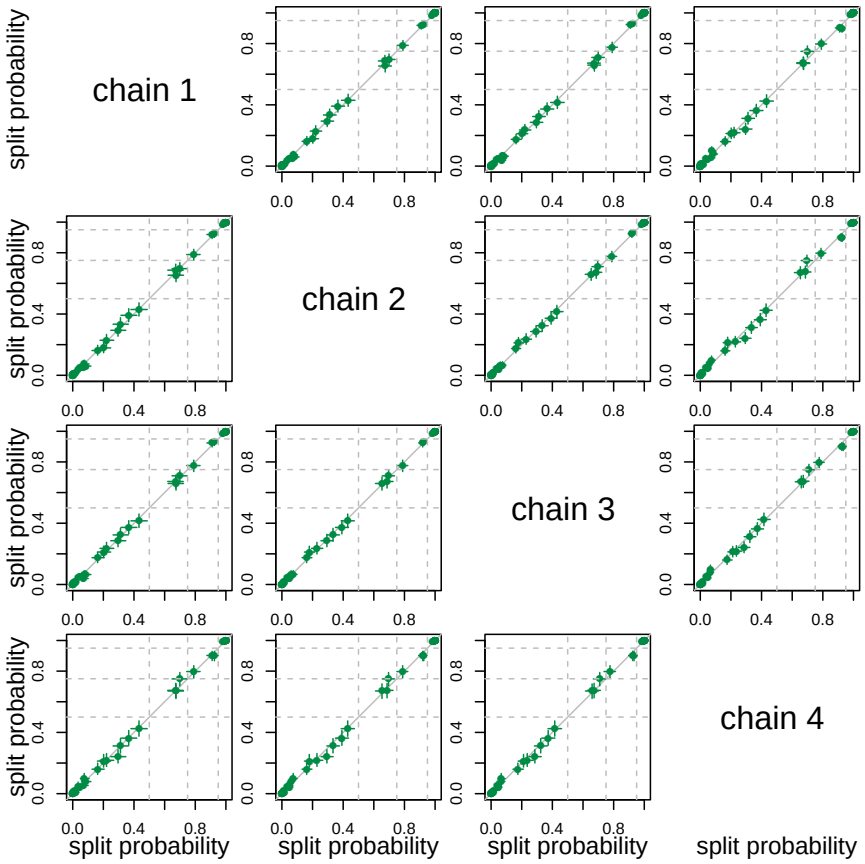


Figure C.12: Split probabilities computed for all chains of the Uroplatus dataset of [127], plotted against the probabilities computed for all other chains, with confidence intervals. For more explanation, see Figure C.8 caption.