On the robustness of Bayesian phylogenetic gene tree estimation

Luiza Guimarães Fabreti

München 2022

On the robustness of Bayesian phylogenetic gene tree estimation

Luiza Guimarães Fabreti

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Geowissenschaften
der Ludwig–Maximilians–Universität
München

vorgelegt von
Luiza Guimarães Fabreti
aus Brasilien

München, den 07. September 2022

Erstgutachter: Dr. Sebastian Höhna

Zweitgutachter: Prof. Dr. Jeremy M. Brown

Tag der Fertigstellung: 07.09.2022
Tag der mündlichen Prüfung: 26.04.2023

Contents

Acknowledgements

I would like to start by thanking my supervisor Dr. Sebastian Höhna for trusting me with the opportunity of this scientific work. His support and guidance were fundamental for the success of this project. Thank you for allowing me to learn so much from you, for always being ready and sharp for a scientific discussion and for all the incentive during these years. I also thank the German Research Foundation (*DFG - Deutsche Forschungsgemeinschaft*) through the Emmy Noether Program for funding this project.

My greatest gratitude to the Höhna Lab members and to Sebastian Höhna for putting together such a great team. I thank Dr. Allison Hsiang and Ronja Billenstein for their endless support, guidance and terrific friendship. I also thank Killian Smith for all his help with computer technologies and Bjørn Kopperud for his always accurate observations. I also thank Dr. Alessio Capobianco, Dr. John Clarke and Wenjie Zhu for their patience and willingness to provide feedback. I am very grateful for all the discussions we could put together and the precious feedback we could give one another.

My gratitude to the `RevBayes` development team, in special to my collaborators and people I could get to know through these years. Dr. Jeremy Brown, Dr. Tracy Heath, Dr. Michael May, Dr. Lyndon Coghill, Dr. Andrew Magee, Dr. John Huelsenbeck, Dr. Isabel Sanmartin, Dr. Rui Borges and Dr. Rachel Warnock thank you very much for inspiring me and showing me the way to great science.

I would like to thank Prof. Dr. Gert Wörheide, Prof. Dr. William Orsi, Dr. Dirk Erpenbeck, Dr. Sergio Vargas and Dr. Oliver Voigt for their contribution in the further development of my scientific path. All past and present members of the Department of Earth & Environmental Sciences, Palaeontology & Geobiology impacted my years in the department and I will always be grateful for getting to know so many magnificent people. Among these people, I thank my colleagues and friends Luis Porras, Ramon Rivera, Nicola Conci, Joëlle van der Sprong, Dr. Michael Eitel, Catalina Garcia, Ksenia Juravel, Dr. Charalampos Kevrekidis, Dr. Ömer Coskun and Dr. Carolin Gut for the inspiring conversations and good moments. I especially thank Dr. Sergio Vargas for the friendship and all the help he gave me when I first arrived in Munich. I knew the big challenge ahead of me when I came here with a child and without your help it would have been a lot harder.

Many people from the Department of Earth & Environmental Sciences, Palaeontology & Geobiology helped in many ways with the accomplishment of this project. Among them I thank René Neumaier and Dominik Bastl for their support with the IT challenges. I also owe many thanks to Monika Brinkrolf, Britta Nüske-Illenberger and Nicole Vieweger for always being so helpful and all their help with German bureaucracy.

Last, but not least, I would like to thank my family in my native language. Aos meus pais, Marian e Luiz Alberto, eu agradeço imensamente pelo suporte e todo investimento em minha educação. Vocês nunca mediram esforços para que eu tivesse acesso aos melhores estudos e recursos e o resultado de todo esse esforço encontra-se aqui nesse trabalho. À minha irmã Tatiana, minha eterna gratidão por ser minha melhor amiga, meu porto-seguro e a pessoa com quem eu sempre posso contar. Ao meu noivo, Norbert, minha gratidão por todo apoio nos últimos anos e principalmente por exercer um excelente papel de pai. Sem a sua ajuda, eu não conseguiria ter sucedido. Ao meu filho, Victor, muito obrigada por ter embarcado nessa aventura comigo e por ser tão paciente e compreensivo. Esse trabalho é dedicado à vocês: Victor, Norbert, Marian, Luiz Alberto, Tatiana, Leonardo, Melina e Clarissa.

List of Figures

List of Tables

List of Published Papers, Preprints, Manuscripts Submitted and in Preparation

Luiza Guimarães Fabreti, Sebastian Höhna (2021). Convergence Assessment for Bayesian Phylogenetic Analysis using MCMC simulation.
Published in: *Methods in Ecology and Evolution*, Volume 13, Issue 1, January 2022, Pages 77-90, https://doi.org/10.1111/2041-210X.13727 (**Chapter 2**)

Luiza Guimarães Fabreti, Lyndon M. Coghill, Robert C. Thomson, Sebastian Höhna, Jeremy M. Brown (2022). The Expected Behavior of Posterior Predictive Tests and its Unexpected Interpretation.
In Review in: *Molecular Biology and Evolution* (**Chapter 3**)

Luiza Guimarães Fabreti, Sebastian Höhna (2022). Nucleotide Substitution Model Selection is not Necessary for Bayesian Inference of Phylogeny with Well Behaved Priors.
In Review in: *Systematic Biology*
Preprint available at: *BioRxiv*, https://doi.org/10.1101/2022.02.17.480861 (**Chapter 4**)

Luiza Guimarães Fabreti, Sebastian Höhna (2022). Evaluating Gene Tree Discordance on Mammalian Orthologous Markers.
This manuscript is in preparation for standalone publication (**Chapter 5**)

Joëlle Barido-Sottani, Joshua A. Justison, Rui Borges, Jeremy M. Brown, Wade Dismukes, Bruno do Rosario Petrucci, **Luiza Guimarães Fabreti**, Sebastian Höhna, Michael J. Landis, Paul O. Lewis, Michael R. May, Fábio K. Mendes, Walker Pett, Benjamin D.Redelings, Carrie M. Tribble, April M.Wright, Rosana Zenil-Ferguson, and Tracy A.Heath (2022). Lessons learned from organizing and teaching virtual phylogenetics workshops.
Published in: *Bulletin of the Society of Systematic Biologists*, Volume 1, No. 2, June 2022, https://doi.org/10.18061/bssb.v1i2.8425 (**Appendix C**)

Authors Contributions

Chapter 2: L.G.F. implemented the R package Convenience and performed the simulation study; S.H. designed the project and algorithms. Both authors wrote the manuscript and approved the final version.

Chapter 3: All authors designed the research. L.G.F. conducted simulations. L.G.F., L.M.C., and J.M.B. conducted analyses. L.G.F., S.H., and J.M.B. wrote the draft manuscript. All authors revised the manuscript.

Chapter 4: L.G.F. and S.H. designed the reasearch. L.G.F. performed the analyses. Both authors wrote the manuscript and approved the final version.

Chapter 5: L.G.F. and S.H. designed the reasearch. L.G.F. performed the analyses. L.G.F wrote the manuscript and both authors approved the final version.

Appendix C: All authors contributed to the teaching of the workshops. J.B.S., J.A.J and T.A.H. wrote the manuscript. All authors approved the final version.

Summary

Bayesian phylogenetic inference uses a model of sequence evolution and a multiple sequence alignment data to estimate posterior probabilities of phylogenetic trees and other parameters. Such estimates are acquired through Markov chain Monte Carlo (MCMC) algorithms. Some additional steps are necessary to infer a robust phylogenetic tree such as convergence assessment of the MCMC, choosing a model of sequence evolution, choosing prior probabilities for the model parameters and assessment of model adequacy between the model of sequence evolution and the data. These steps are often overlooked in the inference, mainly due to the lack of straightforward methods and thorough investigation of their impact.

In this dissertation I investigated these often disregarded steps and proposed new methods to ensure a robust estimation of Bayesian phylogenetic gene trees. In **Chapter 1** some relevant background information on Bayesian inference of phylogeny is provided. **Chapter 2** aimed at the development and implementation of a new method for MCMC convergence assessment in phylogenetics. We proposed a novel method that evaluates continuous and discrete parameters separately, additionally we proposed reliable thresholds for each convergence criterion. Furthermore, we implemented the method in the easy-to-use R package `Convenience`.

Chapter 3 was dedicated to assessing the nature of posterior predictive *p*-values in Bayesian phylogenetic model adequacy tests. We unraveled that *p*-values distributions are rather conservative for phylogenetic tree inference. This finding emphasizes that poor model fit in phylogenetics should be taken seriously.

In **Chapter 4** we assessed the effects of model over-parameterization and prior probability distribution choice for the commonly used generalised time reversible (GTR) family of nested models. We observed that substitution model over-parameterization is not a problem for phylogenetic inference, when a proper set of prior distributions are chosen. Consequently, substitution model selection of common models of nucleotide substitutions becomes an unnecessary step in Bayesian phylogenetic inference.

Finally, in **Chapter 5** we tested the robustness of gene tree estimation with the newly proposed methods of the previous chapters. We estimated the gene trees for a subset of multiple sequence alignment from the OrthoMam database. We observed that the lack of proper convergence assessment impacts the inferred parameters and can lead to erroneous conclusions. Furthermore, we concluded that a considerable amount of gene tree discordance is due to estimation errors and the most complex substitution model is still inadequate to the real data. Therefore, future research should focus on models of sequence evolution that better capture the heterogeneity of real data.

This dissertation contributes to the development and improvement of Bayesian methods to estimate phylogenetic trees. Robust methods to estimate phylogenetic trees are fundamental for complex evolutionary questions such as estimating the tree of life or estimating time in evolutionary trees. It includes a novel method for MCMC convergence assessment with thorough statistical foundation and fully automation. Furthermore, we characterized the distribution of posterior predictive *p*-values for model adequacy tests in Bayesian phylogenetics. Additionally, we demonstrated that substitution model selection is not a necessary step by testing the behavior of phylogenetic estimates under over-parameterized models. Finally, we demonstrated the application of these outcomes on real data and suggested future directions for the field of Bayesian phylogenetics.

Chapter 1

Introduction

'Nothing in biology makes sense except in the light of evolution'

Theodosius Dobzhansky, 1973

'Nothing in evolution makes sense except in the light of phylogeny'

Society of Systematic Biologists

The biodiversity observed on Earth is a result of many years of complex interactions among organisms and the environment. The process which includes adaptation, selection and inheritance receives the name of evolution and the most important work describing its mechanisms is the book *On the origin of species by means of natural selection* [29]. The main concept in evolution is that all forms of life are related through common ancestors and the landscape of biodiversity changes through time with some forms of life disappearing and others emerging. The relationship among these forms of life can be uncovered by looking into their genetic information. The genetic contents that share a common ancestry are called homologous and these are primary to reconstruct evolutionary histories. The reconstructed evolutionary history is a phylogeny, a tree-like configuration as seen in Figure 1.1 where nodes represent the most common ancestor from lineages and the tips represent the forms of life that we can observe (commonly referred as taxa). The root represents the common ancestor among all displayed taxa, and evolutionary time goes from the root to the tips. Reconstructing a phylogenetic tree is not a trivial problem since the number of possible trees increases incredibly fast as the number of taxa increases. But this challenge can be tackled with probabilistic models and modern algorithms [42].

Bayesian phylogenetic inference has become a popular method in molecular phylogenetics since it was implemented in the late 90's [181, 123, 103, 113, 82]. This popularity can be attributed to the robustness of Bayesian data analysis and the availability of user-friendly software [132]. Over the years, the models for Bayesian phylogenetic inference have grown in complexity. Such models are employed in biodiversity studies [25, 176, 161], epidemiology [96, 117, 129], phylogeography [110], estimation of diversification rates [84], divergence time estimation [136, 10]. The underlying methods that enable these complex models to be computationally feasible are the Markov chain Monte Carlo (MCMC) algorithms. In this thesis, I explored and proposed novel improvements for several challenges of the Bayesian method associated with estimating phylogenetic gene trees. These challenges include convergence assessment for MCMC, selection of a substitution model, the choice of the prior probability distribution for the model parameters and the behavior of posterior predictive tests for model adequacy studies. Additionally, we investigated the proposed improvements on empirical data sets. In the following sections we will present the Bayesian phylogenetic inference method in more detail.

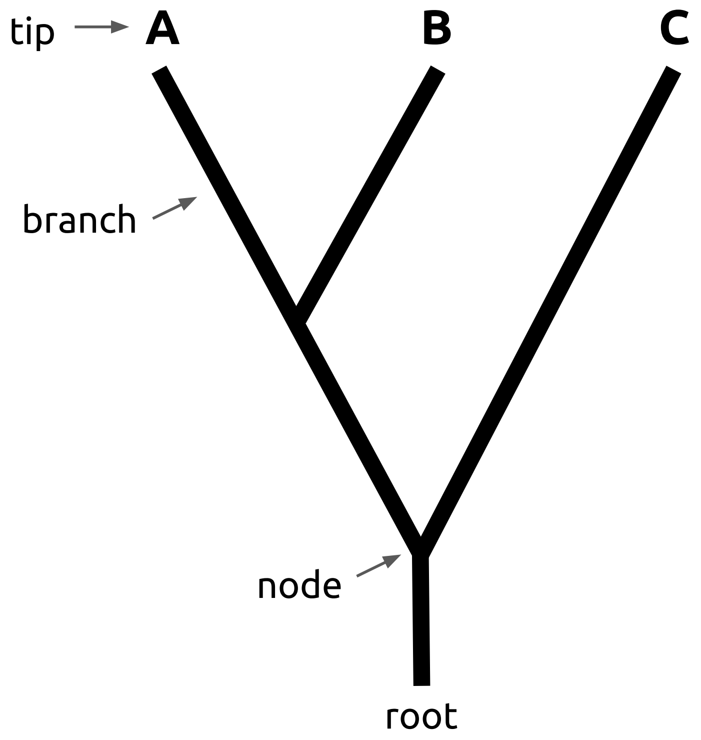


Figure 1.1: A schematic of a phylogenetic tree with tips A, B, C, two internal nodes, and branches connecting the nodes. Time goes from the root in direction to the tips. The internal node represents the most recent common ancestor among the lineages.

1.1 Phylogenetic gene trees

Phylogenetic gene trees represent hypotheses about the evolutionary relationship among genes. Such relationships reflect the evolutionary history that led to the emergence of the given genes. Gene trees and species trees do not necessarily share the same evolutionary history. The process of evolution of genes commonly involves gene loss and gene duplication, incomplete lineage sorting, horizontal gene transfer, besides different genes usually evolve under different rates and moreover genes that are in close spatial regions can undergo recombination [157]. All this together contributes to the incongruence between gene and species trees. Such incongruence imposes an obstacle to reconstruct the evolutionary relationship among species. To overcome this problem and infer species trees, researchers have adopted two possible strategies: 1) assuming that all genes follow the same evolutionary history (super-matrix approach) 2) allow genes to evolve with different histories, in a process known as the multispecies coalescent. In the super-matrix approach, the alignments for different genes are concatenated generating a big matrix of multiple genes. The big matrix is then used to perform the phylogenetic inference. The problem with this approach is that it assumes all genes follow the same evolutionary history. The multispecies coalescent process, on the contrary, accounts for different evolutionary histories among different genes. The multispecies coalescent incorporates ancestral polymorphisms in the process of species tree inference. The most commonly used MSC method, ASTRAL, is a summary method, which means that the gene trees are first estimated and then summarized using quartets to estimate the species tree with branch lengths being the coalescent times. Hence, the multispecies coalescent process requires robust gene trees to infer the species tree.

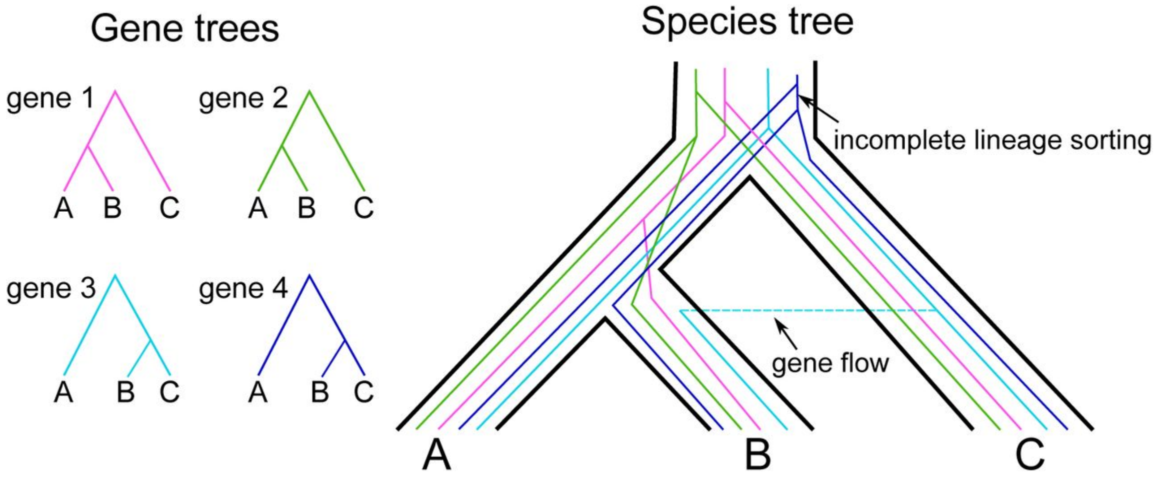


Figure 1.2: The example of incongruent gene trees and the species tree. On the left panel, four gene trees are displayed, with gene 1 and gene 2 showing the same evolutionary history; and gene 3 and gene 4 with same evolutionary history. The right panel shows the species tree with the different gene trees inside it. On the species tree we can observe incomplete lineage sorting and an event of gene flow. Extracted from [119]

Bayesian inference provides a robust method for inferring phylogenetic gene trees. First, because of the requirement of an explicit model of sequence evolution, *i.e.,* the evolutionary assumptions are defined and specified. Second, due to the nature of the method, uncertainties are intrinsically estimated. The results from a Bayesian phylogenetic inference are not simply point estimates, but rather distributions. In that sense, a Bayesian phylogenetic tree is estimated together with its posterior probability. The posterior probability of a phylogenetic tree is the probability that the tree is correct, supposing the model is correct.

Within the different methods to estimate phylogenies, this property is only intrinsic to the Bayesian method [80].

1.2 Bayesian Inference

Bayesian inference makes use of Bayes' theorem [12, 102] to estimate posterior probabilities of model parameters based on the data and prior information. The posterior probabilities are updated with the joint probability of prior and data according to Bayes' theorem as follows:

$$P(H \mid D) = \frac{P(H) \times P(D \mid H)}{P(D)} \tag{1.1}$$

where H is the hypothesis; D is the data; $P(H \mid D)$ is the posterior probability of the hypothesis given the data; $P(H)$ is the prior probability distribution of the hypothesis; $P(D \mid H)$ is the probability of observing the data given the hypothesis, *i.e.,* the likelihood; $P(D)$ is the marginal likelihood of the data.

In phylogenetic inference the parameter of interest to be estimated is frequently the phylogenetic tree. The data are typically DNA nucleotide sequence alignments for the taxa of interest. The description of the process of sequence evolution that gives rise to the phylogenetic tree is comprised in the substitution model. The parameters from the substitution model and the phylogenetic tree are jointly estimated using Bayes' theorem.

The prior probability distribution (or just prior) comprises the former knowledge about the parameters to be estimated. The priors can be classified according to the researcher's degree of certainty about the parameter. In this classification the three categories for priors are: (a) an informative prior, when there is a high degree of certainty about the model parameters; (b) a weakly informative prior, when some information is known; and (c) a diffuse prior, when there is a lot of uncertainty about the parameters [169].

For many practical applications of Bayesian inference, including phylogenetic inference, the marginal likelihood becomes a multi-dimensional integral. This means that calculating the integral is virtually impossible. Other methods are used in combination with Bayes's theorem to overcome this problem. Such a method is the Markov chain Monte Carlo (MCMC) sampling method. We will discuss the MCMC in more detail in the next section.

1.3 MCMC

The Markov chain Monte Carlo (MCMC) algorithms is a method to simulate stochastic processes with an underlying stationary distribution. Such methods are used as a numerical approximation for problems in statistical inference with multi-dimensional integrals [56]. A sequence $X_1, X_2, X_3, ...$ is a Markov chain if the probability of the state X_{n+1} depends only on the state X_n. The most common MCMC method is the Metropolis-Hastings algorithm [148]. The algorithm follows these steps:

1. Generate an initial value for the model parameter (x).

2. Propose a new value $x^{'}$.

3. Calculate the acceptance ratio R:

$$R = \min{(1, \frac{f(x^{'}) \times Q(x \mid x^{'})}{f(x) \times Q(x^{'} \mid x)})}, \tag{1.2}$$

   where $f(x)$ is the target distribution and Q is the transition kernel.

4. Draw a number u from an Uniform distribution between 0 and 1.

5. If $u \leq R$, accept the new value and set $x = x^{'}$. Otherwise keep the value of x.

6. Go back to step 2 and repeat for many iterations.

The algorithm above will result in a collection of values for each parameter, this collection is called sample. A very important step after the MCMC is checking for convergence. Convergence aims at checking that sufficiently many samples were taken and the sample is representative of the posterior distribution [66]. Checking for convergence in Bayesian phylogenetics is particularly challenging. That is due to the discrete nature of phylogenetic trees, which are frequently the parameter of interest that researchers wish to estimate.

The common practice in phylogenetics is to visually inspect the trace plots of the continuous parameters [134, 144, 173]. Trace plots are plots of sampled states per iteration. They are a good tool for checking MCMC mixing status, *i.e.,* the level of autocorrelation between consecutive samples. Figure 1.3 shows an example of a trace plot for the tree length of a single chain MCMC. The Figure shows an example of good mixing, which can be evaluated by the lack of plateaus and directional change. Figure 1.4, in contrast, shows the example of a bad mixing MCMC chain.

1.4 Substitution model

Substitution models are the basis for phylogenetic trees reconstruction. The substitution model describes how changes in nucleotides (for DNA tree-based reconstruction) happen over evolutionary time. The most commonly used substitution models are the family of Generalized Time Reversible (GTR) [165] nested models. These models assume that:

a) neutrality, selection does not play a role in the substitutions;

b) the sites on the alignment evolve independently, *i.e.,* the changes in one site do not affect the probabilities of changes in other sites;

c) the number of sites is finite, over the course of evolution a single site can undergo multiple changes;

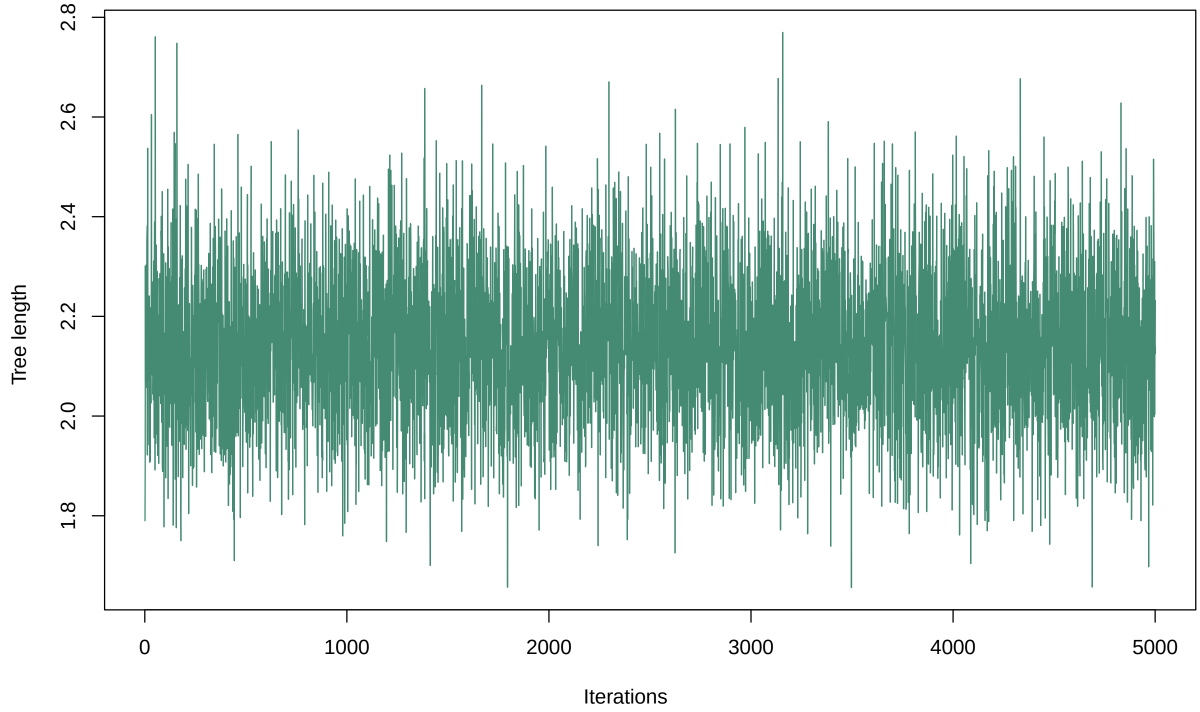


Figure 1.3: An MCMC trace plot with good mixing of the tree length for a single chain. The MCMC iterations are on the x-axis. The tree length values that were sampled during the MCMC are on the y-axis.

d) time-reversibility, *i.e.,* the direction in which evolution occurs does not matter.

These models are also designed to be stationary, meaning that the expected value does not change over time.

The rate matrix Q is a matrix that governs the rates at which the possible states change from one to another. In the case of DNA sequence evolution, there are four possible states which are the nucleotides A, C, G and T. The Q matrix takes the following form:

$$Q = \begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix} \tag{1.3}$$

$$Q_{GTR} = \begin{pmatrix} \bullet & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & \bullet & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & \bullet & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & \bullet \end{pmatrix} \tag{1.4}$$

with μ_{ij} being the instantaneous rate of change from i to j and μ_i the rate of not changing out of state i. The transition rate matrix is a matrix of probabilities of transition

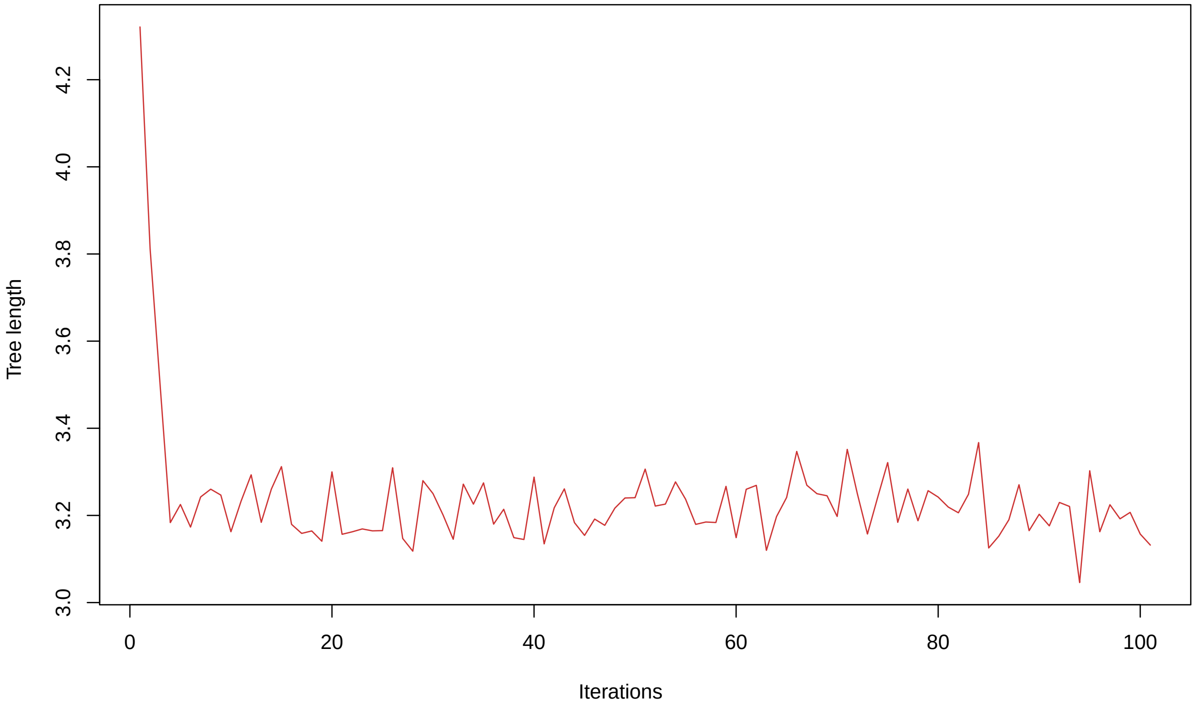


Figure 1.4: An MCMC trace plot with bad mixing of the tree length for a single chain. The MCMC iterations are on the x-axis. The tree length values that were sampled during the MCMC are on the y-axis.

between states for a given time t, it can be computed by exponentiating the rate matrix as follows:

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{TT}(t) \end{pmatrix} = e^{Qt} = \sum_{n=0}^{\infty} Q^n \frac{t^n}{n!} \tag{1.5}$$

The time unit is often measured in expected number of changes per site. This measure is represented in the tree as branches lengths. The branches display evolutionary distances among sequences. They are calculated as the product of the mean rate of substitutions and the time.

Another property of these substitution models is ergodicity, *i.e.,* it is always possible to move from state i to state j (even if it takes more steps). This property results in an equilibrium in which the frequencies of states (π_A, π_C, π_G and π_T) do not change.

The most simple substitution model within the family of Generalized Time Reversible (GTR) nested models is the Jukes-Cantor (JC) model [86]. The JC model assumes that the base frequencies (π_A, π_C, π_G and π_T) are equal and the transition rates among the nucleotides is also equal. The transition matrix for the JC model takes the form:

$$P(t) = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix} \tag{1.6}$$

with μ being the rate of substitution. The relationship between μ and branch length (v) is given by $v = \frac{3}{4}t\mu$. The JC model is the simplest among the GTR family because it has no free parameters to be estimated. On the contrary, the GTR model [165] is the most complex within its family, because all parameters have to be estimated from the data. As a consequence of the complexity of the model, the GTR does not have an algebraic form for the transition probability matrix, therefore, its transition matrix must be computed numerically.

Other models with increased complexity were proposed over the years following the JC model. For the purposes of this thesis, I will focus on both the JC and GTR models. These two models represent the most simplistic and most complex assumptions within a family of nested models.

The assumptions of the models are usually a simplification of the complex process that generates the observed data. For example, the assumption that all sites in an alignment evolve under the same rates is contradicted by the information that researchers have about the different codon positions. Scientists have observed that for protein coding genes the third codon position mutates faster than the first position, which mutates faster than the second position [13]. In order to account for that rate variability, [177] proposed the among site rate variation (ASRV) model. The ASRV model implementation is done by a discrete gamma model with four categories. The underlying gamma distribution has mean one, as a result the shape parameter (α) and the rate parameter (β) have equal values. The amount of rate variation increases with lower values of α [179]. Another model that incorporates observed features of the data is the invariant sites model [2, 62]. This model enables a proportion of the sites to be invariant. In that way, a group of sites will be considered variable, and another group will be considered fixed.

1.5 Aims of the Study

This study aimed at addressing methodological challenges of the Bayesian phylogenetic method. The first Chapter was dedicated to developing a robust and statically grounded method to assess convergence of Bayesian phylogenetic inference using MCMC. The second Chapter focused on characterizing the behavior of model adequacy tests using posterior predictive simulations. Next, we investigated the necessity of performing model selection in Bayesian phylogenetic inference. The last Chapter combined all previous results into estimating phylogenetic gene trees for an empirical dataset. Overall, our aim was to improve the robustness of the Bayesian phylogenetic method and guide the advancements in future work.

Chapter 2

Convergence Assessment for Bayesian Phylogenetic Analysis using MCMC simulation

2.1 Abstract

1. Posterior distributions are commonly approximated by samples produced from a Markov chain Monte Carlo (MCMC) simulation. Every MCMC simulation has to be checked for convergence, i.e., that sufficiently many samples have been obtained and that these samples indeed represent the true posterior distribution.

2. Here we develop and test different approaches for convergence assessment in phylogenetics. We analytically derive a threshold for a minimum effective sample size (ESS) of 625. We observe that only the initial sequence estimator provides robust ESS estimates for common types of MCMC simulations (autocorrelated samples, adaptive MCMC, Metropolis-Coupled MCMC). We show that standard ESS computation can be applied to phylogenetic trees if the tree samples are converted into traces of absence/presence of splits.

3. Convergence in distribution between replicated MCMC runs can be assessed with the Kolmogorov-Smirnov test. The commonly used potential scale reduction factor (PSRF) is biased when applied to skewed posterior distribution. Additionally, we provide how the distribution of differences in split frequencies can be computed exactly akin to standard exact tests and show that it depends on the true frequency of a split. Hence, the average standard deviation of split frequencies is too simplistic and the expected difference based on the 95% quantile should be used instead to check for convergence in split frequencies.

4. We implemented the methods described here in the open-source R package `Convenience` (https://github.com/lfabreti/convenience), which allows users to easily test for convergence using output from standard phylogenetic inference software.

2.2 Introduction

In the last two decades, Bayesian inference has become a widely used framework for performing statistical analyses in phylogenetics and macroevolutionary biology [131, 104]. The goal of a Bayesian analysis is to compute the posterior probability distribution of the parameters given the observed data. Unfortunately, we cannot compute this posterior distribution analytically for virtually all realistic and empirically interesting models. Instead, one often applies sampling based methods such as Markov chain Monte Carlo sampling [MCMC, 125, 69]. MCMC algorithms produce (autocorrelated) samples from the desired posterior distribution. Thus, the frequency of parameter samples corresponds to their posterior probability. As with any stochastic sampling based method, the researcher needs to make sure that (a) sufficiently many samples have been obtained, and (b) the samples are indeed representative of the posterior distribution. The process of determining whether these conditions have been met is called *convergence assessment*.

Convergence assessment is a widespread problem for Bayesian analyses using MCMC methods (although one should check for convergence when using any stochastic search algorithms). Every class and lecture about MCMC methods teaches practitioners to check for convergence but in practice these checks are neither standardized nor consistently performed [66]. Unfortunately, traditional MCMC convergence assessment methods from the statistical community have several shortcomings when applied to Bayesian phylogenetics. First, phylogenetic trees are a very peculiar and difficult type of parameter for which common convergence tests that assume continuous parameter values cannot be applied. Thus, specific methods which transform phylogenetic trees into distances have been proposed [*e.g.,* 101]. Second, all widely used convergence assessment approaches used in phylogenetics require manual interaction, often through visual inspection [134, 173, 144]. Visual inspection renders convergence assessment irreproducible by reviewers and other researchers. Moreover, visual inspection makes it unfeasible to apply convergence assessment for genomic datasets with thousands of parameters (*e.g.,* each gene tree and gene-specific substitution model parameters) and for simulation studies with hundreds or thousands of MCMC runs.

In this manuscript we aim to develop a convergence assessment approach for phylogenetics that fulfills the following criteria: (i) it checks whether a single MCMC run needs to be run longer; (ii) it compares multiple independent MCMC runs to check if one of the runs got trapped in an area of parameter space and thus did not sample from the target posterior distribution; (iii) it uses statistically motivated and mathematically derived thresholds, (iv) with longer MCMC runs the chance of convergence increases towards one and does not plateau at a 5% rejection level; and (v) it can be applied without manual interactions. Our motivation and final goal is to provide a tool that, if used and its output provided in publications that used MCMC simulations in phylogenetics, we as readers or reviewers could easily verify if convergence was achieved.

2.3 Materials and Methods

As we stated above, convergence assessment consist of checking whether (1) enough samples have been obtained, and (2) if these samples represent the true posterior distribution. The first aspect concerns the *precision* of the estimators (*e.g.,* the posterior mean or the 95% credible interval). A common question is: “Has the MCMC simulation run long enough?” Or in other words, “Would more samples and/or a longer MCMC run change the estimates?” The most frequent approach to answer this question in phylogenetics is to assess whether the effective sample size (ESS) is larger than 200 [144]. That is, if samples obtained from the chain are equivalent to 200 or more independent samples from the posterior distribution, then one assumes that the MCMC has been run long enough. However, this threshold of 200 samples is arbitrary, as stated in [144]. Here, we turn the question around and ask instead “How many samples do I need to obtain a sufficiently precise estimate?” We will derive the number of effective samples needed to obtain a specified precision. We assume, following standard statistical practice, that our parameter estimates will not change once this precision has been reached [60]. Thus, as suggested by [144], a single MCMC run has been run sufficiently long if the ESS is larger than our derived threshold value. So far, the ESS for convergence assessment is primarily applied to continuous model parameter but not phylogenetic trees [but see 101, 173]. Thus, we will develop and test a new method to compute the ESS for splits of a phylogeny.

The second aspect concerns the *reproducibility* of the stochastic sampling algorithm. It could be the case that the MCMC simulation got stuck in some area of the parameter space and never converged to the true posterior distribution. Therefore, we will adopt a test that compares samples from two independent MCMC simulations. In phylogenetics, the often used approach to compare samples from two independent MCMC simulations is the potential scale reduction factor [PSRF, 53] and the average standard deviation of split frequencies [ASDSF, 98]. Neither the PSRF nor ASDSF have clear and statistically motivated thresholds. Furthermore, the PSRF is very dependent on the shape of the posterior distribution (see Supplementary Material section S6). Therefore, the PSRF does not fulfill our criteria of a robust convergence assessment method that would eventually accept the samples if the MCMC were run sufficiently long and sampled from the true distribution. Instead, we propose to use the Kolmogorov-Smirnov test [KS-test, 94, 156]. The ASDSF also has shortcomings which we will address here.

2.3.1 Precision of an estimator to assess sufficiently many samples

Ideally we would like to compute our parameter estimates as precisely as possible. However, in many situations, such as when the parameter estimate is computed using numerical methods, we can not obtain the estimated value with arbitrary precision. Instead, we content ourselves if the computed parameter estimates is precise to, for example, a certain number of significant digits. The number of significant digits depends on the sample variance. For example, if we want to estimate the average body size of a population then we might want to have at least a precision on the scale of centimeters, but if we want instead to estimate the average flight distance of migrating birds, then a precision up to a few meters could be completely sufficient. Similarly, no one would trust the mean estimate from only a handful of observations but most people would trust an estimate of a population mean if hundreds of observation have been taken. Given these considerations, we specify the number of samples needed from an MCMC based on the desired precision.

Our threshold value for the ESS is derived from the standard error of the mean ($\sigma_{\bar{x}}$), which is defined as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ where σ is the standard deviation of the sample and N the sample size. $\sigma_{\bar{x}}$ measures the error associated with the estimated mean value; or, in other words, the precision of the mean estimator. Both the variance and the sample impact how precise the mean estimate can be. For example, a sample x from a distribution with a larger variation has a higher error associated with the mean (or lower precision) than a sample from a distribution with lower variation. Therefore, we define an acceptable standard error to have a width smaller than or equal to 1% of the width of the 95% probability interval of the true distribution. If we assume a normal distribution as our reference distribution, then the width of the 95% probability interval is approximately equal to 4σ (see Figure S1). Thus, we derive a threshold value for the ESS based on the specified precision of $\sigma_{\bar{x}}$ as

$$\begin{aligned}
\sigma_{\bar{x}} &\leq \frac{\sigma}{\sqrt{ESS}} \\
1\% \times 4 \times \sigma &\leq \frac{\sigma}{\sqrt{ESS}} \\
ESS &\geq 625 \qquad .
\end{aligned} \tag{2.1}$$

We will use this threshold value of a minimum required ESS of 625 as our reference, but other researchers could derive their own justified threshold for the ESS by specifying a different allowed standard error $\sigma_{\bar{x}}$. Table 4.1 shows some examples of the width of $\sigma_{\bar{x}}$ regarding the 95% probability interval and ESS values.

Table 2.1: List of minimum ESS thresholds based on precision of $\sigma_{\bar{x}}$.

Width of $\sigma_{\bar{x}}$	ESS
0.5%	2500
1%	625
1.77%	200
2%	156.25
5%	25

2.3.2 Assessing ESS estimation of continuous parameters

The ESS plays a fundamental role in assessing convergence. Therefore it is crucial that we can estimate the ESS correctly. The ESS can be defined as the number of samples N taken from an MCMC simulation divided by the autocorrelation time (τ)

$$\text{ESS} = \frac{N}{\tau} \qquad . \tag{2.2}$$

Most methods estimate the autocorrelation time τ instead of computing the ESS directly. Nevertheless, Equation 2.2 shows that both the ESS and τ are directly correlated. The ESS (or τ) can be estimated using different approaches [for an overview, see 167]. We investigated three different commonly used ESS estimation implementations: (1) the `R` package `CODA` [138]; (2) the `R` package `MCMCSE` [43]; and (3) our own re-implementation in `R` of the ESS computation algorithm implemented on the software `Tracer` [144]. `CODA` estimates the autocorrelation as an auto-regressive process and estimates the spectral density at frequency 0 [65, 167]. `MCMCSE` uses a batch means approach to estimate the autocorrelation time [55, 44]. `Tracer` uses the initial sequence estimator [158]. We will address these different methods by the names of the tools in which they were implemented (`CODA`, `MCMCSE` and `Tracer`). We test below common variants of MCMC algorithms used in phylogenetics (*e.g.,* Metropolis-Coupled MCMC [54, 5] and adaptive MCMC [63, 64] as well as specific parameters such as the phylogeny).

Autocorrelated Samples: Samples obtained from an MCMC simulation are rarely (not to say never) drawn independently. For that reason, it is important to evaluate the ESS estimation in the case of autocorrelated samples. Unfortunately, we never know the true autocorrelation τ for a specific MCMC algorithm because it both depends on the implementation and data. Therefore, we developed two new algorithms that mimic an MCMC simulation and produce an autocorrelated sample with specified and known autocorrelation time τ. The first algorithm (Algorithm 1) generates N *iid* values and resamples these values while sampling each value at least once and keeping the order of the original values. The second algorithm (Algorithm 2) generates a sequence of $N * \tau$ samples where the $(i + 1)$th sample is either drawn from the chosen distribution with probability $\alpha = \frac{1}{\tau}$ or set to the same value as the i-th sample with probability $1 - \alpha$. Although these two algorithms work differently, they produce samples with the same characteristics of autocorrelated samples by, on average, τ identical values consecutively. Such a behavior can be observed in MCMC samples specifically for phylogenetic trees, but also for continuous parameters when the acceptance rate is low. We included both algorithms for completeness although we present here only the results based on Algorithm 1.

We simulated 1,000 replicates with an ESS of $N = \{100, 200, 300, 400, 500, 625, 800, 1000\}$ with samples drawn from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and varied the autocorrelation time (ACT) between $\tau = \{1, 5, 10, 20, 50, 100, 250, 500, 750, 1000\}$. An ACT of $\tau = 1$ is equivalent to independent sampling and thus represents a baseline comparison (see Supplementary Material Section S2). Only `Tracer` produced robust ESS

Algorithm 1 Simulating samples from an MCMC algorithm with known autocorrelation time by resampling *iid* values.

```
Inputs:
    $N$: the effective number of samples.
    $\tau$: the autocorrelation time.
Initialize:
    $n \leftarrow N * \tau$                          // the total number of correlated samples
    $m \leftarrow n - N$                             // the number of samples to add
  $X \sim \mathrm{norm}(N)$                          // generate $N$ values from some distribution
  $i \sim \mathrm{sample}(1\!:\!N, m, \mathrm{replace} = \mathrm{TRUE})$    // Draw $m$ indices between 1 and $N$
  $all \leftarrow c(1\!:\!N, i)$       // merge the newly sampled indices $i$ with all possible indices 1 to $N$
  $s \leftarrow sort(all, \mathrm{increasing{=}TRUE})$      // sort the indices in increasing order
  $a \leftarrow X[s]$     // create the trace of samples using the independent sample $X$ and the indices $s$
return vector of autocorrelated samples $a$
```

Algorithm 2 Simulating samples from an MCMC algorithm with known autocorrelation time by accepting new values with probability $\alpha = \frac{1}{\tau}$.

```
Inputs:
    $N$: the effective number of samples.
    $\tau$: the autocorrelation time.
Initialize:
    $n \leftarrow N * \tau$                          // the total number of correlated samples
    $\alpha \leftarrow \frac{1}{\tau}$               // the acceptance probability
$X[1] \sim \mathrm{norm}(1)$                         // generate a single value from some distribution
for $i$ in $2\!:\!n$ do                              // generate correlated samples
    $p \sim \mathrm{unif}(1, 0, 1)$                  // draw a uniform random number between 0 and 1
    if $\alpha \geq p$ then
        $X[i] \sim \mathrm{norm}(1)$                 // generate a new single value from some distribution
    else
        $X[i] \leftarrow X[i-1]$                     // stay at the previous value
    end if
end for
return vector of autocorrelated samples $X$
```

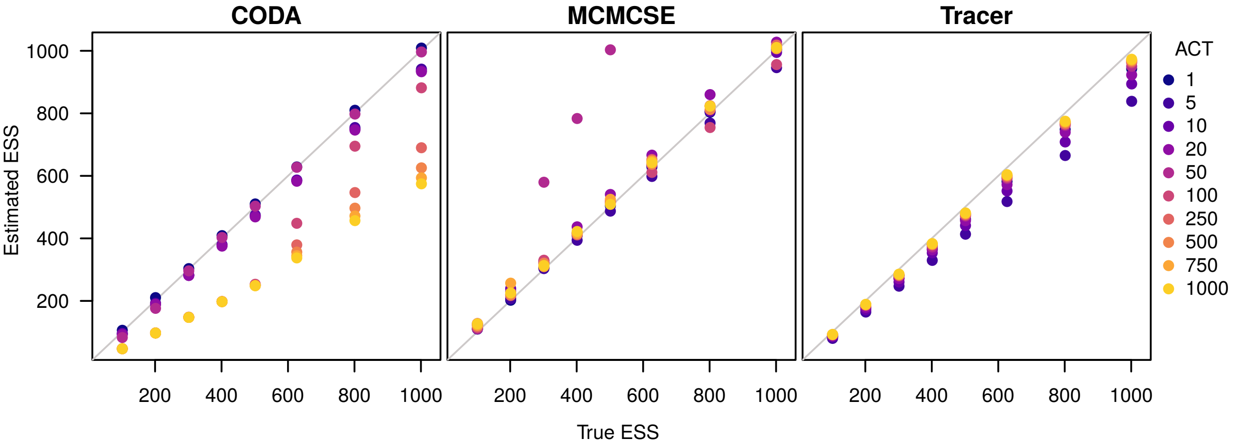


Figure 2.1: The estimated ESS for autocorrelated samples. The x-axis is the true ESS used to generate the sample, the y-axis is the estimated ESS. The left panel displays the ESS estimated using `CODA`, the central panel displays the ESS estimated using `MCMCSE` and the right panel displays the ESS estimated according to `Tracer`. The different colored dots represent different autocorrelation times. The blues dots are the cases that the samples have autocorrelation of 1. As the colors get lighter, the autocorrelation time increases.

estimates for all values of τ (Figure 2.1). `MCMCSE` was robust to most values of the ACT, except for values of τ around 50 where the ESS were overestimated (Figure 2.1 and Figure S3). Our more focused results show that we might have overlooked other problematic ranges for `MCMCSE` and the apparent deviation for $\tau = 50$ is a reproducible outlier. Interestingly, `CODA` performed particularly bad for large τ. Thus, we recommend the use of `Tracer` to evaluate ESS values of MCMC samples over `MCMCSE` and `CODA`.

Samples from Metropolis-Coupled MCMC: In Bayesian phylogenetics, Metropolis-coupled MCMC (MC^3 or MCMCMC) is applied frequently to improve mixing of MCMC chains and thus efficiency [5, 128]. An MC^3 algorithm consists of K independent MCMC chains but samples are only taken from the currently active/cold chain. The MC^3 algorithm proposes swaps between chains every S iterations. For more details about different MC^3 algorithms, *e.g.,* swap frequencies and acceptance frequencies, we refer the reader to [5] and [128].

The samples from an MC^3 algorithm might show different characteristics, for example, when the autocorrelation structure is seemingly broken due to swaps between chains. This behavior could mislead the estimation of the ESS. Therefore, we evaluated the ESS estimation accuracy for MC^3 samples with different swap frequencies. First, we developed a new algorithm with a know autocorrelation τ to mimic MC^3 (Algorithm 3). Our MC^3 algorithm generates K independent MCMC chains using either Algorithm 1 or Algorithm 2. Then, swapping between these independent chains is performed by randomly selecting one of the chains after every S iterations. Note that our Algorithm 3 treats all chains as cold chains and thus always accepts a jump. However, the important feature that we want to test here is the discontinuity introduced by the jumps between chains, and if this discontinuity introduces problems for ESS estimation methods. Thus, we choose to neglect that heated chains traverse the parameter space faster than cold chains. The effect would most likely be irrelevant since our Algorithm 1 and Algorithm 2 produce independent samples once a new sample is accepted.

Algorithm 3 Simulating samples from an MC^3 algorithm with known effective sample size.

```
Inputs:
    $N$: the effective number of samples.
    $\tau$: the autocorrelation time.
    $K$: number of chains.
    $S$: swap frequency between chains.

Initialize:
    $n \leftarrow N * \tau$                    // the total number of correlated samples
    $c \leftarrow 1$                           // the current chain index

for $i$ in $1{:}K$ do                          // generate independent chains
    chain[$i$] $\leftarrow$ sampleMCMC($N, \tau$)   // use Algorithm 1 or Algorithm 2
end for
$X[1] \leftarrow$ chain[$c$][1]                // start with the first value of the first chain
for $i$ in $2{:}n$ do                          // generate the Metropolis-coupled chain
    if $(i \,\%\, S) == 0$ then                // check if we swap at this iteration
        $tmp \leftarrow c$                     // initialize the new chain index
        while $tmp == c$ do
            $tmp \sim$ sample($1 : K$)         // randomly draw a chain new index
        end while
    end if
    $X[i] \leftarrow$ chain[$c$][$i$]          // take the $i$-th value of the current chain $c$
end for

return vector of autocorrelated samples $X$
```

In this evaluation of the three ESS estimation methods, we simulated 1,000 replicates of MC^3 samples using Algorithm 3 with the standard number of chains $K = 4$ and an arbitrarily chosen ACT of $\tau = 20$. We varied the swap frequency $S = \{1, 2, 5, 10, 20, 50, 100\}$ in our simulations.

To derive an expectation of how many effective samples we should get using our MC^3 algorithm, let us consider a window of τ samples. In our independent MCMC chains it takes, on average, τ iterations until we obtain a new value. We swap exactly every S iterations and therefore we swap $\frac{\tau}{S}$ times within a window of size τ. For every time that we swap to a chain that we did not visit yet within the window τ we obtain a new value. Hence, in our sample from the MC^3 algorithm we expect that we have, on average, at least one independent value and at most K independent values within τ iterations. We derive our expectation of the ESS under the MC^3 algorithm as

$$ESS_{MC^3} = \min(K, \max(1, \frac{\tau}{S})) \times N \qquad . \tag{2.3}$$

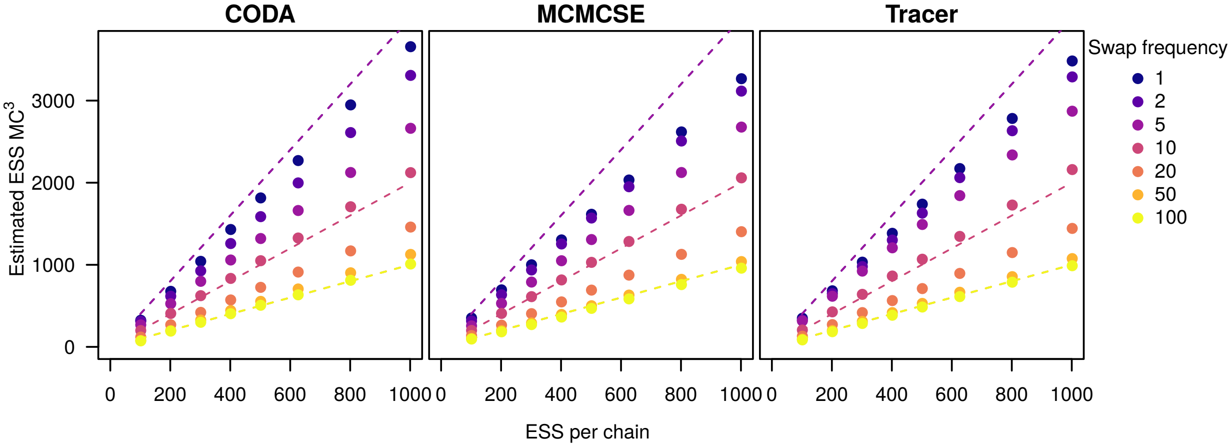


Figure 2.2: Evaluation of ESS estimation when samples are generated using our MC^3 algorithm (Algorithm 3). The x-axis shows the true ESS used to generate the $K = 4$ different chains. The y-axis shows the estimated ESS value for samples from the MC^3 chain. The colored dots show the average ESS estimate over 1,000 replicates. The dashed lines show the expectation for the ESS for each swap value, the expectations for the swaps of 1, 2 and 5 have equal values, thus they are represented by the same dashed line, the same happens for the swaps 20, 50 and 100.

We observed that all three methods perform virtually identical for samples from our MC^3 algorithm (Figure 2.2). Surprisingly, we never estimated the ESS of the MC^3 algorithm as high as $K \times N$ (the number of chains times the ESS per chain). This, in fact, is very reassuring because it indicates that too frequent swapping will not artificially inflate ESS values.

Samples from adaptive MCMC: Almost all implementations of MCMC algorithms in phylogenetics and macroevolution use some type of adaptive MCMC [151, 16, 74, 78]. An adaptive MCMC algorithm changes the tuning parameter of the proposal distribution every κ iterations. For example, the window size of the sliding window proposal could be updated to achieve a target acceptance probability of 0.45 [182, 73]. Thus, the autocorrelation time τ changes during the MCMC simulation. Tuning is performed during a pre-burnin phase where no samples are taken from the MCMC simulation or during the actual MCMC simulation if the tuning parameter is guaranteed to stabilize (*i.e.,* converge) and tuning is performed with low frequency, *e.g.,* $\tau < \kappa$.

Here we only consider the second case where tuning is performed during the actual MCMC simulation while taking samples because the first case is equivalent to samples from a standard MCMC simulation. We sampled from 1,000 replicate MCMC simulations where the total chain length was broken into five intervals. During each interval, we generated samples from an MCMC simulation with $\tau = \{50, 40, 30, 20, 10\}$, that is, in the first interval we had a higher autocorrelation and in the last interval we had the lowest autocorrelation. The size of the intervals varied between $\kappa = \{50, 100, 200, 500, 1000\}$. We expect a true ESS of

$$ESS_\kappa = \sum_{i=1}^{5} \frac{\kappa}{\tau_i} \tag{2.4}$$

because in every interval i of size κ we should obtain $\frac{\kappa}{\tau_i}$ independent samples.

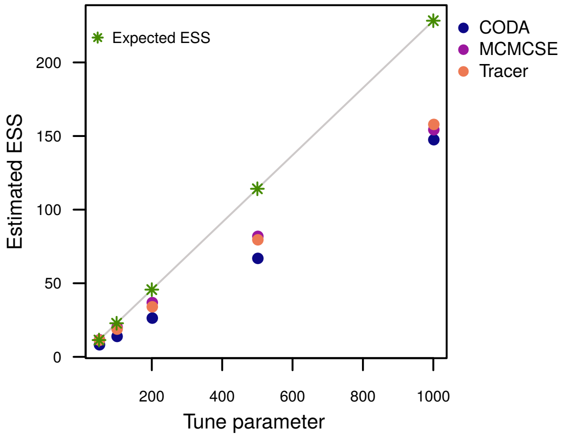


Figure 2.3: Evaluation of the estimated ESS for samples from adaptive MCMC simulations. The MCMC simulation was tuned every κ iterations in a total of four times. Within each of the five intervals we used an autocorrelation of $\tau = \{50, 40, 30, 20, 10\}$. The x-axis shows the tuning frequency and the y-axis shows the estimated ESS. The colored dots show the average ESS of 1,000 replicates. The grey line shows the expected ESS values.

We observed that all three methods produce comparable estimated ESS values when samples are taken from an adaptive MCMC algorithm (Figure 2.3). However, for the adaptive MCMC samples `CODA` was slightly more conservative than `Tracer` and `MCMCSE` compared to our previous experiments. Furthermore, all three methods produced lower ESS estimates than our analytical expectation. This underestimation could be due to fact that all ESS methods compute an average autocorrelation time τ for all samples and not per-window estimates. Our observed estimated ESS values are closer to

$$ESS_\kappa = 5 \times \frac{\kappa}{\bar{\tau}} \tag{2.5}$$

where $\bar{\tau}$ is the average autocorrelation time. Thus, ESS estimates are not inflated and instead conservative when samples are taken from adaptive MCMC algorithms.

Assessing ESS estimation of discrete parameters

Tree topologies are arguably the most important but also most difficult parameter of Bayesian phylogenetic analyses [66]. Tree topologies can be considered as categorical parameters and thus standard ESS estimation algorithms for continuous parameters do not apply. Therefore, we transform the samples of tree topologies into traces of absence/presence of splits (see also [49]). For each split that was sampled at least once during the

MCMC simulation, we construct a trace that has a 1.0 if the split is present in the currently sampled tree and 0.0 otherwise. Hence, we obtain one discrete trace per split.

The three ESS estimation methods introduced above are derived for parameters drawn from continuous distributions. Whether these methods also work well for discrete, binary samples has not been studied. We approximate the sampled split frequencies using samples from a binomial distribution with probability p corresponding to the true posterior probability of the split. Different splits are clearly not independent because they are extracted from the same tree topologies. We show the results assuming independence of splits here for simplicity. In the Supplementary Material we provide the results where we simulated draws from the posterior distribution of trees and extracted splits from the trees.

Following our approach for continuous parameters, we simulated 1,000 replicates and sampled from our MCMC algorithm (Algorithm 1) with a true ESS of $N = 625$ (see Equation 2.2) and an autocorrelation of $\tau = \{1, 5, 10, 20, 50\}$. We varied the true probability of the split p between 0.001 and 0.999 in increments of 0.001.

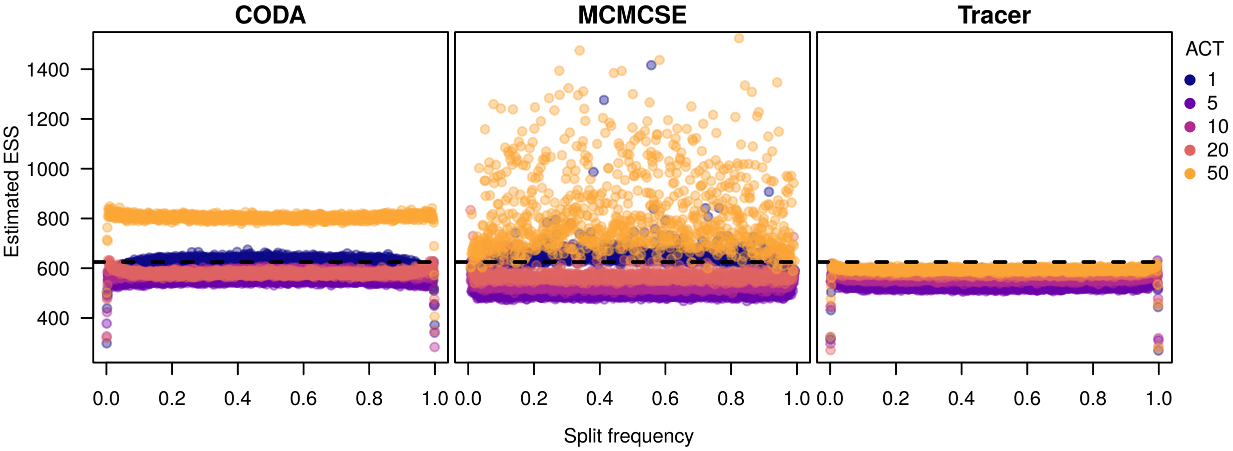


Figure 2.4: Evaluation of ESS estimation for samples taken from a binomial distributions with different autocorrelation times and varying probability p. The true ESS was $N = 625$ and p varied from 0 to 1 with steps of size 0.001. The colored dots show average ESS estimates for 100 replicates. The dashed line shows the true ESS value for the initial sample.

The efficiency of all methods dropped drastically once the true posterior probability gets close to either zero or one (Figure 2.4). This is expected because once the posterior probability is very close to zero or one, it is very likely that all samples either include or exclude the split. The ESS estimation methods are not designed to work well if all or all but one sample are zero (or one, respectively). It is intrinsically impossible to tell if all samples were identical because they are autocorrelated or because they truly should be the same. Thus, we suggest to exclude splits and parameters that have a posterior probability of $p < 0.01$ or $p > 0.99$.

In conclusion, `Tracer` had an overall high precision in recovering the true ESS also when samples are binary (Figure 2.4). As before, we observed that `CODA` performs badly when the ACT was high. Similarly, MCMCSE performed poorly for an ACT of 50. Thus, Tracer is the only method we tested that is robust under all circumstances.

2.3.3 Reproducibility of MCMC runs

In the previous section we focused on the precision of parameter estimates, which concerns the question of whether we have run the MCMC simulation long enough or if we need to run the MCMC simulation longer. In this section, we focus on the second question: "Are our parameter estimates *reproducible*?" Our estimates are reproducible if the MCMC simulation has converged to the true posterior distribution and did not get stuck in some other area of the parameter space. For example, single MCMC runs can achieve a high effective sample size but have sampled posterior probabilities that do not reflect the true posterior probabilities because the MCMC simulation got trapped in an "island in tree-space" [72, 175].

In practice, we can never know with absolute certainty that our MCMC simulation has converged to the true posterior distribution. Instead, one commonly compares multiple independent MCMC runs. In phylogenetics, the two most commonly used approaches to compare independent MCMC runs are the potential scale reduction factor [PSRF, 53] for continuous parameters and the average standard deviation of split frequencies [ASDSF, 134, 98] for tree topologies.

The PSRF computes the ratio of the variance of samples between chains over the variance of samples within a chain. If this ratio converges towards one, *i.e.,* if the variance between independent chains is the same as the variance within a chain, then all MCMC simulation have presumably sampled from the same distribution and thus have converged to the true posterior distribution. However, the PSRF is problematic for two reasons: (1) there is no clear threshold and in practice values between 1.003 and 1.3 have been used [171]; and (2) the PSRF is very sensitive to the shape and variance of the posterior distribution and large independent samples ($N > 1,000$) from the same distribution can yield a PSRF clearly larger than one (see Supplementary Figure S6). Thus, we discourage the use of the PSRF for assessing convergence of Bayesian phylogenetic MCMC simulations. Instead, we suggest using the Kolmogorov-Smirnoff test (KS), as described in [20]. The KS test assesses if two samples are drawn from the same distribution but has —to our knowledge— not been used for convergence assessment in phylogenetics [but see 20, for examples of the KS test for convergence assessment outside phylogenetics]. Below, we discuss and explore below the behavior of the KS test to assess convergence for continuous parameters.

The ASDSF computes the posterior probability of each sampled split in a Bayesian phylogenetic MCMC simulation. Then, the difference between the posterior probabilities per split for two runs are computed. We support the underlying idea of the ASDSF to break the sampled phylogenies into splits and use the frequencies of observing each split. However, computing the average difference between splits is problematic because (1) if the frequency of one split differs strongly (*e.g.,* a frequency of one in the first run but zero in the second run) and all other splits have identical frequencies, then the ASDSF will be low enough to wrongly signal convergence; and (2) the expected difference between two

MCMC samples for the same split depends on the true posterior probability of the split (see below). We introduce our alternative version of the ASDSF below.

2.3.4 Assessing reproducibility of continuous parameter estimates

The two-sided KS test for two samples constructs the empirical cumulative distribution function of the samples. The test statistic D is the largest difference of the two empirical cumulative distribution functions F_1 and F_2, $D = \max_x |F_1(x) - F_2(x)|$. The D statistic shows a significant departure from the expectation that both samples were drawn from the same distribution if $D > \sqrt{-\ln(\frac{\alpha}{2}) \times \frac{1}{2} \times \frac{m+n}{m \times n}}$ at a significance level α, assuming that the first sample has an ESS of n and the second sample an ESS of m.

If we would use the standard approach to define the threshold for D based on the number of samples, then we would reject on average α pairs of MCMC simulation regardless of how long we ran the MCMC simulations. To circumvent this problem, we fix $\alpha = 0.01$ and $N = 625$ (see Equation 2.2) so that $D_{crit} = 0.0921$. Thus, with more effective samples we should obtain a more precise estimate of D and therefore avoid incorrectly rejecting runs that had truly converged as often.

We assessed the false-rejection rate and power of the KS test with our threshold D_{crit} by simulating 1,000 replicated pairs of samples with $N = \{100, 200, 625, 1000\}$ *iid* values drawn from a normal distribution. The first normal distribution had a mean $\mu_1 = 0$ and standard deviation $\sigma = 1$. The mean of the second normal distribution differed by $\mu_2 = \{0.0, 0.04, 0.08, \ldots, 0.8\}$, thus representing that the two means differed by 0% to 20% of the 95% probability interval.

When the samples where drawn from the same distribution ($\mu_1 = \mu_2$), then we observed a false rejection rate of 0.01 for a sample size of 625 (Figure 2.5). This rejection rate is exactly expected for an $\alpha = 0.01$. If we increased the sample size to 1,000, then the false rejection rate decreased to 0.0003. Thus, more samples, *i.e.,* longer MCMC runs, will increase the chances that the chains are assessed as converged if the samples are truly from the same distribution. When the mean of the distributions have a 10% difference ($\mu_2 = \mu_1 + 0.1 \times 4 \times \sigma$), then we correctly rejected convergence with a rate of 0.9974 for a sample size of 625 and 0.9995 for a sample size of 1,000.

In the previous section we defined that our mean estimate is precise enough if the standard error is smaller than 1% of the 95% probability interval. Here we showed that the KS test has very strong power to reject runs if the true means of the samples was different by 10% or more, and has an acceptable power of 0.95 when the means are different by 8% (Figure 2.5). Increasing the ESS further would both decrease the standard error of the mean, *i.e.,* increase our precision, and slightly increase the power to correctly reject convergence when the samples are truly from different posterior distributions.

The KS test is well established for testing if two samples are drawn from the same underlying distribution, which we also illustrated in Figure 2.5. However, the KS test is less established as a convergence assessment tool for autocorrelated samples drawn from

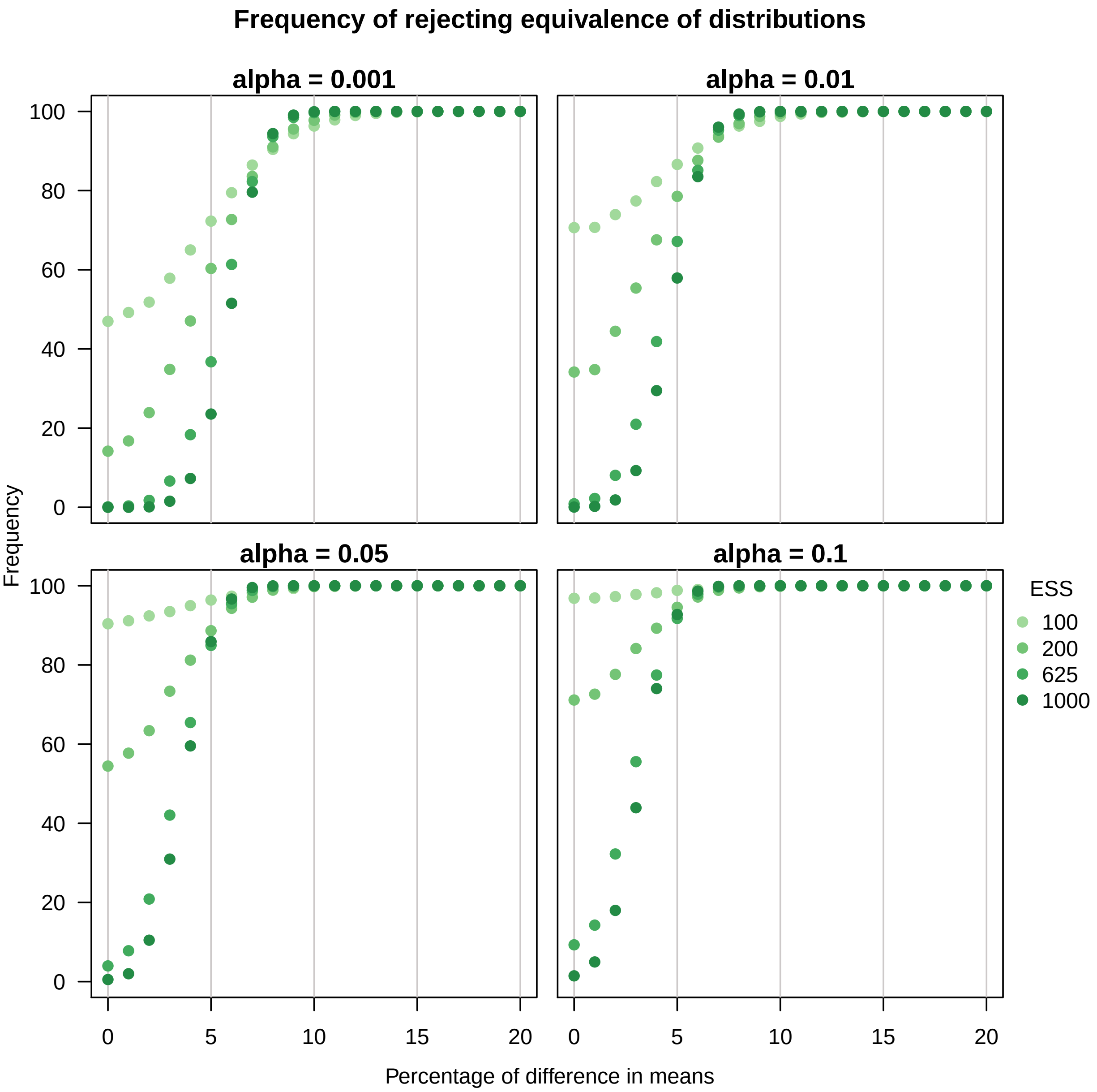


Figure 2.5: Testing the power of the Kolmogorov-Smirnov test statistic to distinguish between samples from two different distributions. The two samples were drawn from different normal distributions with different mean values. For each combination of difference in means, 1,000 replicates were tested and the frequency of rejecting the null hypothesis that both samples were drawn from the same distribution was computed (y-axis). The x-axis displays the difference in the means of the distributions with regard to the 95% probability interval of the normal distribution with mean 0 and standard deviation 1.

MCMC simulations [but see 20]. Therefore, we explored if autocorrelation affects the behavior of the KS test. We performed the same test as before (1,000 replicates, two samples from normal distributions where the mean changed by $\mu_2 = \mu_1 + X \times 4 \times \sigma$) but introduced autocorrelation of $\tau = \{1, 5, 10, 20, 50\}$ into the samples using Algorithm 1. We observed that there is no impact of using autocorrelated or uncorrelated samples for the false rejection rate and power of the KS test (Figure 2.6 for $N = 625$).

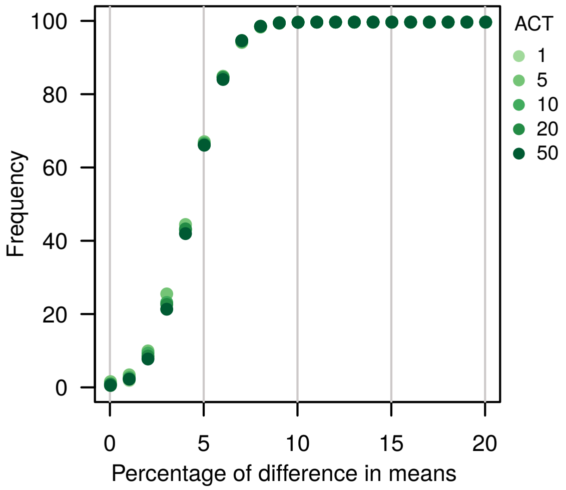


Figure 2.6: Kolmogorov-Smirnov test to detect difference in distribution for autocorrelated samples. The two samples were drawn from different normal distributions with different mean values and different ACT values. Here we only show the results for $N = 625$. For each combination of difference in means, 1,000 replicates were tested and the frequency of rejecting the null hypothesis that both samples were drawn from the same distribution was computed (y-axis). The x-axis displays the difference in the means of the distributions with regard to the 95% probability interval of the normal distribution with mean 0 and standard deviation 1.

2.3.5 Assessing reproducibility of discrete parameter estimates (split frequencies)

Phylogenetic trees and split frequencies do not yield continuous distributions which can be compared using the KS test. Instead, one often plots the split frequencies of one run against a second run (xy-plot). If there is a strong deviation from the diagonal line, then the two runs are assessed as non-converged. Quantitatively, one can compute the average [98] or maximum [72] deviation of split frequencies. However, an under-appreciated characteristic is that the expected difference, or standard error, of split frequencies depends strongly on the true split frequency.

In our view, the ASDSF is not sensitive enough to detect outliers in estimated split frequencies. For large trees with many splits, there will be many splits with a very low frequency. These splits overwhelm the computation of ASDSF and outliers, such that even a difference in posterior probability as large as 1.0 in one run and 0.0 in the second might not be detected. Another unsolved issue with the ASDSF is that no theory for a threshold is provided and the default thresholds are applied to all tree sizes. Additionally, the difference for each split has equal weight although the stochastic difference in split frequencies depends on the true split frequency.

Samples of split frequencies can be treated as a series of 1.0 and 0.0 (presence and absence, as we did above). Thus, we can consider samples of split frequencies as draws from a binomial distribution. For a binomial distribution, we can actually apply an exact test and compute the expected difference based on a given quantile analytically. We call this test the expected difference of split frequencies (EDSF) because it represents the difference in split one expects to obtain from two samples and given a specific quantile and true split frequency. Note, the EDSF should not be interpreted as the expected value (first moment) of the distribution of difference of split frequencies.

Let us denote the difference of split frequencies between two samples for a split with true frequency p as Δ_p^{sf}. We compute the EDSF for Δ_p^{sf} as follows. First, we compute all possible outcomes of split frequencies between two runs with absolute difference $|\frac{i}{N} - \frac{j}{N}|$ and their corresponding probabilities $P_{binom}(i|N,p) \times P_{binom}(j|N,p)$. Then, we order the pairs of split frequency differences and probabilities by the size of the split frequency difference and sum the probabilities. Thus, we can compute any quantile of expected split frequency differences. For our purposes, we define that two samples of phylogenetic trees are from the same underlying posterior distribution if all splits have a difference smaller than the 95% quantile of Δ_p^{sf} for $N = 625$ and $p = \frac{\hat{p_1}+\hat{p_2}}{2}$, where $\hat{p_i}$ is the estimated split frequency for run i.

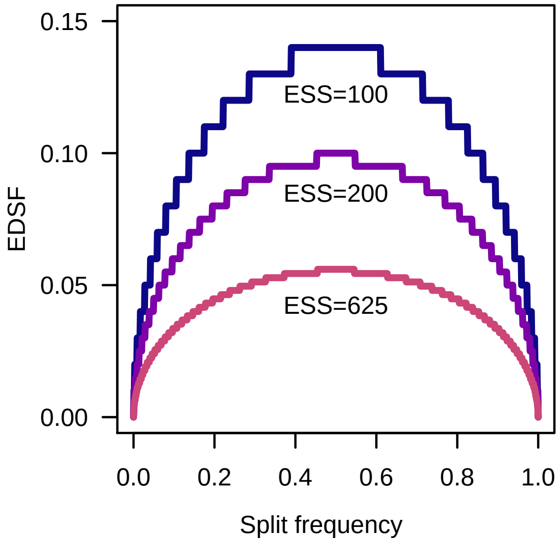


Figure 2.7: The expected difference in split frequencies for ESS of 100, 200 and 625. The x-axis is the true value of the split frequency. The y-axis is the expected difference in split frequencies. The effect of increasing the ESS is the decrease of differences in frequency of sampled splits.

Figure 2.7 shows curves of the expected difference in split frequencies (EDSF) for different samples sizes. As expected, the EDSF is smallest when the true split frequency is close to the boundaries zero or one, and largest when the true split frequency $p = 0.5$. Moreover, the EDSF decreases for larger sample sizes (*i.e.,* larger ESS). Interestingly, for all possible true split frequencies we expect at most a difference of 0.056 (or 0.1 and 0.14) if we have 625 independent samples (or 200 and 100 respectively).

2.4 Convenience: Implementation of convergence assessment and interpretation of output

We implemented the methods described here in the stand-alone R package Convenience. Convenience is open-source and can be downloaded and installed from https://github.com/lfabreti/convenience. Currently, Convenience supports the output file formats from RevBayes [78], MrBayes [151], BEAST [16] and PhyloBayes [107]. Additionally, it is also possible to assess convergence in outputs containing only continuous or discrete parameters (trees).

The main function of Convenience is checkConvergence(), which runs the complete convergence assessment pipeline and the thresholds established in this article.

```
> test_convergence <- checkConvergence("convenience/example/")
```

The checkConvergence() function checks first the best burn-in value. If the burn-in is greater than 50%, the function stops and tell the user that the burn-in is too large. Otherwise, the function continues the convergence assessment by applying the described methods to the continuous and discrete parameters. In addition, the user has the possibility to use each method separately in different functions and change the thresholds as suited. A more detailed explanation of the functions can be found in the tutorial at https://revbayes.github.io/tutorials/convergence/.

Once the convergence assessment has been performed, the user has the option to print or plot the results. Convenience produces four main plots:

```
> plotEssContinuous(test_convergence)
> plotEssSplits(test_convergence)
> plotKS(test_convergence)
> plotDiffSplits(test_convergence)
```

plotEssContinuous displays the ESS values for all continuous parameters and all MCMC replicates within one plot (Figure 2.8a). If MCMC convergence has been achieved, then all ESS values in this histogram are on the right side of the minimum ESS threshold. plotEssSplits displays the ESS values for all splits and all MCMC replicates (Figure 2.8b). Again, if MCMC convergence has been achieved, then all ESS values in this histogram are on the right side of the minimum ESS threshold. plotKS displays the KS-score for all continuous parameters and all pairwise comparisons of MCMC replicates (Figure 2.8c). If MCMC convergence has been achieved, then all KS scores in this histogram are

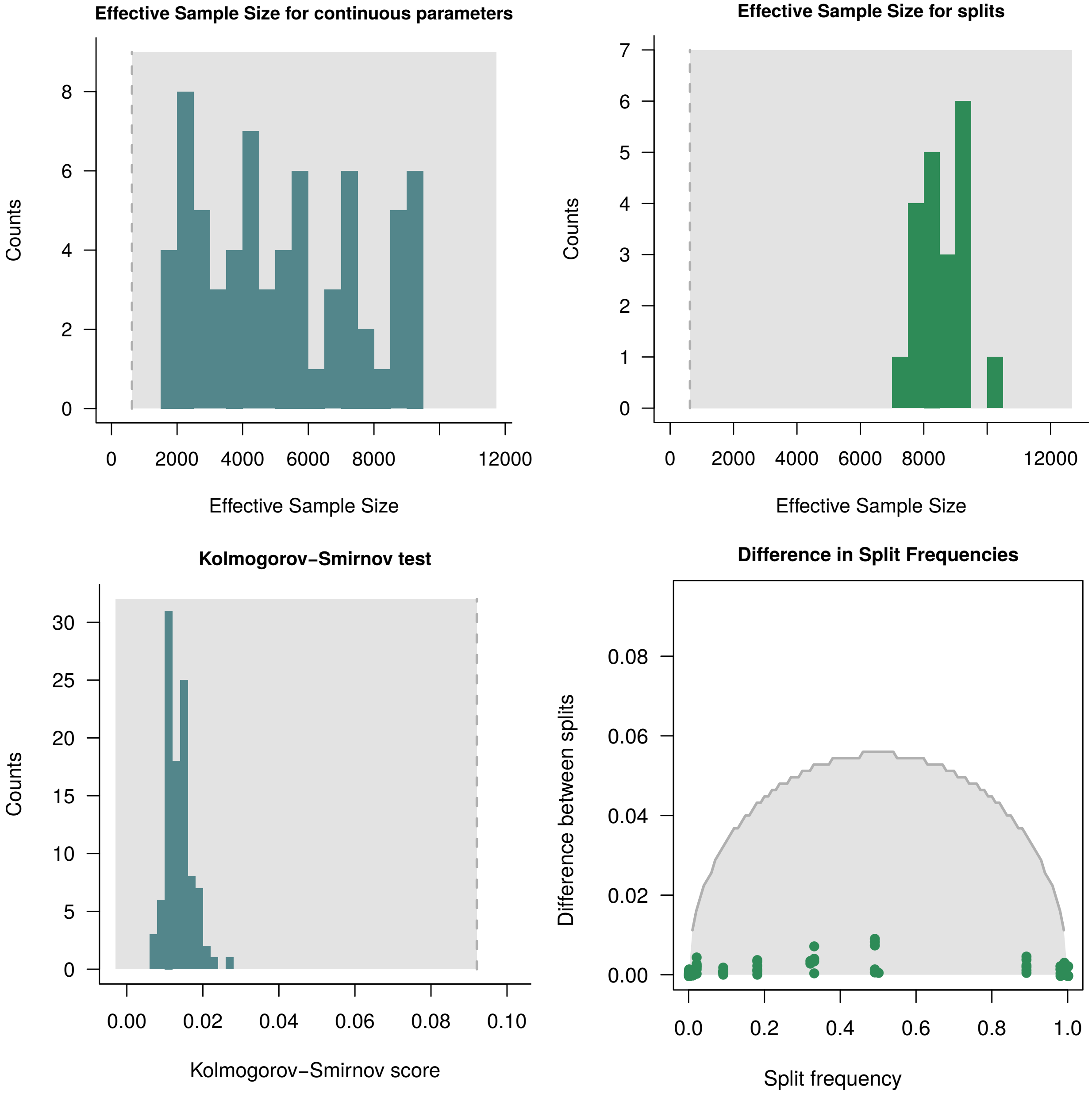


Figure 2.8: The plots generated with `Convenience` for summarizing and visualizing the results from the convergence assessment. Here we used the MCMC example in the `RevBayes` tutorial https://revbayes.github.io/tutorials/ctmc/, see [73]. Top-left: the histogram of estimated ESS values for the model parameters (continuous parameters). Top-right: the histogram of calculated ESS for the splits. In both histograms the dashed lines represents the minimum ESS threshold of 625. Bottom-left: histogram of the Kolmogorov-Smirnov (KS) test for the model parameters, the dashed line represents the threshold for the KS test. Bottom-right: the observed difference is split frequencies in the green dots and the maximum threshold for split frequencies based on the expected difference between split frequencies (EDSF) in the gray curve. For all plots the gray area shows where the values should be if the analysis achieved convergence.

on the left side of the maximum KS threshold. plotDiffSplits displays the difference in split frequencies for all splits and all pairwise comparisons MCMC replicates (Figure 2.8d). If MCMC convergence has been achieved, then all differences in split frequencies are below the maximum split frequency threshold.

In summary, applying Convenience is fully automatic. Thus, the package can be used interactively or in batch-mode (*e.g.,* on computer clusters). If an MCMC analysis includes a plot similar to Figure 2.8, then it is easy to verify convergence assessment. The text output of Convenience can be easily parsed to perform hundreds or thousands of convergence assessments.

2.5 Discussion and conclusions

2.5.1 Convergence thresholds for nuisance parameters

In many phylogenetic analyses, there are some focal parameters and other parameters are nuisance parameters. For example, in a traditional phylogeny estimation analysis, the phylogenetic tree is the focal parameter and the substitution model parameters might be nuisance parameters. So far, the thresholds used in convergence assessment are applied equally to all parameters. That is, one requires that all parameters have an ESS>625 (or whichever other threshold was used). One might argue that checking for convergence for the nuisance parameters is not as relevant as checking for convergence for the focal parameters. Thus, it could be possible to use more relaxed thresholds for the nuisance parameters. However, we find no theoretical support for treating nuisance and focal parameters differently. Whether relaxing the precision, and consequently the ESS, for the nuisance parameters can affect the convergence of the focal parameters needs to be further investigated. For now, we advise on using the same criteria for all underlying parameters of the model.

2.5.2 Future directions

Our approach and evaluation presented here has several limitations and is only another small step towards robust and automatic convergence assessment. First, we did not test how well either of these methods perform when the posterior distribution is multi-modal. Second, parameter non-identifiability might confuse convergence assessment tests. For example, hidden Markov models can produce seemingly multi-modal posterior distributions if the hidden state is arbitrarily labelled and not ordered [105, 7]. Third, we currently reject convergence if a single MCMC run did not converge or produced a different posterior distribution. When MCMC mixing is very challenging, it might happen that many replicated MCMC runs are performed and only a small subset converged. Thus, it would be desirable to automatically identify the subset of MCMC runs that converged.

2.5.3 Conclusions

Convergence assessment should be a mandatory, objective, simple and reproducible step in any Bayesian analysis that relies on samples from the posterior distribution. In this manuscript we presented and explored our approach, which is implemented in the R package `Convenience`. We identified two crucial aspects when running an MCMC simulation: (i) Has the MCMC ran long enough?; and (ii) Do the samples represent the true posterior distribution?

We addressed the first question by focusing on the precision of the posterior mean estimate. If we have sufficiently many samples from the posterior distribution, then our standard error of the mean estimate will be sufficiently small. Thus, one only needs to check if the effective sample size is large enough and we provide some objective criteria to choose a threshold for the minimum ESS. If we accept an SEM of 1% of the 95% credible interval, then a minim ESS of 625 is required. We tested three commonly used methods to estimate the ESS: spectral density estimators of an auto-regressive process (`CODA`), batch means (`MCMCSE`), and initial sequence estimator (`Tracer`). Our assessment included: (a) independent samples; (b) autocorrelated samples; (c) samples from Metropolis-Coupled MCMC simulations; and (d) samples from adaptive MCMC simulations. We found that only the initial sequence estimator (`Tracer`) was robust in all scenarios and for all ranges of autocorrelations.

Focusing on phylogenetic applications, we showed that samples from the posterior distribution of phylogenetic trees can be converted into binary traces of absence/presence of splits. We tested such approach for cases where the ESS of the trees was known and concluded that the estimated ESS of splits is a good proxy for the ESS of trees. The ESS estimation works robustly on these discrete, binary traces and can be applied in the same way.

We addressed the second question by focusing on reproducibility of multiple MCMC runs. We observed that the commonly used potential scale reduction factor (PSRF) is not robust to the shape of the posterior distribution. For example, samples from a lognormal distribution yield a PSRF that is asymptotically significantly larger than 1.0. We suggest the Kolmogorov-Smirnov test instead, which we showed to work well also for autocorrelated samples.

We modified the average standard deviation of split frequencies (ASDSF) to use instead an analytically derived expected difference between split frequencies (EDSF). We demonstrated that the EDSF depends on the true frequency of a split, and thus the same thresholds for all splits cannot be used.

Chapter 3

The Expected Behavior of Posterior Predictive Tests and its Unexpected Interpretation

3.1 Abstract

Poor fit between models of sequence or trait evolution and empirical data is known to cause biases and lead to spurious conclusions about evolutionary patterns and processes. Bayesian posterior prediction is a flexible and intuitive approach for detecting such cases of poor fit. However, the expected behavior of posterior predictive tests has never been characterized for evolutionary models, which is critical for their proper interpretation. Here, we show that the expected distribution of posterior predictive p-values is generally not uniform, in contrast to frequentist p-values used for hypothesis testing, and extreme posterior predictive p-values often provide more evidence of poor fit than typically appreciated. Posterior prediction assesses model adequacy under highly favorable circumstances, because the model is fitted to the data, which leads to expected distributions that are often concentrated around intermediate values. Non-uniform expected distributions of p-values do not pose a problem for the application of these tests, however, and posterior predictive p-values can be interpreted as the posterior probability that the fitted model would predict a dataset with a test statistic value as extreme as the value calculated from the observed data.

3.2 Introduction

Statistical models are mathematical abstractions of reality that employ simplifying assumptions to capture important features of complex systems. As long as such assumptions do not depart from reality too strongly, statistical models can provide important insights into the systems they represent. However, if assumptions violate reality in meaningful ways, models lose both utility and reliability [51, 24].

Applied statistical fields, including phylogenetics and molecular evolution, need tools to assess when their models fail as meaningful abstractions of reality. The use of these tools is often referred to as testing absolute model fit or testing model adequacy. In a Bayesian framework, one way to test a model's absolute fit is through posterior prediction [152, 14].

Posterior prediction involves fitting a Bayesian model with parameters θ to observed data y. We then draw S values of θ from the posterior distribution, $p(\theta|y)$, and based on these posterior draws $(\theta_1 \cdots \theta_S)$, we simulate S predictive datasets $(y_1^{rep} \cdots y_S^{rep})$ of the same size as y. To perform a posterior predictive check of our model, we start by selecting a test statistic, $T(y)$, that can be calculated on the observed and predictive datasets in order to compare them. One way to summarize the comparison between $T(y)$ and $T(y^{rep}_{1 \ldots S})$ is to calculate the fraction of predictive datasets that have test statistic values smaller or larger than the observed. If smaller, we can define the posterior predictive p-value as $P(T(y^{rep}) < T(y)|y)$ and, if larger, $P(T(y^{rep}) > T(y)|y)$ [see 83, for a description of different posterior predictive p-values]. In either case, particularly large or small p-values indicate poor fit between the model and data.

The steps outlined above describe the mechanics of performing posterior prediction, but the more formal mathematical description of the quantity being estimated by this procedure is given by

$$p = \int_{-\infty}^{T(y)} \left(\int_{-\infty}^{\infty} p(T(y^{rep})|\theta)p(\theta|y)d\theta \right) dT(y^{rep}). \tag{3.1}$$

Here, integration inside the parentheses describes the posterior predictive distribution of test statistic values, T, based on the posterior distribution of θ, while the outer integration describes calculation of the lower tail-area probability of this distribution with an upper limit defined by the empirical test statistic value, $T(y)$.

Despite statistical literature discussing the behavior of posterior predictive tests in general (e.g., [124]), expected distributions have never been characterized for posterior predictive p-values in phylogenetics and molecular evolution. Therefore, we aim to characterize the expected distributions of posterior predictive p-values for phylogenetics, compare such distributions across different types of test statistics, and understand how different parameters affect these expectations. To do so, we performed a broad set of simulations and posterior predictive analyses. We used the same model for simulation and analysis, and we drew parameter values for simulation from the prior distributions of the model parameters.

Our results convincingly demonstrate that posterior predictive p-values should not be interpreted like p-values from frequentist hypothesis tests. If misinterpreted in this way, posterior predictive tests will not be used to greatest effect and the strength of evidence for poor model fit will be underestimated because the expected distributions of posterior predictive p-values are, in many cases, highly non-uniform with a concentration of values near 0.5.

3.2.1 Definition and Comparison of p-values

While posterior predictive p-values are called p-values because they involve the calculation of tail-area probabilities, they are distinct from several other types of p-values that we describe here for clarity.

The traditional frequentist p-value used in a hypothesis testing framework is defined as the probability of obtaining a test statistic value, $T(y^{rep})$ that is as or more extreme than the observed test statistic value, $T(y)$, if the null hypothesis (with a value of θ fixed *a priori*) is true. If we focus on the probability of obtaining observations that are smaller than the observed, the frequentist hypothesis testing p-value can be described by the cumulative distribution function,

$$p = \int\limits_{-\infty}^{T(y)} f(T(y^{rep})|\theta) dT(y^{rep}). \tag{3.2}$$

Note that θ, and correspondingly the distribution of $T(y^{rep})$, does not depend at all on y in this case.

The parametric bootstrap p-value is similar in formulation to the frequentist p-value for testing a null hypothesis, but with estimated parameter $\hat{\theta}$. That is, instead of assuming a value of θ that is fixed *a priori*, we use the maximum-likelihood estimate, $\hat{\theta}$, based on y:

$$p = \int\limits_{-\infty}^{T(y)} p(T(y^{rep})|\hat{\theta}) dT(y^{rep}) \tag{3.3}$$

Parametric bootstrapping is a frequentist analogue to posterior predictive model checking, but does not involve prior distributions or integration across different values of θ. The estimated value of $\hat{\theta}$ and the distribution of $T(y^{rep})$ do depend on y in this case.

The prior predictive p-value [19] is the Bayesian equivalent of the traditional frequentist hypothesis test, in the sense that the (probabilities of) parameter values defining the model are fixed *a priori* and do not depend on the observed data, y. The main difference is that, in the case of the prior predictive p-value, the cumulative distribution function is computed while integrating over different values of θ weighted by the *prior* probability of each, $p(\theta)$,

$$p = \int\limits_{-\infty}^{T(y)} \left(\int\limits_{-\infty}^{\infty} p(T(y^{rep})|\theta) p(\theta) d\theta \right) dT(y^{rep}). \tag{3.4}$$

A graphical depiction of the similarities and differences across p-values is given in Figure 1.

Table 3.1: Settings for simulations and posterior predictive analyses.

Setting	Substitution Model	Number of taxa	Number of sites	Mean branch length
1 (Baseline)	JC	16	100	0.1
2	JC	64	100	0.1
3	JC	16	1000	0.1
4	JC	16	100	0.02
5	GTR++I	16	100	0.1

3.3 Results

The expected distribution of posterior predictive p-values varies by both test statistic and simulation condition, but is typically non-uniform (Figs. 2 and 3). Instead, these distributions are more concentrated around intermediate values, with fewer values near 0 or 1. This expectation has gone unappreciated in the discussion and applications of these tests to phylogenetics and molecular evolution [e.g., 14, 21, 24], but has important consequences for how results are interpreted.

In this study, we investigated the expected behavior of both data- and inference-based test statistics. Briefly, data-based test statistics can be calculated directly based on the properties of sequence alignments (e.g., the variance in GC content across sequences), while inference-based test statistics are calculated based on the properties of inferences conditional on those alignments and a model (e.g., the 99th percentile in the ordered vector of RF distances describing distances between trees sampled from the posterior distribution). Despite these differences, both types of test statistics have expected distributions that exhibit the same concentration of posterior predictive p-values near intermediate values.

While most test statistics have non-uniform expected distributions, ancillary test statistics (those statistics whose probability distributions do not depend on model parameters) should have uniform expected distributions [124, 50], because fitting the model has no effect for these statistics. This expectation explains the distributions of p-values for statistics based on GC content in our results (Fig. 2). Mean GC content is an ancillary statistic of the Jukes-Cantor model (JC), since this model assumes equal nucleotide frequencies, and we see roughly uniform expected distributions for Mean GC when using JC. However, Mean GC content is not ancillary for the GTR+I+G model, so the expected distribution is more concentrated around 0.5 in this case (bottom right of Fig. 2). Variance in GC content across sequences is ancillary for both models, since both assume that GC content does not vary across the tree, resulting in consistently uniform expected distributions.

Inference-based test statistics, by definition, depend on parameters of the model and cannot be ancillary. As a result, expected distributions of posterior predictive p-values for these statistics are never uniform (Fig. 3) and are always more concentrated near 0.5 than 0 or 1. Expected distributions for inference-based statistics tend to be more consistent than for data-based statistics, although some become markedly more peaked when dataset size increases either in terms of number of sites or number of taxa.

Several statistics, both data- and inference-based, have expected p-value distributions that are essentially fixed at 0.5 for some conditions (e.g., the effect of more taxa on the number of invariant sites or the topological entropy; Figs. 2 and 3). Posterior predictive p-values can be interpreted as the posterior probability of observing a test statistic value that is as extreme as the observed value [50], so these (nearly) invariant distributions may indicate that fitted models almost always predict datasets with (nearly) the same test statistic value as the observed. This interpretation makes sense for both the number of invariant sites and entropy test statistics with large numbers of taxa in our simulations ("Setting 2ïn Table 3.1, "More Taxaïn Figs. 2 and 3). For datasets simulated with these conditions, nearly every site in the alignment will have some variation, causing the number of invariant sites to be approximately 0 for all observed and posterior predictive datasets. Similarly, these conditions lead to very diffuse posterior distributions of phylogenetic topologies, such that every topology sampled from the posterior distribution is unique and the estimated entropy is the same across datasets.

Expected p-value distributions for some test statistics are multimodal (e.g., the minimum pairwise difference statistic; Fig. 2). Multimodal distributions typically occur with discrete test statistics that adopt a small number of possible values. Such distributions are not unique to phylogenetics and molecular evolution and present no particular difficulties for interpretation [50]. However, these expected distributions are worth bearing in mind when interpreting such values in empirical studies. In these cases, small changes in test statistic values can lead to seemingly large changes in p-values.

While p-values have received the most attention as a way to summarize the results of posterior predictive tests, an alternative approach is the use of effect sizes [33, 83]. Briefly, effect sizes measure the distance between the empirical test statistic value and the median of the posterior predictive distribution, normalized by the standard deviation of the posterior predictive distribution. Effect sizes are useful for understanding the magnitude of the discrepancy between the observed and predicted values, even when the observed value is highly improbable given the model. We used the same set of simulations and analyses to characterize the expected distributions of effect sizes (Figs. 4 and 5). These expected effect size distributions make sense in light of the expected distributions of p-values, although there is a preponderance of values near 0 rather than near 0.5. Due to the way effect sizes are calculated, their expected distribution is not uniform even when the expected distribution of p-values is uniform. As an example, see the distributions of expected effect sizes for Mean GC content for any of the analyses employing a JC model (Fig. 4). As with expected distributions of p-values, many of the distributions of effect sizes are multi-modal, due to the discrete nature of many test statistics. However, in all cases, these values are nearly always < 2.0. This result stands in contrast to our experiences analyzing empirical data sets, where effect sizes are frequently $\gg 10.0$ [33].

3.4 Discussion and Conclusions

P-values by definition represent the probability, conditional on the model, of observing data that are more extreme than what has actually been observed. A p-value that is very small or very large indicates that the observed dataset is an outlier relative to model expectations and possibly reflects poor absolute model fit. In a standard frequentist hypothesis test, the model corresponds to the null hypothesis and poor model fit would lead to its rejection. Frequentist p-values for hypothesis testing are explicitly constructed to have uniform distributions in order to control false positive rates. Importantly, this uniformity of p-values stems from the use of fixed (*i.e.,* not fitted) parameter values.

Posterior predictive p-values, on the other hand, use model parameter values that have been fitted to the observed data (Fig. 1). The probability that the observed data are more extreme than expected is always reduced relative to tests using fixed values, because the model is given the opportunity to explain the data as well as possible. Thus, expected distributions of posterior predictive p-values tend to be concentrated around 0.5 [124, Figs. 2 and 3], although the precise shape can vary by both test statistic and analysis condition. In practice, non-uniform distributions can be precisely what we want if our goal is to assess the ability of our model to capture certain aspects of our observed data [50]. If our model always does a good job of predicting these features, then the expected distribution of posterior predictive p-values should reflect that.

We focused here on posterior predictive p-values, computed in a Bayesian framework, but similar considerations apply to p-values from parametric bootstrapping analyses (Fig. 1) conducted in a frequentist framework. Since parametric bootstrapping also involves fitting a model to a dataset, it should also produce expected distributions of p-values that are non-uniform. In fact, expected distributions from parametric bootstrapping may be much more concentrated than those from posterior prediction, because the effect of posterior uncertainty keeps the expected distributions from becoming too peaked in a Bayesian setting.

If posterior predictive p-values are misinterpreted as frequentist hypothesis testing p-values, the evidence for poor model fit will usually be underestimated. A posterior predictive p-value of 0.05 typically has a $< 5\%$ probability of occurring when the assumptions of an analysis exactly match the data-generating process. However, again, it is best to avoid framing posterior predictive tests in frequentist hypothesis testing terms. The goal of posterior predictive tests should not be to reject a model as “true“[51], since we know that none of the models fully represent the complexity of real evolutionary processes. Rather, these tests indicate the extent to which the model’s simplifications are problematic for explaining important features of the data.

In the course of this study, we simultaneously characterized expected distributions of posterior predictive p-values for multiple test statistics and our results demonstrate that many of these test statistics are correlated. Strong correlations mean that a count of the number of statistics with small p-values is not an effective way to measure the overall degree of fit between model and data. Small p-values for posterior predictive tests with two uncorrelated test statistics would provide more insight than small p-values for many such tests with highly correlated test statistics.

Empirical application of posterior predictive tests in phylogenetics and molecular evolution has frequently resulted in extremely small p-values across many different datasets using a variety of different test statistics [e.g., 106, 45, 185, 33, 146, 36, 83]. Based on

Table 3.2: Parameters of phylogenetic models and their associated prior distributions.

Parameter	Description	Prior Distribution	Parameters of the distribution
Ψ	Topology of the tree	Uniform	Num. of Taxa=16 or 64
bl	Branch lengths	Exponential	λ_{bl} =10 or 50
π	Equilibrium base frequencies	Dirichlet	α_{π}=(1,1,1,1)
er	Exchangeabilities	Dirichlet	α_{er}=(1,1,1,1,1,1)
α	Shape of the Gamma distribution	Exponential	λ_{α}=0.05
I	Proportion of invariant sites	Beta	(α_I, β_I)=(10,20)

the nature of the expected distributions that we have characterized here, these empirical results often represent even stronger evidence than has been appreciated that commonly applied models in phylogenetics are seriously inadequate. An important future direction will be to more comprehensively characterize those aspects of empirical datasets that are consistently fit poorly by commonly employed models of sequence and trait evolution. This characterization can guide the most efficient development of effective new models.

To our knowledge, this paper is the first characterization of the expected distribution of posterior predictive p-values for models commonly used in phylogenetics and molecular evolution. Our hope is that the results presented here clarify the interpretation of empirical assessments of absolute model fit using posterior predictive tests. These tests can highlight important mismatches between model assumptions and the actual biological processes that shape genome sequences. Critical thought must be given to how models are applied in order to gain insight into evolutionary patterns and processes [24].

3.5 Materials and Methods

3.5.1 Data Simulation

To understand the expected distributions of posterior predictive p-values when analysis conditions precisely match those under which the data were generated, we first simulated alignments of DNA sequences using a baseline set of conditions: a JC model [87] of sequence evolution, a 16-taxon tree from a uniform distribution, alignments with 100 sites, and exponentially distributed branch lengths with a mean of 0.1 (Table 3.1, Setting 1). We then simulated alignments under four additional sets of conditions that varied each baseline setting individually. We increased the size of the tree to 64 taxa (Setting 2), increased the length of the alignment to 1,000 sites (Setting 3), reduced the mean branch length to 0.02 (Setting 4), and used the General Time-Reversible model (GTR) [166] with Gamma-distributed rate variation among sites as four discrete rate categories (Γ) [178, 180] and a proportion of invariable sites (I) [3, 61] (Setting 5). For each setting, we simulated

10,000 alignments in RevBayes [75] by randomly drawing parameter values from the prior distribution associated with each parameter (see Table 3.2 for details about the parameters and their prior distributions). Parameter values were drawn separately for each dataset.

Once datasets were simulated, we conducted Bayesian Markov chain Monte Carlo (MCMC) analyses using RevBayes [75] to estimate posterior distributions of tree topologies and model parameter values for each simulated dataset. We then drew samples from these posterior distributions to generate posterior predictive datasets and compared each original dataset to its corresponding posterior predictive distribution using a variety of test statistics [83]. Details of these analyses are provided below.

Table 3.3: Moves used during Markov chain Monte Carlo (MCMC) analyses. Jukes-Cantor analyses used only the first three moves.

Function in RevBayes	Description	Parameter to change	Weight
mvNNI	Nearest-neighbor interchange move	Ψ	num. of taxa
(e.g., 16, 64)			
mvSPR	Subtree prune-and-regraft move	Ψ	num. of taxa x 0.1
(e.g., 1.6, 6.4)			
mvBranchLengthScale	Scaling move on the branch lengths	bl	num. of taxa
(e.g., 16, 64)			
mvBetaSimplex	Scaling move on nucleotide frequencies	π	2.0
mvDirichletSimplex	Scaling move on nucleotide frequencies	π	1.0
mvBetaSimplex	Scaling move on exchangeabilities	er	3.0
mvDirichletSimplex	Scaling move on exchangeabilities	er	1.5
mvScale	Scaling move on shape parameter	α	2.0
mvBetaProbability	Scaling move on proportion of invariable sites	I	2.0

3.5.2 Markov Chain Monte Carlo Analyses

We performed MCMC analyses in RevBayes [75] for each simulated dataset using the same conditions under which they were simulated (see Table 3.1). Prior distributions were the same as those from which parameter values were drawn for simulation (Table 3.2). For all analyses, we estimated the tree topology and branch lengths. For analyses of datasets simulated under Setting 5 with a GTR+Γ+I model, we also estimated the equilibrium base frequencies, the exchangeabilities, the shape parameter of the Γ distribution, and the proportion of invariable sites (I; Table 3.2). Each analysis involved a burn-in phase of 200 iterations, followed by MCMC sampling for 10,000 iterations. The moves used for each parameter, and their associated weights, are provided in Table 3.3. A subset of runs from different conditions were spot checked to ensure that the MCMC settings were sufficient to achieve good convergence of both scalar parameter values and tree topologies. MCMC analyses conducted for use with posterior predictive analyses involving data-based test statistics used two independent replicate analyses and automatic tuning of moves every 200 generations during both the burn-in and sampling phases. Analyses conducted for use with posterior predictive analyses involving inference-based test statistics used a single replicate and only used automatic tuning during the burn-in phase.

3.5.3 Posterior Predictive Analyses and *p*-values

To perform posterior predictive analysis on each of the simulated datasets, we used the P^3 (Phylogenetic Posterior Prediction) workflow implemented in RevBayes [83]. Phylogenetic posterior predictive analyses involve four steps: (1) estimating posterior distributions of phylogenetic trees and model parameters from input data (see above), (2) simulating new (posterior predictive) data using parameter values drawn from the estimated posterior distributions, (3) computing test statistics for both the original and simulated data, and (4) calculating p-values and effect sizes to summarize the (dis)similarity between original and simulated data. Some test statistics, known as inference-based (see below), may depend on characteristics of the inferences drawn from data. To calculate these, an additional step (3a) is necessary that involves running MCMC analyses on each simulated, posterior predictive dataset. For step (2), we simulated 1,001 posterior predictive datasets when using data-based test statistics and 501 posterior predictive datasets when using inference-based test statistics.

The P^3 workflow has a number of test statistics (Tables 3.4 and 3.5) available that summarize characteristics of alignments. Some of these statistics (data-based, Table 3.4) are calculated directly from the alignment itself, while others (inference-based, Table 3.5) are calculated based on characteristics of inferences drawn from the alignment. We used all test statistics currently implemented in P^3 in RevBayes. For any of these statistics, p-values can be used to assess whether the "observedälignment is similar to the posterior predictive alignments [33, 83]. P-values indicate what percentage of posterior predictive test statistic values are more extreme than the observed value.

P-values near 0 or 1 indicate that the observed value falls in a tail of the posterior predictive distribution. Midpoint p-values are particularly useful for discrete test statistics, where ties can be observed between posterior predictive and observed values. In such a case, the midpoint p-value will consider half of the tied posterior predictive values to be more extreme than observed and half to be less extreme than observed. In this study, we specifically focused on the lower, one-tailed, midpoint p-value. All 10,000 simulated datasets were analyzed to characterize the behavior of posterior predictive analyses for data-based test statistics, while 1,000 datasets were analyzed for inference-based test statistics due to their more computationally intensive calculation.

Table 3.4: Descriptions of data-based test statistics.

Test Statistic	Description	Reference
Number of invariant sites	Number of columns in the alignment that show no variation in nucleotide content	[83]
Max invariant block length	The maximum number of consecutive sites with no variation	
Max pairwise difference	The scaled number of mismatches between the pair of sequences with the greatest number of mismatches	[83]
Max variable block length	The maximum number of consecutive sites with variation	
Min pairwise difference	The scaled number of mismatches between the pair of sequences with the fewest number of mismatches	[83]
Mean GC content	GC content averaged across all sequences	[83]
Variance in GC content	Variance in GC content across sequences in an alignment	[83]
Theta	Watterson's measures the genetic diversity in a given population	[174]
Tajima's D	Accounts for how much the variability observed is due to chance	[164]
Tajima's	Average number of pairwise differences across sequences in an alignment	[133, 93]
Multinomial likelihood	Measures the ability of the model to account for different site pattern frequencies	[59]

3.5.4 Effect Sizes

While we have largely focused our attention in this study on the distribution of posterior predictive p-values, because such values have received the most attention in the statistical literature, an alternative measure of absolute model fit is the posterior predictive effect size [PPES; 33, 83]. Complementary to posterior predictive p-values, posterior predictive effect sizes capture the magnitude of differences between observed and expected test statistic values on a broader scale. While posterior predictive tests using two different test statistics for the same dataset may both produce p-values of 0, one observed value may fall just outside the tails of the corresponding posterior predictive distribution, while the other observed value may be very, very far away from its predicted values. Effect sizes differentiate between these two situations, and are calculated as

$$PPES = \frac{|T(y) - M(p(T(y^{rep})|y))|}{\sigma(p(T(y^{rep})|y))} \tag{3.5}$$

Table 3.5: Descriptions of inference-based test statistics, originally described by Brown (2014).

Test Statistic	Description
Mean RF	Mean RF [149] distance between trees sampled from the posterior distribution
Quant 25	25th percentile in the ordered vector of RF distances between trees sampled from the posterior distribution
Quant 50	50th percentile in the ordered vector of RF distances between trees sampled from the posterior distribution
Quant 75	75th percentile in the ordered vector of RF distances between trees sampled from the posterior distribution
Quant 99	99th percentile in the ordered vector of RF distances between trees sampled from the posterior distribution
Quant 999	999th 1000-quantile in the ordered vector of RF distances between trees sampled from the posterior distribution
Entropy	Gain in information about the tree topology provided by the data
Mean TL	Mean length of trees sampled from the posterior distribution
Var TL	Variance in the length of trees sampled from the posterior distribution

where y is the observed dataset, y^{rep} is a posterior predictive dataset, $T(y)$ is a test statistic calculated with y, $p(T(y^{rep})|y)$ is the posterior predictive distribution of T, M is the median of a distribution, and σ is the standard deviation of a distribution. In other words, a posterior predictive effect size is the absolute value of the difference between the observed test statistic value and the median of the posterior predictive distribution of test statistic values, normalized by the posterior predictive distribution's standard deviation. Using the same simulations and analyses that we used to understand the expected behavior of posterior predictive p-values, we also examined the expected distributions of posterior predictive effect sizes.

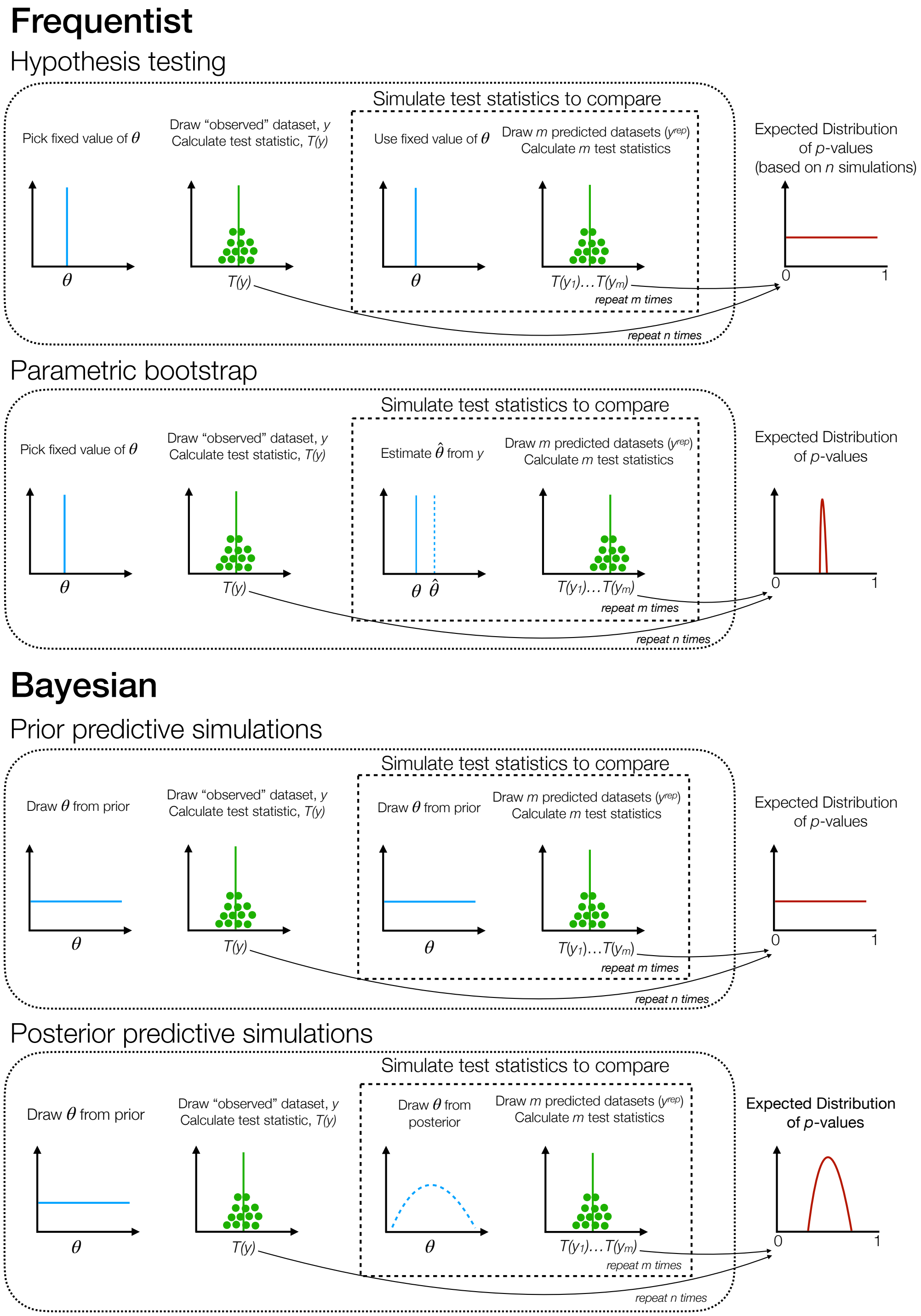

Frequentist
Hypothesis testing
Simulate test statistics to compare
Pick fixed value of θ
Draw "observed" dataset, y
Calculate test statistic, $T(y)$
Use fixed value of θ
Draw m predicted datasets (y^{rep})
Calculate m test statistics
Expected Distribution of p-values (based on n simulations)
θ
$T(y)$
θ
$T(y_1)\ldots T(y_m)$
repeat m times
0
1
repeat n times
Parametric bootstrap
Simulate test statistics to compare
Pick fixed value of θ
Draw "observed" dataset, y
Calculate test statistic, $T(y)$
Estimate $\hat{\theta}$ from y
Draw m predicted datasets (y^{rep})
Calculate m test statistics
Expected Distribution of p-values
θ
$T(y)$
θ $\hat{\theta}$
$T(y_1)\ldots T(y_m)$
repeat m times
0
1
repeat n times
Bayesian
Prior predictive simulations
Simulate test statistics to compare
Draw θ from prior
Draw "observed" dataset, y
Calculate test statistic, $T(y)$
Draw θ from prior
Draw m predicted datasets (y^{rep})
Calculate m test statistics
Expected Distribution of p-values
θ
$T(y)$
θ
$T(y_1)\ldots T(y_m)$
repeat m times
0
1
repeat n times
Posterior predictive simulations
Simulate test statistics to compare
Draw θ from prior
Draw "observed" dataset, y
Calculate test statistic, $T(y)$
Draw θ from posterior
Draw m predicted datasets (y^{rep})
Calculate m test statistics
Expected Distribution of p-values
θ
$T(y)$
θ
$T(y_1)\ldots T(y_m)$
repeat m times
0
1
repeat n times

Figure 3.1: Schematic of workflows to estimate expected distributions for different types of p-values. The depictions of expected distributions are generalizations, intended to highlight important differences among different types of p-values.

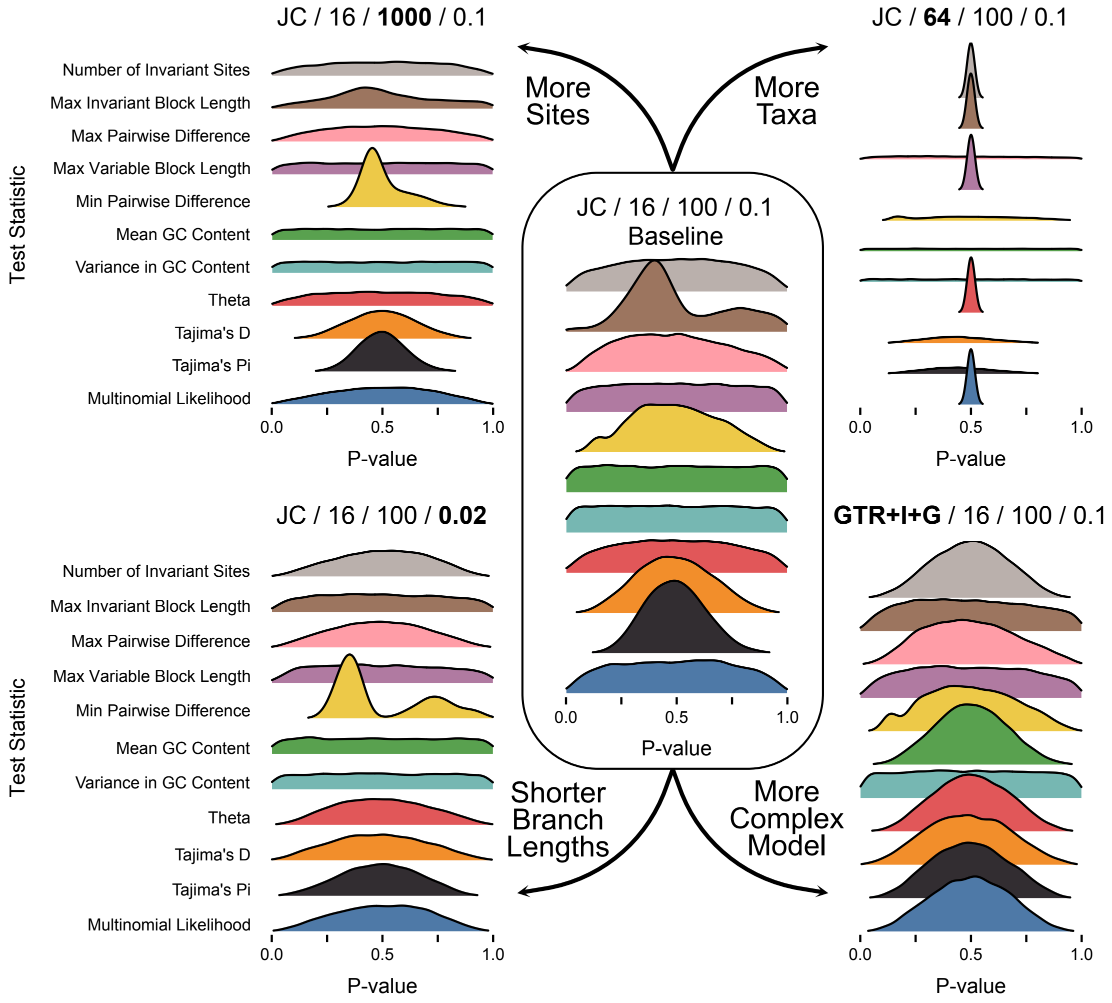


Figure 3.2: Distributions of posterior predictive p-values for data-based test statistics. The conditions for simulation and analysis are shown above each relevant portion of the figure as: Model / Number of Taxa / Number of Sites / Mean Branch Length. Results from the baseline setting (Setting 1 in Table 3.1) are shown in the middle. The other settings modified one condition of the baseline, indicated by the labels next to arrows.

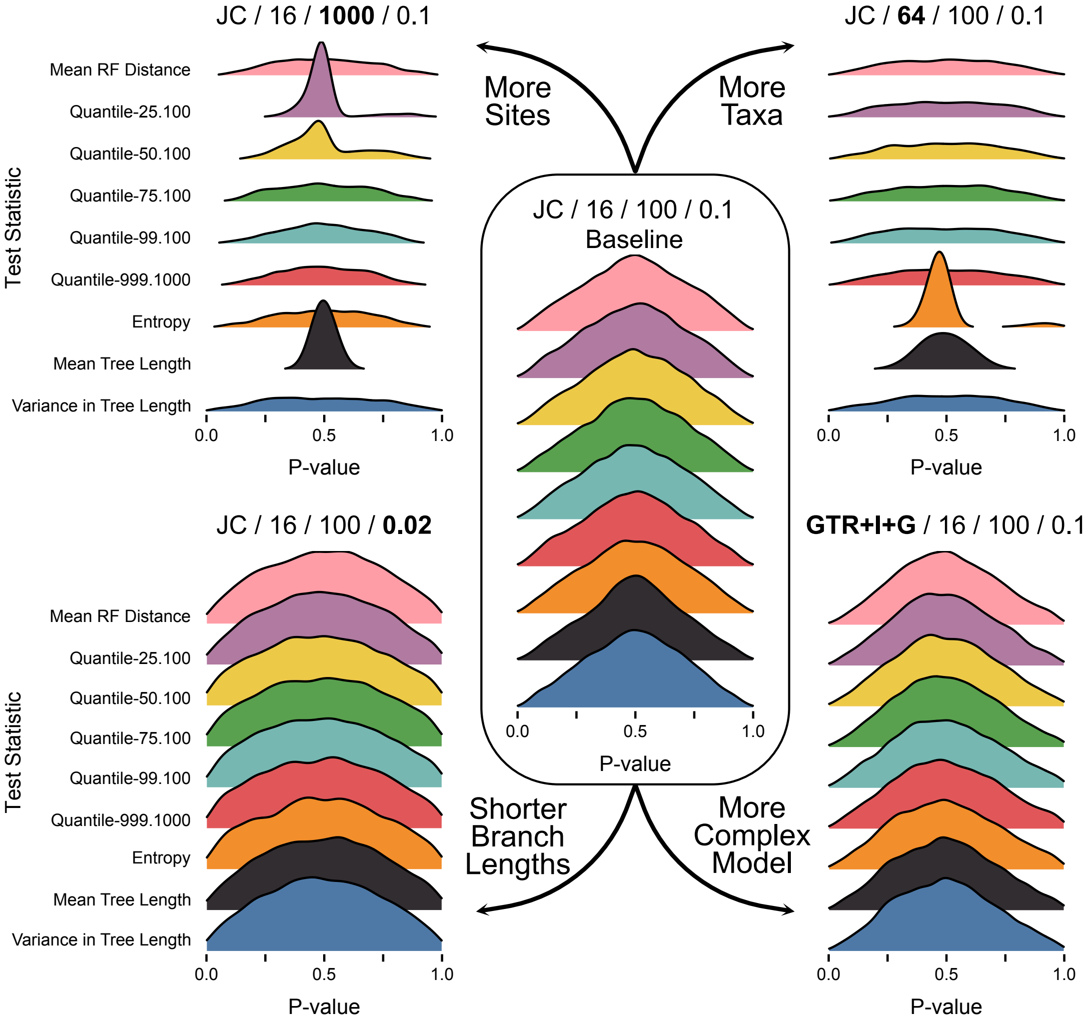


Figure 3.3: Distributions of posterior predictive p-values for inference-based test statistics. The labels and layout are the same as in Fig. 2.

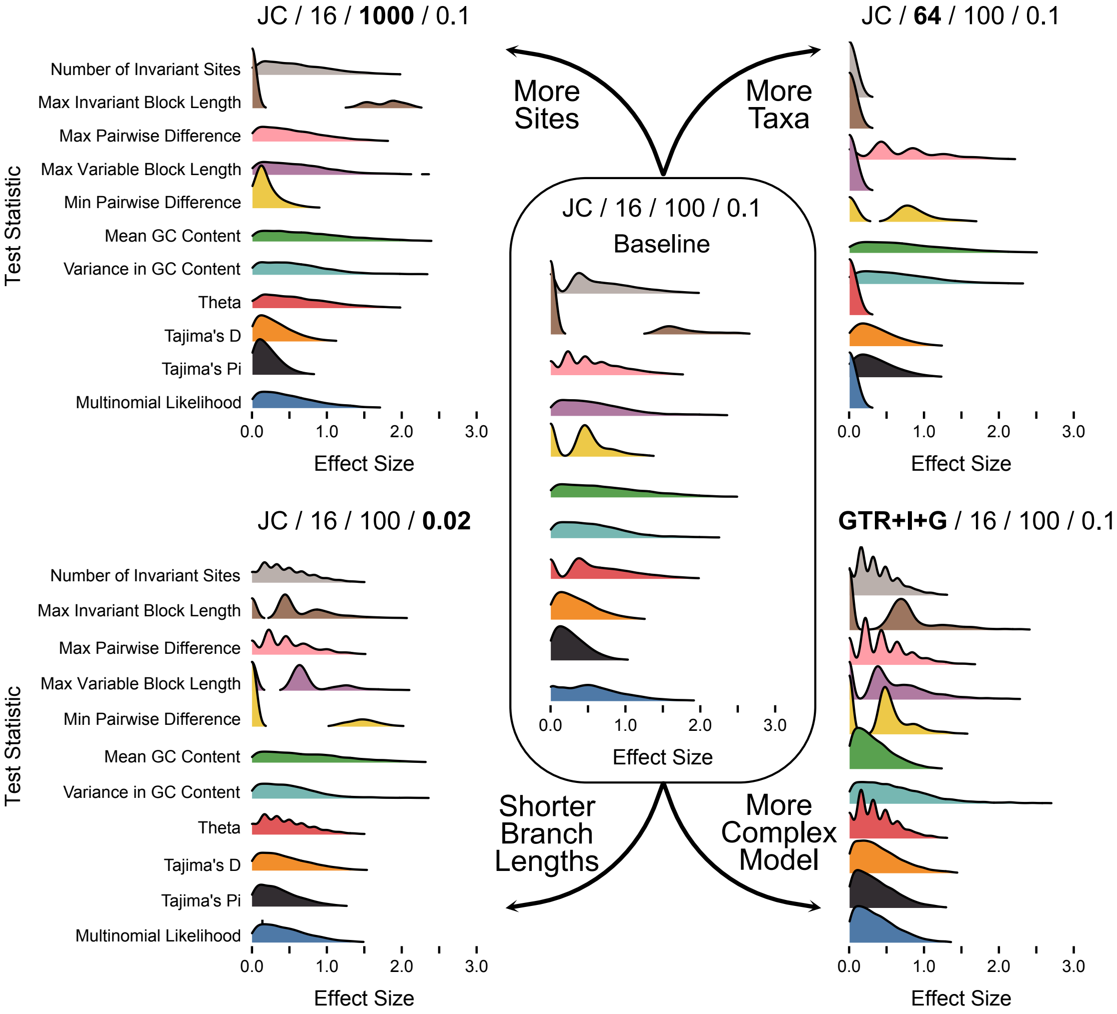


Figure 3.4: Distributions of effect sizes for data-based test statistics. The labels and layout are the same as in Fig. 2.

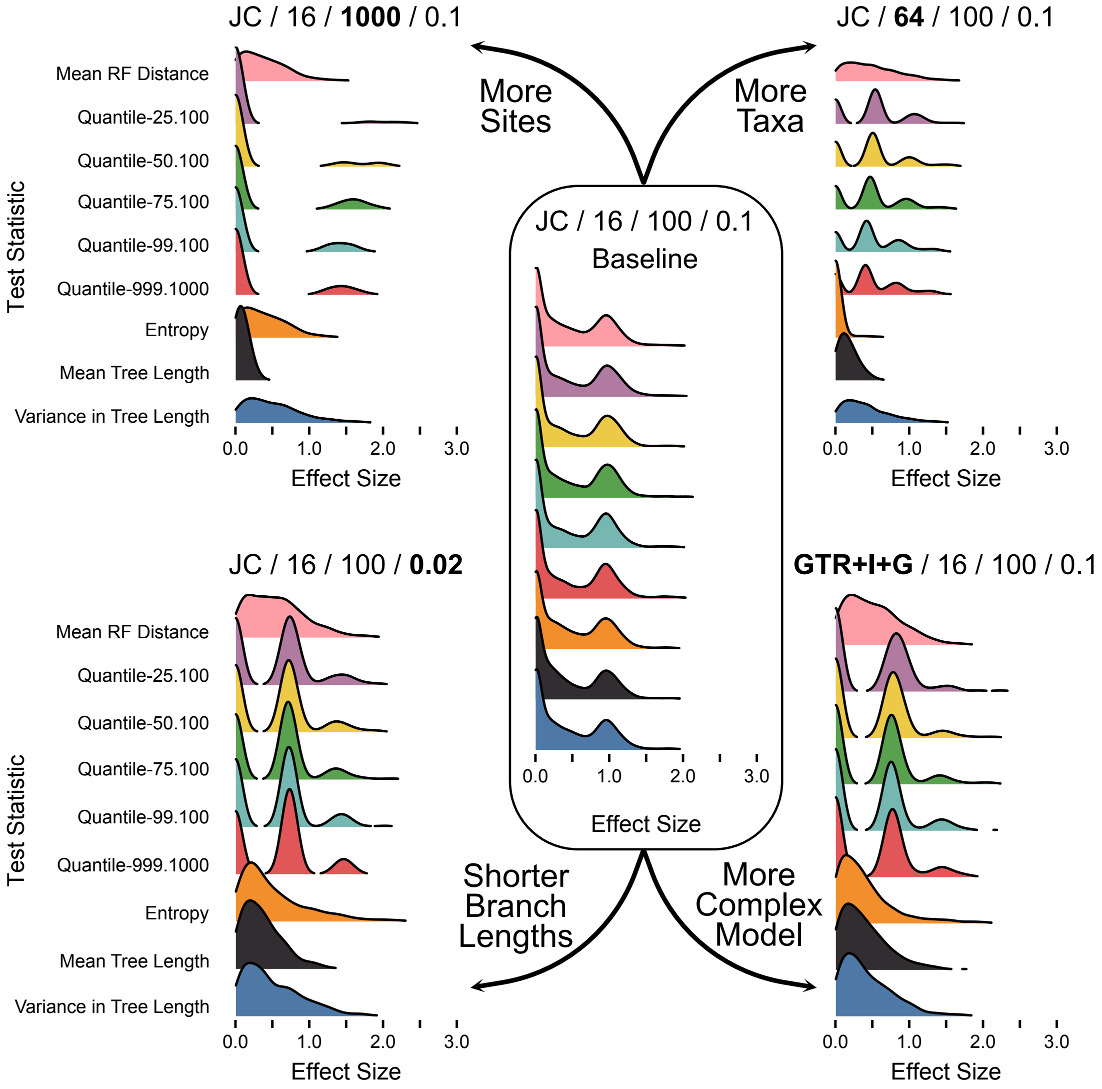


Figure 3.5: Distributions of effect sizes for inference-based test statistics. The labels and layout are the same as in Fig. 2.

Chapter 4

Nucleotide Substitution Model Selection is not Necessary for Bayesian Inference of Phylogeny with Well Behaved Priors

4.1 Abstract

Model selection aims to choose the most adequate model for the statistical analysis at hand. The model must be complex enough to capture the complexity of the data but should be simple enough to not overfit. In phylogenetics, the most common model selection scenario concerns selecting an appropriate substitution and partition model for sequence evolution to infer a phylogenetic tree. Here we explored the impact of substitution model over-parameterization in a Bayesian statistical framework. We performed simulations under the simplest substitution model, the Jukes-Cantor model, and compare posterior estimates of phylogenetic tree topologies and tree length under the true model to the most complex model, the GTR+Γ+I substitution model, including over-splitting the data into additional subsets (*i.e.,* applying partitioned models). We explored four choices of prior distributions: the default substitution model priors of `MrBayes`, `BEAST` and `RevBayes` and a newly devised prior choice (`Tame`). Our results show that Bayesian inference of phylogeny is robust to substitution model over-parameterization but only under our new prior settings. All three default priors introduced biases for the estimated tree length. We conclude that substitution and partition model selection are superfluous steps in Bayesian phylogenetic inference pipelines if well behaved prior distributions are applied.

4.2 Introduction

At the heart of all model-based phylogenetic inferences lies the substitution model. The substitution model defines the rate of substitutions between any pair of states (*e.g.,* between nucleotides) and thus the probabilities of substitutions given a branch length. Many different substitution models have been suggested over time, *e.g.,* the Jukes-Cantor (JC) model [86], the Kimura two parameter (K2P) model [90], the Kimura three parameter (K3P) model [91], the Felsenstein (F81) model [41], the Hasegawa-Kishino-Yano (HKY85) model [68], and the general time reversible (GTR) model [165]. Additionally, phylogenetic substitution models often incorporate different rates among sites [the +Γ model, 177] and/or a proportion of invariable site [the +I model, 2, 62]. With this diversity of substitution models, a researcher is left with the daunting task to choose the "best" substitution model for the specific dataset at hand.

An *under-parameterization* (*i.e.,* oversimplified) substitution model can bias phylogeny estimation [159]. This problem was demonstrated in several applications, *e.g.,* [109], [160], [27], [89]. These studies concluded that phylogenetic inference with different substitution models can result in a significant difference of the tree topology and/or branch lengths, with the more complex models performing better in all cases. The observed biases introduced due to under-parameterized substitution models has led to the development of methods for substitution model selection [140, 79, 139, 28, 88]. It has become common practice to select the best fitting substitution model before estimating a phylogenetic tree. For example, in 2014, the paper describing the software `ModelTest` [140] was among the the 100 most cited papers of all time [170].

There are several problems with the current approach of substitution model selection in phylogenetics pipelines. First, the current approach is circular because the model selection step [*e.g.,* in `ModelTest` and its successors, 140, 139, 28] requires a phylogeny [18]. This phylogeny is often estimated using fast but less accurate models and methods [*e.g.,* using Neighbor-Joining, 141]. Using the wrong phylogeny could lead to biased model selection. Second, the current approach does not incorporate uncertainty in the estimated phylogeny, branch lengths and substitution model parameters [135, 18]. The crucial issue with substitution model selection is that considerable shortcuts are taken because the analysis using a single substitution model can take weeks to months. Performing full inferences and model selection, *e.g.,* computing Bayes factors for each substitution model [73], increases the computation time by several factors even using parallel computations [77]. Full substitution model selection is infeasible and (almost) never applied because of this computational demand.

Is substitution model selection in Bayesian phylogenetic inference a necessary step, or could simply the most complex, *e.g.,* the GTR+Γ+I substitution model, be used? It has been shown that an *under-parameterized* substitution model can bias phylogeny estimation [see 159, for a review] but does an *over-parameterized* (*i.e.,* too complex) substitution model also biases phylogenetic inference? Surprisingly, this question has received rather little attention and only two studies have partially addressed this question [81, 111]. First, [81] studied the posterior probabilities of bipartitions under simple and complex substitution models. They concluded that Bayesian inference is more sensitive to under-parameterization than to over-parameterization with regard to tree topology. Second, [111] specifically studied the impact of model misspecification on Bayesian inference of phylogeny. They show that model over-parameterization does not bias bipartition posterior probabilities and has little to no effect on branch lengths. The observed effect on branch lengths was a decrease in precision. Therefore, [111] conclude with a warning to not assume the most complex model because of the “imprecision that may result from over-parameterization”. Thus, substitution model selection remains common practice in phylogenetic inference.

Furthermore, in most phylogenetic analyses, the sequence alignment is divided into several data subsets, *e.g.,* a multi-locus dataset divided by gene or codon position. Each data subsets can either receive its own substitution model or share the substitution model with another data subset (so-called partition models, [135]). The partition model accommodates process heterogeneity along molecular data and improper data partitioning and application of under-parameterized substitution model can bias phylogenetic inferences [23]. The number of possible partition models to choose from is significantly larger than the number of available substitutions models. If selecting the best substitution model for a single locus is already burdensome and computationally extremely demanding, then selecting the best partition model is clearly infeasible without even more drastic shortcuts. Nevertheless, several methods have been developed for partition model selection [*e.g.,* 99, 100] and are commonly applied in phylogenetic inference pipelines. Until today, there has been no study to evaluate if over-parameterization, *i.e.,* assuming a separate GTR+Γ+I substitution model per data subset, biases phylogenetic inference.

In this study, we will investigate the effect of model over-parameterization on Bayesian phylogenetic analysis. Specifically, we focus on the question if substitution model selection is a necessary step for Bayesian phylogenetic inference. Can we simply use the most complex substitution model and partition model and thus avoid the danger of under-parametrized models? Here, we explore this question using simulations. We simulated data under the simplest model and inferred the phylogenies under (a) the true model, (b) an over-parameterized substitution model, and (c) an over-parameterized partition model. Moreover, we tested different choices of prior distributions for the over-parametrized substitution model. The advantage of using simulations over previous studies using empirical data [*e.g.,* 1] is that we know the true parameters (*i.e.,* the true phylogeny and branch lengths) and the true model. Therefore, we are able to assess if over-parametrization biases our results or leads to less precise estimates (*i.e.,* higher uncertainty and larger credible intervals). We assessed biases by comparing bipartition posterior probabilities and tree lengths between analyses under the true substitution model and an over-parametrized model.

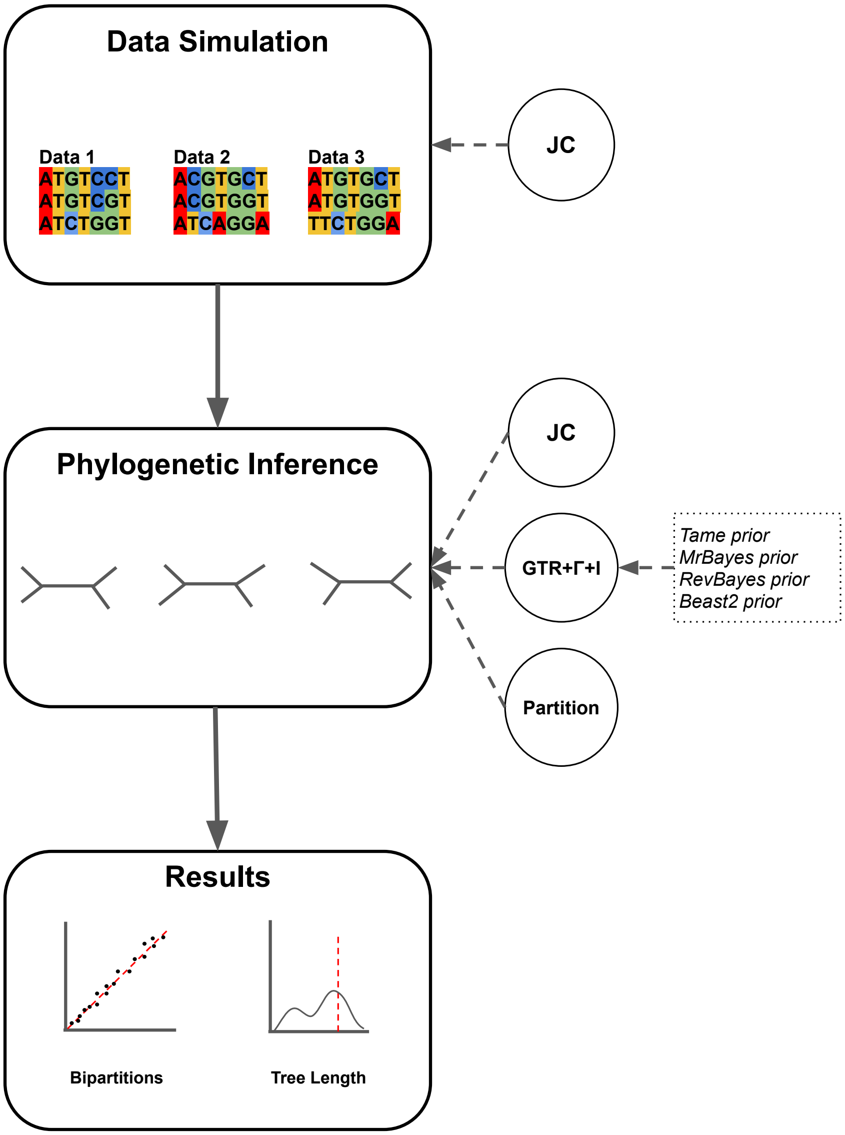


Figure 4.1: Summary of the simulation and analyses used in this study. The first step contains the simulation of data matrices under the Jukes-Cantor (JC) substitution model. The next step contains the phylogenetic inference under the different settings (GTR+Γ+I with four different prior settings and the partition model). The comparisons between the true and over-parameterized models are summarized in the results.

4.3 Methods

We performed a simulation study to understand how the estimated tree topology and tree length are affected under an over-parameterized substitution model (within the GTR family of nested substitution models). Our approach is depicted in Figure 4.1. The data sets were simulated under the simplest substitution model, the Jukes-Cantor substitution model (JC). We over-parameterized the substitution model by using the most parameter rich commonly used substitution model, the GTR+Γ+I substitution model, to perform the phylogenetic inference. The simulation with the simplest substitution model in combination with the inference with the most complex comprise the most extreme case of over-parameterization of common substitution models, and therefore the most conservative scenario to evaluate the effects of substitution model over-parameterization. If we find no impact of substitution model over-parameterization for this extreme scenario, then there is no impact of substitution model over-parameterization for less extreme cases.

We varied the prior probability distributions for the over-parameterized model according to the default settings of three commonly used Bayesian phylogenetic software (`MrBayes`, `BEAST` and `RevBayes`). These default prior distributions for the Bayesian phylogenetic software were extracted for: a) `MrBayes` v3.2 [151]; b) the `RevBayes` protocol described in [73]; c) `BEAST` using BEAUTi [17]. The difference in default prior distributions among these popular software packages reflects the uncertainty about good prior choices in the field, and our choice of these three example does not represent a favor for or against any of these choices. We included another prior distribution (called “`Tame`” in this manuscript). Additionally, we performed inference under an over-partitioned scheme for the novel prior setting where each data subset received its own separate GTR+Γ+I substitution model. Next, we compared posterior probabilities of bipartitions and tree length credible intervals between the models used for inference. The following sections explain in more detail the methods used in this study.

4.3.1 Simulation Settings for the Datasets

We simulated data matrices by first drawing all parameters from their prior distribution. The substitution model was set to the JC model of sequence evolution [86]. The JC substitution model is the simplest model within the GTR family of nested models and has no free substitution model parameters (all base frequencies are fixed to $\frac{1}{4}$ and all relative exchangeability rates are fixed to $\frac{1}{6}$). Therefore, the prior distributions for the simulations were the tree topology prior and the branch lengths prior. We assumed a uniform distribution on tree topology, *i.e.,* each topology had equal prior probability. We chose two different tree sizes, 16 and 64 taxa, to explore the impact of tree size. The prior distribution for the branch lengths was an exponential distribution with either a mean of 0.1 or 0.02, which impacts the total number of substitutions expected along the phylogeny and therefore the informativeness of the data sets (Figure S1). We defined the number of sites to be 100 or 1000 to explore the impact of the amount of data. In summary, we simulated data sets under all possible combinations of mean branch length, number of taxa and number of sites, yielding eight different simulation scenarios which are displayed in Table 4.1. We simulated 1000 replicates for the data set with 16 taxa and 500 replicates for the data sets with 64 taxa. The simulations were performed using `RevBayes` [74] and all scripts are available at https://github.com/lfabreti/SM-over-parameterization.

Table 4.1: The different settings for the simulation of data matrices under the Jukes-Cantor substitution model. The first column shows the index of the simulation setting. The second column is the number of taxa, followed by the number of sites and the mean branch length. The gray cells highlight the changes in the settings for each simulation setting.

Simulation setting	Number of taxa	Number of sites	Mean branch length
1	16	100	0.1
2	16	100	0.02
3	16	1000	0.1
4	16	1000	0.02
5	64	100	0.1
6	64	100	0.02
7	64	1000	0.1
8	64	1000	0.02

4.3.2 Prior Distributions on the Substitution Model Parameters

In Bayesian inference, the prior probability distribution defines the researcher's belief about the parameter quantity before the data are taken into account. All parameters and hyper-parameters from a model need to be assigned to a prior probability distribution. The best approach to define a prior probability distribution for a given parameter is an open debate in Bayesian inference [114, 127, 52, 112, 9]. In some situations in phylogenetics, the researcher has reliable information to make strong assumptions about the prior distribution for a given parameter. For example, informative prior distribution are commonly used in divergence time estimation using node calibrations where fossil information is used to define minimum and maximum bounds on the age of a given node [137, 172]. If no reliable information about parameter values is available, then the prior should be designed to have little effect on the estimated parameters [186, 4]. Next, we will discuss the default priors for the GTR+Γ+I model adopted by three commonly used Bayesian phylogenetic software (`MrBayes`, `BEAST` and `RevBayes`) and our proposed prior scheme (`Tame`).

The GTR model has four equilibrium base frequencies and six rates of changes between bases (exchangeability rates). The among site rate variation model (ASRV) and the invariant sites model have each one parameter, the shape parameter α and the probability of a site being invariant, respectively. All default prior distributions assign equal probabilities for all four base frequencies (Table 4.2). The main differences in the prior schemes assessed here lie in the exchangeability rates and the α parameter of the among site rate variation model. To aid grasping the impact of the different prior distribution choices, we provide figures depicting the behavior of each induced parameter given the prior settings (Figures 4.2 and 4.3).

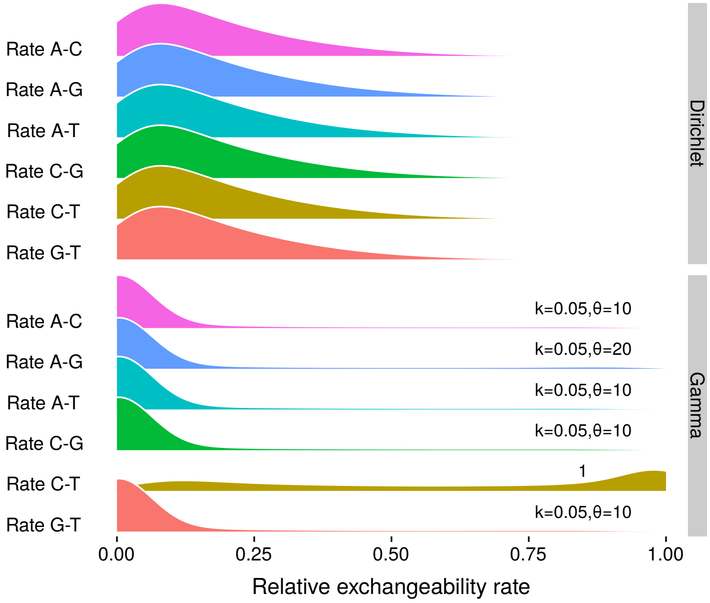


Figure 4.2: The prior distribution of the exchangeability rates. Here we show the two commonly used prior distributions on exchangeability rates; the Dirichlet distribution and the gamma distribution. The gamma distribution has a shape 0.05 and scale parameter 10 for the rates A↔C, A↔T, C↔G and G↔T; the rate A↔G has a different scale of 20; the rate C↔T is set to 1 for normalizing the rates. These densities are produced by simulating 1×10^6 samples from the corresponding distributions.

We compared the profile of two different prior distributions on the exchangeability rates, a flat Dirichlet(1,1,1,1,1,1) distribution and a gamma(k=0.05, θ=10 or 20) distribution. The prior schemes for the `Tame`, `MrBayes` and `RevBayes` use the former, while `BEAST` assumes the latter. Note that the prior settings `Tame`, `MrBayes` and `RevBayes` require that the sum of all rates equals to 1.0 while `BEAST` rescales the rates internally and fixes the rate between C↔T to 1.0. Figure 4.2 shows the distribution for the six exchangeability rates for each prior assumption. The Dirichlet prior distribution results in an equal distribution with a mean of ≈ 0.16 for all six rates, which is expected since the concentration parameter is the same for all categories. The gamma prior distribution used in `BEAST` yields four equal distributions with a mean of ≈ 0.07 for the rates A↔C, A↔T, C↔G and G↔T. The induced prior distribution for the rate A↔G is slightly larger due to the scale parameter θ=20, which results in a mean of ≈ 0.1. We observed the largest difference for the relative rate C↔T with a mean of ≈ 0.6 and the values are concentrated on higher rates. Note that the rate C↔T is originally set to 1.0 and used to normalize the exchangeability rates. All other unscaled exchangeability rates have a prior mean of 1.0 (for A↔G) and 0.5 (for all other rates) but an induced relative mean which is clearly different from $\frac{1}{6}$.

The prior distributions for the shape parameter, α, of the gamma distribution used to model among site rate variation, differs considerably among standard Bayesian phylogenetic inference software (Table 4.2). `MrBayes` v3.2 uses a uniform(0,200) distribution as a prior distribution for α, which has a mean of 100. Note that more recent versions of `MrBayes`, starting with version 3.2.2, assume an exponential prior distribution. The `RevBayes` protocol establishes a biologically motivated prior distribution as a lognormal distribution with median 1.5 and standard deviation 0.587405 [73]. This prior distribution was designed to specify a 95% prior distribution which ranges from a 3-fold to 100-fold difference between the lowest and highest rate categories [73]. `BEAST` uses by default an exponential(λ=1) prior distribution for α with a mean of 1.0.

The induced prior distributions of the four rate categories for each prior distribution on α are shown in Figure 4.3. One might expect that small values for α, or even converging towards 0.0, result in no rate variation. However, the contrary is true. Smaller values of α result in more distinct distributions for the four discretized gamma quartiles, which translates into more rate variation across sites. Both the `RevBayes` and `BEAST` prior settings assign more prior probability to smaller values for α resulting in larger variation between the rate categories. The `MrBayes` prior setting results in more uniformity among the categories, but still expecting some rate variation *a priori*. We propose a prior that results in less *a priori* expected rate variation between the rate categories, namely a uniform(0, 1×10^8) distribution (Figure 4.3 upper panel). This exploration shows the usefulness to examine the induced prior distribution of the model parameters.

4.3.3 Phylogenetic Inference Settings

Phylogenetic inference was performed in a Bayesian Markov chain Monte Carlo (MCMC) framework implemented in `RevBayes` and MCMC convergence was assessed using the `R` [143] package `Convenience` [38]. Each simulated dataset was analyzed under two substitution models: either JC or GTR+Γ+I. The JC substitution model represents the true model whereas the GTR+Γ+I represents the over-specified substitution model. We applied four different prior schemes to the GTR+Γ+I, as described in Table 4.2. The total number of replicated inferences per simulated condition was 1000 for the data with 16 taxa and 500 for the data with 64 taxa. Additionally, we used a partition model, for the data sets with 1000 sites, with two equal-sized data subsets evolving independently under a GTR+Γ+I with the `Tame` prior scheme. For these partition models we only used the simulated data with 1000 because splitting 100 sites results into unrealistically small data subsets. In this case, we performed 500 inference replicates for the data with 16 taxa and 300 for the data with 64 taxa. We used two replicated MCMC runs for each inference and samples were taken at every 20th iteration. The criteria for convergence assessment were the default from `Convenience` as described in [38]. This strict convergence assessment turned out very use-

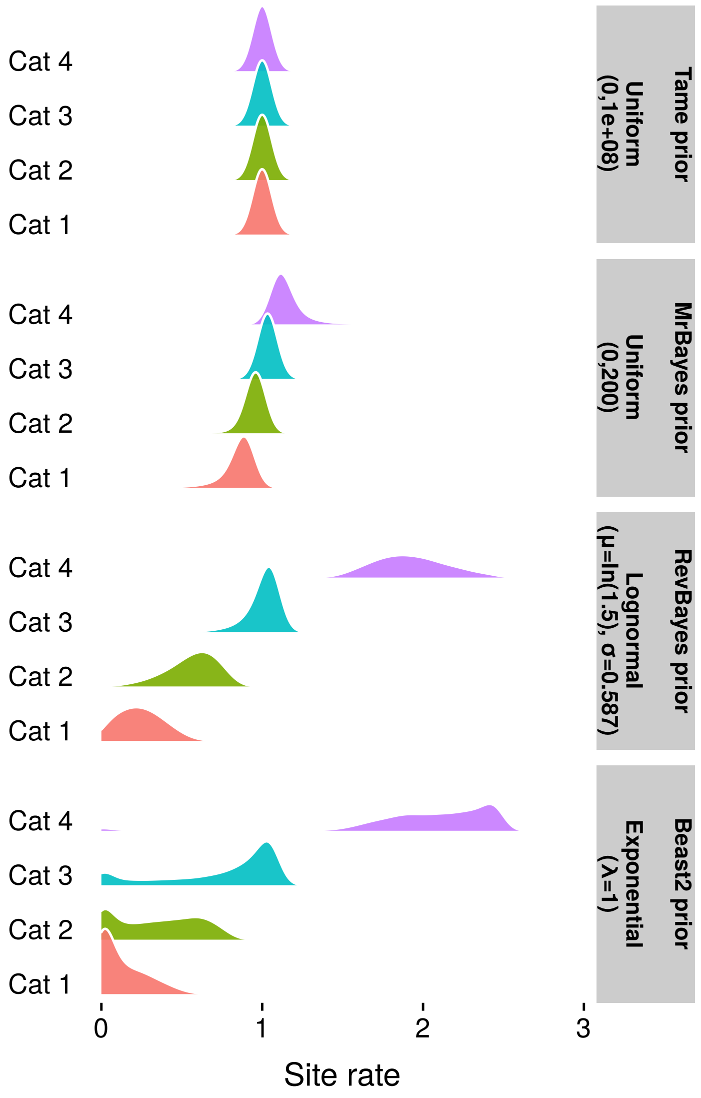


Figure 4.3: The induced prior distributions of the four rate categories from the among site rate variation model for the different α priors. The first panel draws α from a uniform distribution from 0 to 1×10^8 and it is the proposed prior distribution in this study. The second panel uses a uniform form 0 to 200 as the prior on α, which is implemented in `MrBayes`. The third panel shows the default behavior for `RevBayes`, where α is drawn from a lognormal distribution with μ=ln(1.5) and σ=0.587405. The last panel shows the behavior of the four gamma categories when the prior on α is set to an Exponential distribution with λ=1, as it is done in `BEAST`.

ful as several runs showed outlier behavior which could be attributed to non-convergence. The total number of MCMC iterations varied based on the convergence status. We started with 100,000 iterations and increased this value for analyses that did not converge. The maximum number of iterations used was 400,000. The moves during the MCMC followed the scheme on Table 4.3. The moves on the tree parameter varied according to the size of the dataset. For a dataset with 16 taxa, each iteration had 76 moves, whereas a dataset with 64 taxa had 244 moves per iteration. The inference with the `BEAST` prior setting yielded a different move scheme for the exchangeability rates because each parameter had its independent prior distribution instead of a compound prior distribution. In this case, each of the estimated rates (A↔C, A↔G, A↔T, C↔G and G↔T) was assigned with a scale move with weight two.

4.3.4 Evaluation of bias and uncertainty in estimated parameters

The two key parameters of interest for most phylogenetic studies are the tree topology and the branch lengths. Therefore, we focused our evaluation of the effect of over-parameterization on these two parameters. The tree topology was translated into its bipartitions, which are subsets of the full tree. We analyzed two outcomes of the MCMC output: (1) the posterior probability of bipartitions and (2) the 95% credible interval for the tree length. These estimates were compared between the inference under the true model (JC) and the over-parameterized models (GTR+Γ+I). If over-parameterization is not problematic, then the estimates under the GTR+Γ+I model will not deviate from the estimates under the JC substitution model.

The posterior probability of any given bipartition was compared to whether the bipartition was true, *i.e.,* present in the true tree. Thus, we obtained the frequency of a bipartition being true given its posterior probability. For a more stable computation of the frequencies, we binned bipartition into 20 equal-sized bins for posterior probabilities between 0.0 and 1.0, *e.g.,* the first bin for bipartitions with posterior probabilities $0.0 \leq PP < 0.05$. The expected behavior is that the overall frequency over all replicates of a bipartition being true is equal to the posterior probability [81]. For example, if we observe thousands of bipartitions with probability ξ, *e.g.,* 0.2, then we expect ξ percent of these bipartitions to be true, *e.g.,* 20% of the bipartitions. We evaluated the behavior of the posterior probabilities vs. frequency of being true for all five inference scenarios.

The 95% credible interval was used as a measure of the precision of the estimated tree length. Larger credible intervals imply more variance in the estimated tree length. The reference for the size of the credible interval was the inference under the true model, *i.e.,* the model used for simulating the data (JC). We compared the size of the credible interval for the tree length between the analysis under JC and GTR+Γ+I, with the four different prior schemes. Additionally, we explored further the posterior probability distributions of the tree length across the inference settings by examining one example dataset for each simulated scenario. If over-parameterization is not a problem, then the estimated 95%

credible interval between the inferences under the JC substitution model (the true model) and the GTR+Γ+I substitution model (the over-parameterized model) should match and the estimates should be seen on the 1:1 line. Conversely, we would expect that if over-parameterization adds uncertainty in our estimates, then we should obtain larger 95% credible intervals for the analyses under the GTR+Γ+I substitution model.

4.4 Results

4.4.1 Accuracy of Posterior Probability of Bipartitions

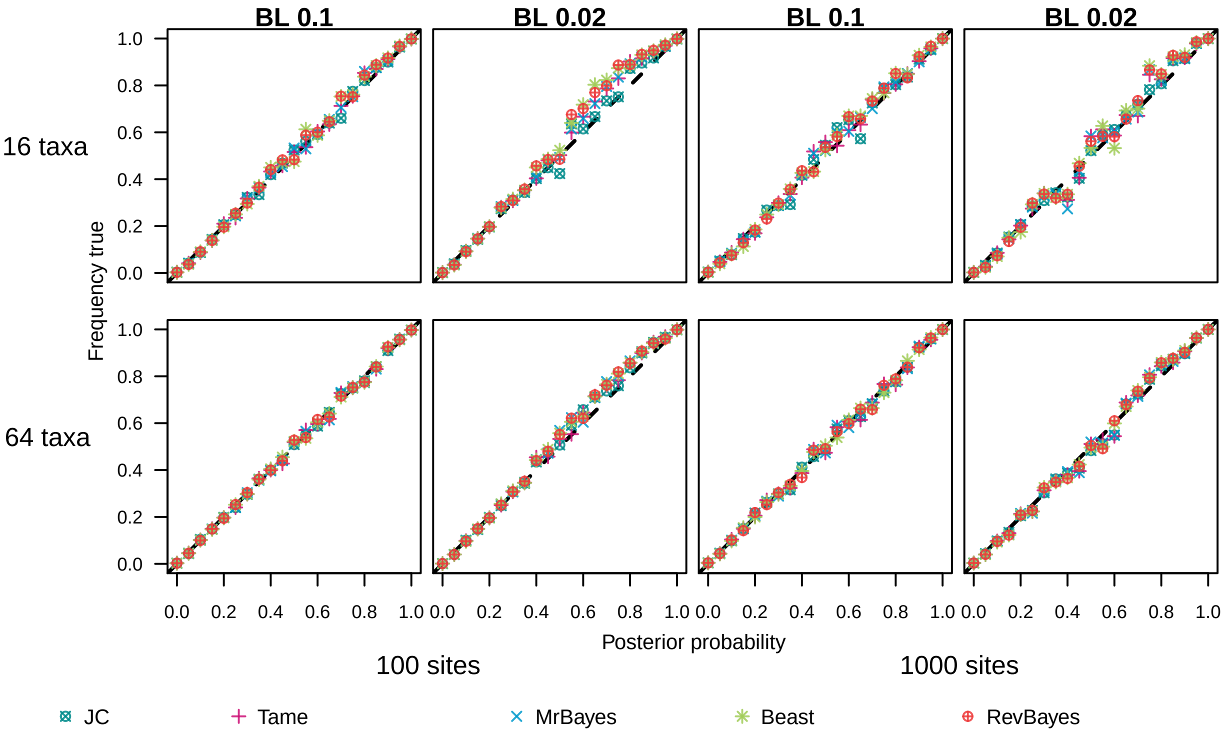


Figure 4.4: The relationship between posterior probability of bipartitions and the bipartition was correct. The expected behavior is that, on average, the posterior probability and the frequency how often a bipartition with this posterior probability is indeed correct are exactly correlated following the 1:1 line. The different panels show each of the 8 settings in which the data were simulated. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa. The symbols represent the different models and prior settings used for the inference (Table 4.2). All models, including the over-parameterized models, produced statistically consistent estimates of the bipartition posterior probabilities.

We observed no major deviation from the expected behavior under all conditions (Figure 4.4, the expected behavior was that the bipartitions fall on the 1:1 line). The minor variation observed in Figure 4.4 is due to the intrinsic randomness of the bipartition posterior probability estimates obtained from the MCMC samples and is also observed for the inference under the true model (JC). Adding more replicates would improve the fit to the diagonal line, but the underlying behavior is already observable with the current amount of replicates, which took several months on our local High Performance Cluster to complete. We note that estimated bipartition probabilities for the simulations with 64 taxa are more accurate, which is due to the larger number of bipartitions in each dataset.

Our results agree with previous observations by [81] and [111]. These two previous studies also showed that substitution model over-parametrization has no effect on estimated bipartition posterior probabilities. Our study clearly corroborates this finding; even a severe over-parameterization of the substitution model does not impact estimated bipartition posterior probabilities. Furthermore, we observed no difference in the estimated bipartition posterior probabilities based on the prior distribution setting used. We also did not observe any impact of the alignment length, number of taxa and branch length prior on the accuracy of the estimated bipartition posterior probabilities. This indicates that, at least for our simulations, the choice among common default prior distributions on the substitution model parameters does not impact the accuracy of bipartition posterior probabilities.

Similarly, we observed no biases in the posterior probabilities of bipartitions (Figure 4.5) for the inference under the partition model. Figure 4.5 shows the same variation around the diagonal line as Figure 4.4 due to the intrinsic stochasticity of the MCMC samples and is also observed for the inference under the true model (JC). This effect is larger for the simulated data sets with 16 taxa, because these simulations have less bipartitions in each replicate. These results for the over-splitted partition model are not surprising as we have seen in Figure 4.4 that even very small datasets of only 100 sites are not impacted by substitution model over-parameterization. Therefore, it is logical that over-splitting and over-parameterization of partition models is not a problem if the size of each data subset is not too small. We conclude that over-parameterization of the substitution and partition models does not affect the posterior probability of bipartitions, and therefore, the tree topology.

4.4.2 Accuracy of Estimated Tree Length

We observed an impact of substitution model over-parametrization on the width of the credible interval of the tree length (Figure 4.6). The 95% credible interval of the tree length was very similar between JC and GTR+Γ+I for all simulated matrices with 1000 sites (Figure 4.6). However, with less data (100 sites), the choice of prior distribution had an observable impact on the estimated tree length and the resulting 95% credible interval (Figures S2-S5). Only our newly proposed `Tame` prior settings produced unbiased estimates of the 95% credible interval.

We noticed the largest deviation between the true model (JC) and the over-parameterized

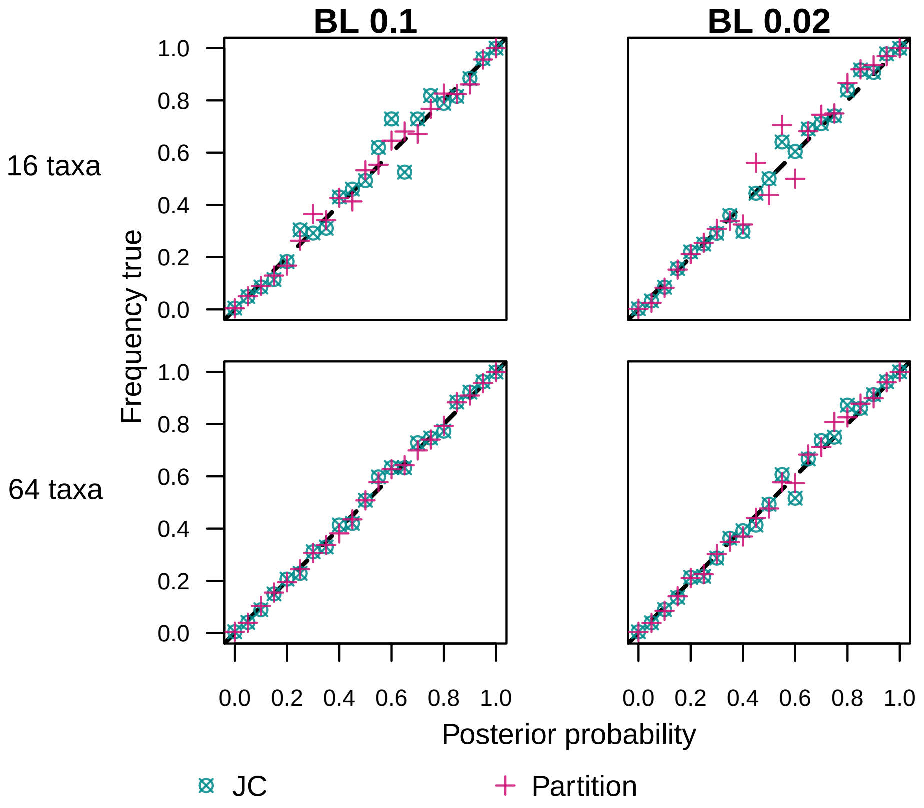


Figure 4.5: The relationship between posterior probability of bipartitions and the bipartition being correct under the partition model. The expected behavior is that the posterior probabilities of the bipartitions and the frequency that a bipartition with this posterior probability was correct follows the 1:1 line. The different panels show each of the 4 settings in which the data was simulated. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. For the over-partitioned model, we only used data sets with 1000 sites. The symbols represent the models for the inference, JC or the over-partitioned GTR+Γ+I model (using the `Tame` prior settings for both data subsets).

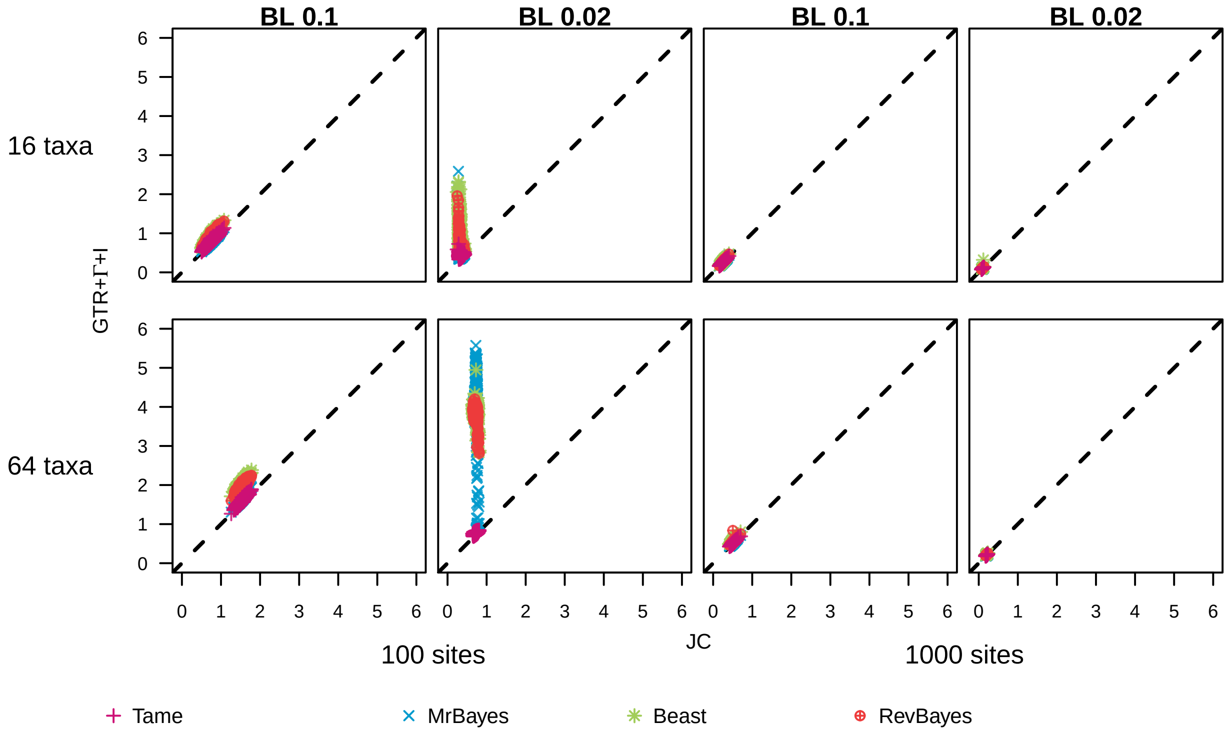


Figure 4.6: The width of the 95% credible interval for the tree length for the inference with Jukes-Cantor (JC) against the inference with GTR+Γ+I. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-parametrized substitution model. If over-parametrization has no impact, then all 95% credible interval sizes would follow the diagonal line (dashed line). The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa. The symbols represent the different prior settings for the inference under GTR+Γ+I (Table 4.2). Separate plots for each model are shown in Figures S2-5.

models (GTR+Γ+I) when we used a mean of 0.02 for the branch lengths in our simulated trees. The estimated 95% credible interval was larger for the over-parameterized models. The trees simulated with a branch length prior with mean 0.02 had shorter branches and the prior distribution on branch lengths was further away from the true values (Table 4.1 and 4.2, Figure 4.6 and 4.7). The trees simulated with a branch length prior with mean 0.1 had a matching prior distribution in the inference. Thus, we observed some interaction between the branch length prior and the prior distribution on the substitution model parameters.

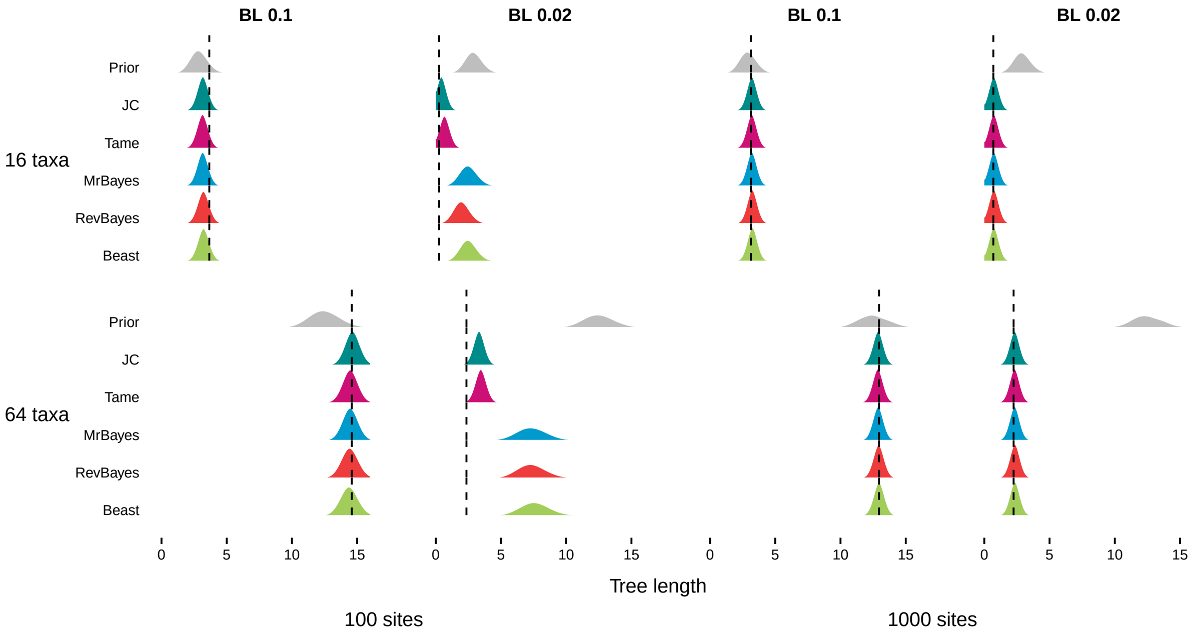


Figure 4.7: Comparison between prior and posterior distributions for the tree length. The first row corresponds to a simulated tree with 16 taxa, while the second row corresponds to a simulated tree with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa. The posterior distributions represented correspond to the inference performed under the true model (JC) and the over-parameterized model (GTR+Γ +I) with the 4 different prior settings (Table 4.2).

We showed one example data set for each simulated scenario to demonstrate more clearly the impact of substitution model over-parameterization on the estimated tree length (Figure 4.7). We observed that the posterior distribution inferred under the GTR+Γ+I substitution model with the `Tame` prior settings matches the posterior distribution inferred under the JC substitution model (the true model) for all scenarios. We observed the most extreme deviation between the posterior distribution inferred under the JC substitution model (the true model) and the inferred posterior distribution of the tree length under the `MrBayes`, `RevBayes` and `BEAST` prior scheme for 64 taxa, 100 sites and mean branch length 0.02. In this scenario, the posterior distribution of the over-parameterized substitution model is widest and shifted in location. Thus, for the MrBayes, RevBayes and BEAST prior schemes we observed biases and more uncertainty in the estimated parameters.

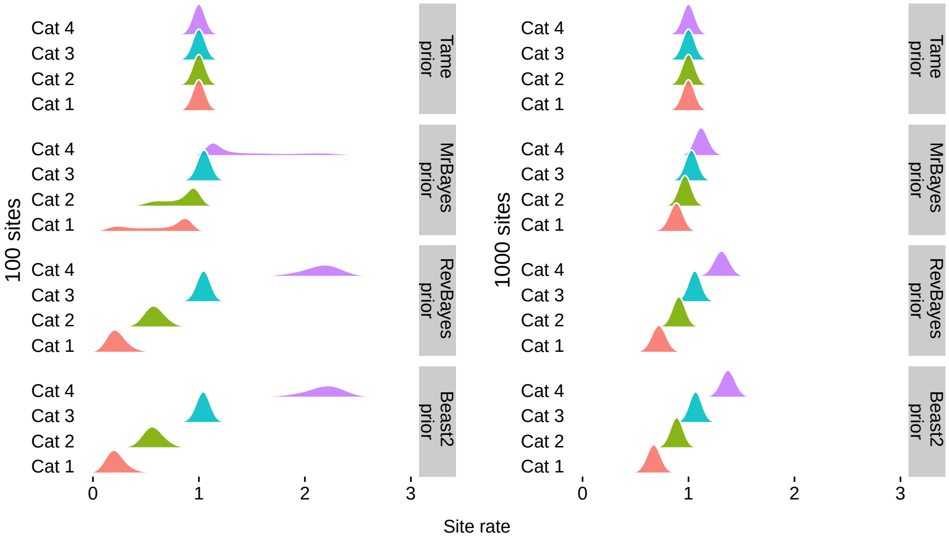


Figure 4.8: Posterior distributions for the rate categories for the among site rate variation model. The posterior distributions correspond to the inference performed under the overparameterized model (GTR+Γ +I) with the four different prior settings (Table 4.2). The left panel shows the posterior distributions for one replicate data with 64 taxa, 100 sites and mean branch length 0.02, whereas the right panel the replicate data has 1000 sites. The posterior distributions are more sensitive to the prior (Figure 4.3) when the dataset was smaller.

We plotted the estimated posterior distribution of the rate categories for the among site rate variation model to elucidate the problem of overestimated posterior distribution of the tree length and interaction of parameters (Figure 4.8). Note that under the true model, all four rate categories should be equal to 1.0. The MrBayes, RevBayes and BEAST prior settings yielded four gamma quartiles with slight (MrBayes) to large (RevBayes and BEAST) differences *a posteriori* in relative site rates (Figure 4.8). When the sites are estimated to fall into the lower rate category, then longer branches are required to obtain the same amount substitutions. This results in larger branch lengths, as seen in Figures 4.6 and 4.7.

Additionally, we tested all possible combinations of the JC and GTR substitution models with the ASRV model and/or the invariant sites model to further evaluate the effect of the inappropriate prior distribution on α (Figures S8-S11). For each simulation setting (Table 4.1) we randomly selected one example simulated dataset and then performed the inference under the following models: JC+I, JC+Γ, JC+Γ+I, GTR, GTR+I, GTR+Γ, GTR+Γ+I. The results show that the 95% credible interval for the tree length was overestimated only when we added the ASRV model with the MrBayes, RevBayes and BEAST prior settings. This result further corroborates that the bias on the tree length is exclusively linked with the prior distribution on α of the ASRV model.

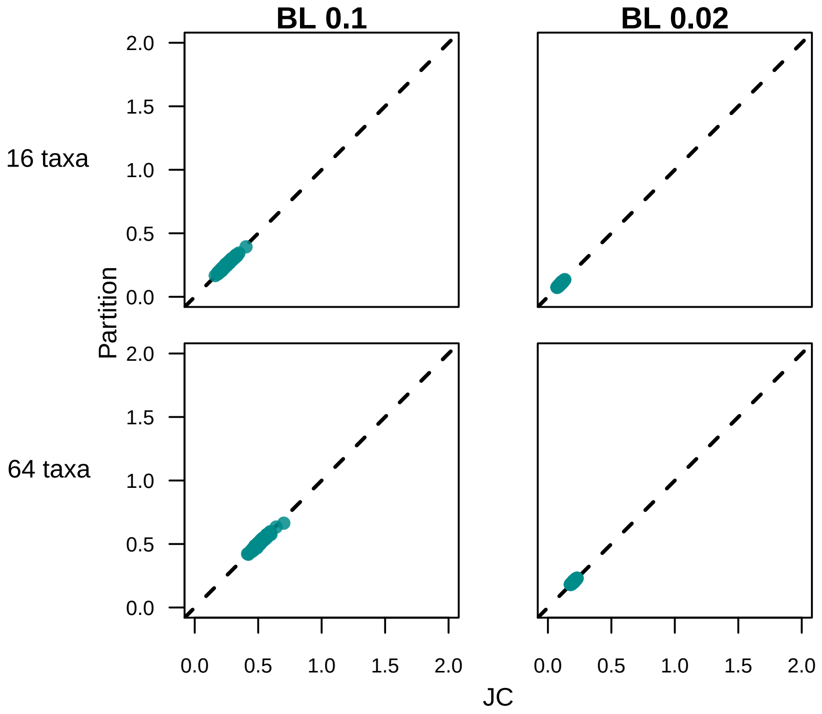


Figure 4.9: The width of the 95% credible interval for the tree length for the inference with Jukes-Cantor (JC) against the inference with the over-splitted GTR+Γ+I partitioned model. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. For the over-partitioned model, we only used data sets with 1000 sites because splitting 100 sites results into unrealistically small data subsets. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-partioned model. If over-parametrization has no impact, then all 95% credible interval sizes would follow the diagonal line (dashed line).

Finally, we observed no biases in the accuracy of the tree length (Figure 4.9) for the inference under the partition model. Recall that we used the `Tame` prior settings for each data subset in our partition model, and used only datasets with 1000 sites equally divided into two subsets of each 500 sites. Since our results using the `Tame` prior on very small datasets of 100 sites showed no impact of substitution model over-parameterization (Figure 4.9), it is expected that the over-splitted and over-parameterized substitution model produces robust estimates of the tree length. In conclusion, Figures 4.6 and 4.9 demonstrate that Bayesian inference of phylogeny is robust to substitution and partition model over-parameterization; however, only if well behaved prior distributions are chosen.

4.5 Discussion and Conclusions

The main purpose of selecting a substitution model is to avoid model misspecification. Assuming an under-parameterized substitution model has been shown to bias phylogenetic inference [159]. However, the question whether over-parameterization of substitution models biases phylogenetic inferences has received much less attention. Under- and over-parameterization of substitution models could be avoided if we would know the true substitution model. Since we do not know the true substitution model, it is common practice to perform substitution model selection before estimating a phylogeny. Common substitution model selection approaches, *e.g.,* `ModelTest` [140], `jModelTest` [139] and `jModelTest` 2 [28], employ shortcuts (*e.g.,* do not take uncertainty in parameter estimate into account and optimize only some parameters) and are philosophically questionable (*e.g.,* they require a phylogeny to perform the model selection step). In this manuscript we argue that substitution model selection is not necessary and can be avoided if the most complex substitution model (*e.g.,* the GTR+Γ+I substitution model) with well behaved prior distribution is applied.

Applying the most complex substitution model comes with the cost that both the likelihood calculation is slower and there are more parameters to be estimated. If we could be absolutely certain that a simpler substitution model suffices, then we could save computation time by applying this simpler substitution model. However, given the shortcuts and philosophical shortcomings of substitution model selection procedure and uncertainty/disagreement in selection substitution models [1], we argue that it is safer to err on the side of a too complex substitution model at a small amount of extra computational cost. Similarly, we argue that model averaging approaches [*e.g.,* 79, 18] are not necessary —because our results show that there is no difference in estimated phylogenies between different models if well behaved prior distributions are chosen— and only unnecessarily increase computational demands due to more complex model averaging algorithms (*e.g.,* reversible jump MCMC which is prone to poor MCMC convergence).

Furthermore, multi-locus datasets are used in phylogenetics since several decades where each locus evolves under a substitution process that is either independent or shared among loci [135]. The shortcuts taken to select the best partition model are more severe due to computational reasons [*e.g.,* 47] and their impact is less explored. In our simulation study we explored an extreme scenario where both the substitution model was over-parameterized and the locus was over-splitted. In line with our results on substitution model over-parameterization of single loci, we found that Bayesian analysis with a well behaved prior distribution does not bias phylogenetic inference. That means, one does not need to worry about assuming too many division of the data into subsets, although more data subsets come with a higher computational cost owed to the additional parameters.

In general, our results show that over-parameterization is not a problem in Bayesian inference if well behaved prior distributions are chosen, and these results could apply to models beyond substitution models. On the contrary, standard phylogenetic models are likely to be over-simplified and inadequate [34, 71, 147]. There exist several extensions to standard phylogenetic substitution models, such as the CAT model [108] and Markov modulated substitution models [7]. None of these more complex models are contained in substitution model selection approaches and our efforts should go into developing and testing more realistic substitution models.

Several previous studies have shown that Bayesian phylogenetic inferences can produce unrealistic long trees [22, 120, 145, 184]. These studies identified the prior distribution on the tree length as being responsible for the unrealistically large posterior estimates of the tree length. In our simulations, we noticed that if the true value for the tree length was far outside the center of prior distribution, then estimates of the tree length using phylogenetic analysis with the default prior settings for `MrBayes`, `BEAST` and `RevBayes` were significantly biased. These previous studies by [22], [120] and [184] used `MrBayes` and therefore the results are likely influenced by an interaction of the tree length prior distribution and ASRV prior distribution (see also [37]). Our proposed prior distribution (the `Tame` prior setting) does not show this interaction between tree length prior distribution and ASRV prior distribution and might alleviate the problem of unrealistically long trees.

In this manuscript we focused exclusively on over-parameterization of substitution models and the choice of prior distributions in a Bayesian inference framework. Our results may not be directly comparable to a Maximum likelihood (ML) framework and over-parameterization could still be a problem. In a ML framework, nuisance parameters, such as the substitution model parameters, are estimated to a single value that maximizes the likelihood instead of integrating over the uncertainty, which is done in a Bayesian framework. Therefore it is possible that a Bayesian inference is less impacted by over-parameterization as the uncertainty in the rate variation among sites could be large, but a ML inference on the other hand is impacted. A similar simulation study as ours could provide an answer to the question if over-parameterization is a problem for phylogenetic inference in a Maximum likelihood framework.

In conclusion, we found that over-parameterization is not a problem for Bayesian phylogenetic inference and does not bias tree topology estimates or branch length estimates if well behaved prior distributions are chosen. Our results corroborate previous findings by [81] and [111] on the robustness of estimating phylogenetic trees using over-parameterized substitution models. Here we additionally explored the impact of substitution model over-parameterization on estimates of the tree length (and thus by proxy on branch lengths) under different prior settings. We show that tree lengths estimates are more sensitive to substitution model over-parameterization and the choice of prior distributions. These prior distribution might result into unforeseen side-effects, for example, an informative prior distribution on the among site rate variation leads to more sensitivity to the prior on the tree length. We propose and tested a new choice of prior distribution, the `Tame` priors, which are well behaved. Our new choice of prior distribution can be applied to partition models and renders selection of the best partition scheme unnecessary. In general, substitution models are most likely too simple and our worries should focus on developing more realistic substitution models instead of selecting between a set of unrealistic substitution models.

Table 4.2: Description of the (default) prior settings for the commonly used phylogenetic tools and the proposed prior. The first column displays the name of the prior setting. The following columns are the parameters for the GTR+Γ+I model.

Prior	Branch length	Equilibrium base frequencies	Exchangeability rates	Alpha parameter	Proportion of invariant sites
Tame	Exponential (λ=10)	Dirichlet (1,1,1,1)	Dirichlet (1,1,1,1,1,1)	Uniform (0, 1×10^8)	Beta (α=1,β=1)
MrBayes	Exponential (λ=10)	Dirichlet (1,1,1,1)	Dirichlet (1,1,1,1,1,1)	Uniform (0,200)	Uniform (0,1)
RevBayes	Exponential (λ=10)	Dirichlet (1,1,1,1)	Dirichlet (1,1,1,1,1,1)	Lognormal (μ=ln(1.5), σ=0.587)	Beta (α=1,β=1)
BEAST	Exponential (λ=10)	Uniform (0,1)	Gamma (k=0.05, θ=10 or 20)	Exponential (λ=1)	Uniform (0,1)

Table 4.3: Description of the move settings during the MCMC. Each row corresponds to a model parameter, the second column shows the move applied to the parameter and the third column shows the weight for the moves. The tree topology moves were the Nearest Neighbor Interchange (NNI) and Subtree Pruning and Re-Grafting (SPR).

Parameter	Move	Weight
Tree	NNI (Nearest Neighbor Interchange)	# taxa
	SPR (Subtree Pruning and Re-Grafting)	$\frac{\#\ taxa}{2}$
Branch length	Branch length scale	# branches
Equilibrium frequencies	Beta simplex	4.0
	Dirichlet simplex	2.0
Exchangeability rates	Beta simplex	6.0
	Dirichlet simplex	3.0
Exchangeability rates (BEAST)	5 scale moves	2.0
Alpha	Scale	4.0
Proportion of invariant	Beta probability	4.0

Chapter 5

Evaluating Gene Tree Discordance on Mammalian Orthologous Markers

5.1 Abstract

Gene tree discordance imposes a challenge for species tree reconstruction. Although there are biological reasons to cause gene trees to disagree, the methods that estimate these trees are not error free. We investigated the possible reasons for gene tree discordance for a subset of multiple sequence alignments of the OrthoMam database. The first step was to estimate the gene trees under the Bayesian phylogenetic method. We started our investigation by exploring the Markov chain Monte Carlo (MCMC) performance with the methods for convergence assessment implemented in `Convenience`. Next, we assessed the inferred posterior probability of clades for well-established mammalian orders. Then, we evaluated the possibility of incomplete lineage sorting for the orders with most conflicting gene tree signal and performed posterior predictive tests of model adequacy for all gene trees. Our results showed that incomplete lineage sorting is improbable for the analyzed orders and, furthermore, the model failed to adequately describe the data for all tested genes.

5.2 Introduction

Reconstructing the species evolution history is of major interest for taxonomic, conservation and evolutionary research [162]. An imposing challenge for species tree inference in the phylogenomic era is that different genes do not necessarily share the same genealogy [32]. Some explanations for gene tree discordance are methodological such as errors in gene tree inference and false recovered orthology. Other explanations are biological events such as horizontal gene transfer and incomplete lineage sorting (ILS) [155, 126]. Current phylogenetic methods do not incorporate all possible sources of gene tree discordance, but some methods account for biological discordance reasons. Such a method is the multi-species coalescent (MSC). The MSC models the species tree and the gene trees in a joint process, while allowing gene trees to follow different histories. Recent studies have demonstrated that the MSC method is statistically consistent, in contrast to the concatenation (or super-matrix) method [118, 150, 95].

A drawback from the MSC method is that its statistical consistency had been demonstrated for simulation studies where the gene trees were known, and ILS played an important role. Therefore, the MSC requires the presence of ILS and the true gene trees to infer a robust species tree. While biological reasoning for gene tree discordance has received much attention, methodological reasons have rather been under-looked. A primary source of error for the Bayesian phylogenetic inference is the poor assessment of Markov chain Monte Carlo (MCMC) convergence. Here we investigate the robustness of Bayesian phylogenetic gene tree estimation using mammalian orthologous markers. Our approach includes assessing the convergence status of the analyses using the method developed in Fabreti and Höhna [39]. Then, we assessed the estimated posterior probabilities of some mammalian orders for different gene trees. These orders are widely accepted as monophyletic, and the gene trees should reflect this expectancy. Next, we calculated the coalescent unit times leading to some orders to verify how likely was that such lineages present ILS. Furthermore, we performed model adequacy posterior predictive simulations for each gene to evaluate how fit was the model to the data set.

Our results showed that proper convergence assessment has an impact on the recovery of phylogenetic relationships. Moreover, the discordance in gene trees could not be explained alone with ILS. We observed poor model fit for all tested genes, suggesting that current commonly used models of sequence evolution fail to capture the complexity of real data.

5.3 Methods

We explored the robustness of gene tree inference for a subset of orthologous markers from mammals. Mammals are a vastly studied group with well-curated available data sets. The subset of genes was taken from OrthoMam v10c database [154]. OrthoMam is a database of mammalian orthologous markers. The database includes 14509 CDS (coding sequences) alignments with up to 116 taxa. To test the robustness of current gene tree estimation methods, we selected a sample of 180 CDS alignments with the greatest number of sampled taxa. We performed the Bayesian phylogenetic gene tree estimation with the software `RevBayes` [74]. The substitution model for the gene tree inference was set to GTR+Γ+I [165, 177, 2, 62], which is the most complex model within the commonly used GTR family of nested models. The prior distributions for the model parameters were set as described in chapter 4. The MCMC settings were varied to improve the MCMC convergence success rate, as explained in the next section. The convergence assessment was performed using the R package `Convenience` [39]. The methods implemented in `Convenience` provide a robust assessment of Effective Sample Size (ESS) for both the continuous parameters and the topology. Besides, `Convenience` compares multiple MCMC replicates to test whether independent runs converge to the same posterior distributions. After trying different MCMC settings, we ended up with a total of 29 analyses that achieved the convergence status according to `Convenience`.

After the evaluation of convergence for the gene tree inferences, we explored the accuracy of the inferred trees by assessing the estimated posterior probability of mammalian orders that have been established as monophyletic. The criterion to include orders was those with at least two specimens present in the list of all taxa. This resulted in 12 orders, namely: Afrotheria, Artiodactyla, Carnivora, Chiroptera, Eulipotyphla, Lagomorpha, Marsupialia, Perissodactyla, Primates, Rodentia, Scandentia and Xenarthra. The following taxa were single representatives of their order: *Galeopterus variegatus*, *Manis javanica* and *Ornithorhynchus anatinus*. We did not include these taxa in the analysis, because the posterior probability of a clade with a single taxon is always one. We further explored the gene trees by visually inspecting the maximum a posteriori (MAP) tree.

To investigate the possibility of ILS in some lineages leading to the orders of the mammal evolutionary tree, we estimated the branches leading to Rodentia and Eulipotyphla in coalescent unit times. Rodentia and Eulipotyphla were the two orders with most gene trees showing low support for their monophyly. Three gene trees resulted in posterior probabilities below 0.5 for the monophyly of Rodentia, whereas Eulipotyphla had a total of five gene trees not supporting its monophyly. ILS is more probable to occur when the divergence time among lineages is very short and the effective population size (N_e) is large. We can calculate the distance among lineages in terms of coalescent unit times (t_{coal}) by dividing the number of generations since the divergence of the lineages (t) by N_e [32]:

$$t_{coal} = \frac{t}{N_e} \tag{5.1}$$

The coalescent unit times are a normalization of branch lengths by population sizes. Small coalescent unit times correspond to higher probability of observing ILS for the corresponding branch.

Additionally, we assessed the model adequacy for each gene tree by performing posterior predictive tests [152, 14] on `RevBayes` [70]. Posterior predictive simulations consist of simulating data sets based on the posterior distributions of the inference of the empirical data. Then, summary test statistics compare the simulated data with the empirical one to evaluate how they deviate from one another. This comparison is done by calculating p-values, *i.e.,* the fraction of simulated data test statistics that are smaller or larger than the empirical test statistic. If the p-value is particularly small or large, the model is poorly fitted to the empirical data. The test statistics can be classified in focal and ancillary statistics. The focal statistics are those that are directly affected by the model, *e.g.,* the number of invariant sites when the Invariant sites model (I) [2, 62] is applied. On the other hand, the ancillary statistics are those that are not particularly accounted for in the model, *e.g.,* the variance of GC content across sequences, since the model does not assume variance in GC content across the tree.

5.3.1 Assessing MCMC performance

We estimated the unrooted gene trees with an initial MCMC setting of two replicate runs each with 50×10^3 iterations and the moves strategies described in Table 5.2. We assessed the convergence status of the MCMC output with the R package `Convenience` [39]. This initial analysis was performed for 178 different alignments, however for this batch of analysis only 19 achieved convergence. This MCMC settings yielded a 10.6% rate of convergence success. This low rate of convergence success showed that the MCMC settings were unsatisfying for the tested data set. Next, we investigated the MCMC performance for different settings of iterations and moves.

We chose four alignments that had the maximum number of taxa within the data set to initially test the MCMC settings. Our first approach was to increase the MCMC iterations and assess the convergence status of the output. Table 5.1 shows the characteristics of the four chosen genes, such as number of taxa and number of sites, and the number of MCMC iterations each gene required for achieving convergence.

Table 5.1: Description of the characteristics of the four alignments used to test the MCMC settings. The last column shows the number of MCMC iterations necessary for each alignment to achieve convergence.

Gene name	Number of taxa	Number of sites	MCMC iterations to converge
COL14A1	116	5397	100×10^3
PDK4	116	1236	150×10^3
FBXL3	116	1290	$> 200 \times 10^3$
HNF1B	116	1680	$> 200 \times 10^3$

Since half of our small, tested data set converged for an MCMC with up to 150×10^3 iterations, we increased the data set to 20 alignments to check the rate of success of convergence for this MCMC setting. Only four out of the 20 alignments achieved convergence, yielding a rate of success of 20%. Table 5.3 shows the number of parameters that failed for each of the tested criteria for the 16 analysis that did not converge.

For the initial set of four alignments, two failed to converge even with 200×10^3 iterations. Figure 5.1 shows the convergence assessment for these alignments. The continuous parameters have achieved the minimum ESS and the comparison between their distributions for different MCMC runs are within the threshold. But the tree parameters (splits)

Table 5.2: Description of the move settings during the MCMC. Each row corresponds to a model parameter, the second column shows the move applied to the parameter and the third column shows the weight for the moves. The tree topology moves were the Nearest Neighbor Interchange (NNI) and Subtree Pruning and Re-Grafting (SPR).

Parameter	Move	Weight
Tree	NNI (Nearest Neighbor Interchange)	# taxa
	SPR (Subtree Pruning and Re-Grafting)	# taxa
Branch length	Branch length scale	# branches
Equilibrium frequencies	Beta simplex	3.0
	Dirichlet simplex	3.0
Exchangeability rates	Beta simplex	3.0
	Dirichlet simplex	3.0
Alpha	Scale	2.0
Proportion of invariant	Beta probability	2.0

Table 5.3: The number of parameters that failed each convergence criterion for all 16 gene trees that failed convergence. ESS is the effective sample size. KS is the Kolmogorov-Smirnov test.

Gene tree	# ESS fails for continuous parameters	# KS fails for continuous parameters	# ESS fails for splits	# difference between splits fails
CCDC136	1	0	40	14
CC2D1A	0	0	66	84
SLC9A5	0	0	52	1
TOR1A	0	0	41	22
TPMT	0	0	1	0
ITPKA	0	0	3	7
LOXL4	0	0	9	0
SCARB2	0	51	0	262
WDFY2	0	0	89	22
SERBP1	0	0	102	88
STRA8	0	0	1	1
TMEM219	0	0	8	0
FXYD4	0	0	0	2
GABRA2	4	0	18	15
CCL28	0	0	10	0
HNF1B	0	0	36	0

have failed to achieve the minimum ESS and the comparison between runs shows that the difference in split frequencies is above the accepted threshold. This is an expected behavior since phylogenetic tree spaces are very large and the most challenging parameter to sample in phylogenetic analyses [98]. For this reason, we investigated next MCMC settings that would improve the sampling of trees.

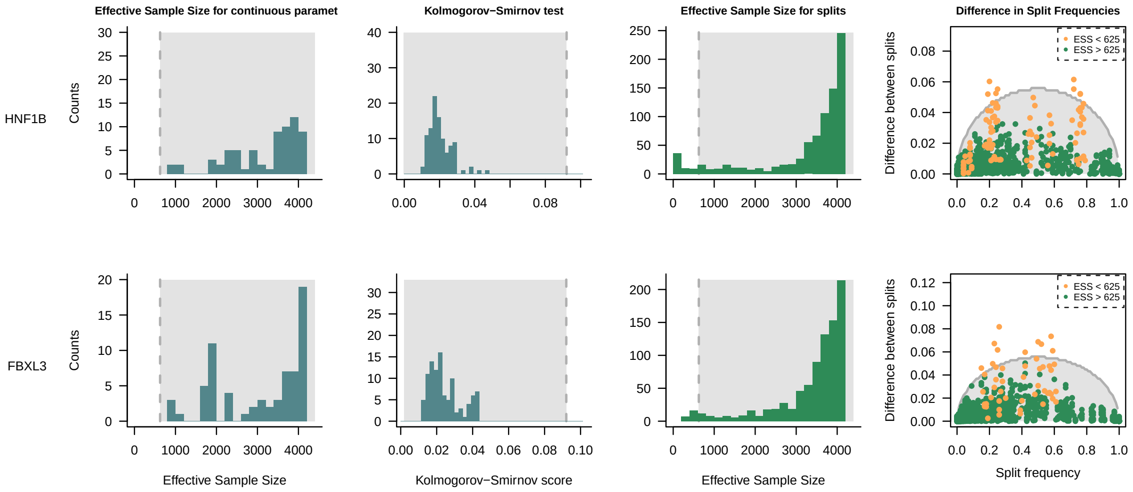


Figure 5.1: Convergence assessment for two analysis that failed to converge with 200×10^3 iterations. The top row shows the results for the gene HNF1B, while the bottom row for the gene FBXL3. The first two columns present the plots for the continuous parameters and the last two columns for the splits.

We tested two MCMC alternatives for one of the alignments that did not converge, namely HNF1B. First, we performed a Metropolis-coupled MCMC (MC^3 or MCMCMC) [5, 128]. Second, we used the original MCMC settings and added the Subtree Swap [35] move with weight # *taxa*. We implemented a modified Subtree Swap move that performs on unrooted trees, since the original implementation was set for ultrametric trees. Figure 5.2 shows the convergence assessment of the tree parameters for these two alternative MCMC configurations. The analysis including the Subtree Swap move improved the reproducibility between MCMC replicates, but the ESS for the single replicates still showed values below the desired threshold. The MC^3 analysis showed a performance worse than the original MCMC settings (Figure 5.1 bottom row) with the comparison between MCMC replicates presenting values higher than the original setting.

Our next attempt was to change the weights on the moves and run the MCMC for 100×10^3 iterations. Table 5.4 summarizes the weights on all the moves used in the MCMC. We changed the strategy of the moves on the tree by increasing the weight on the NNI move and decreasing the weight on the SPR move. We also slightly increased the weight on the branch length. For the equilibrium frequencies and exchangeability rates we chose to use just one move called Simplex element scale with weight 2.0. We kept the strategy on α and decreased the weight on the proportion of invariants to be the drawn

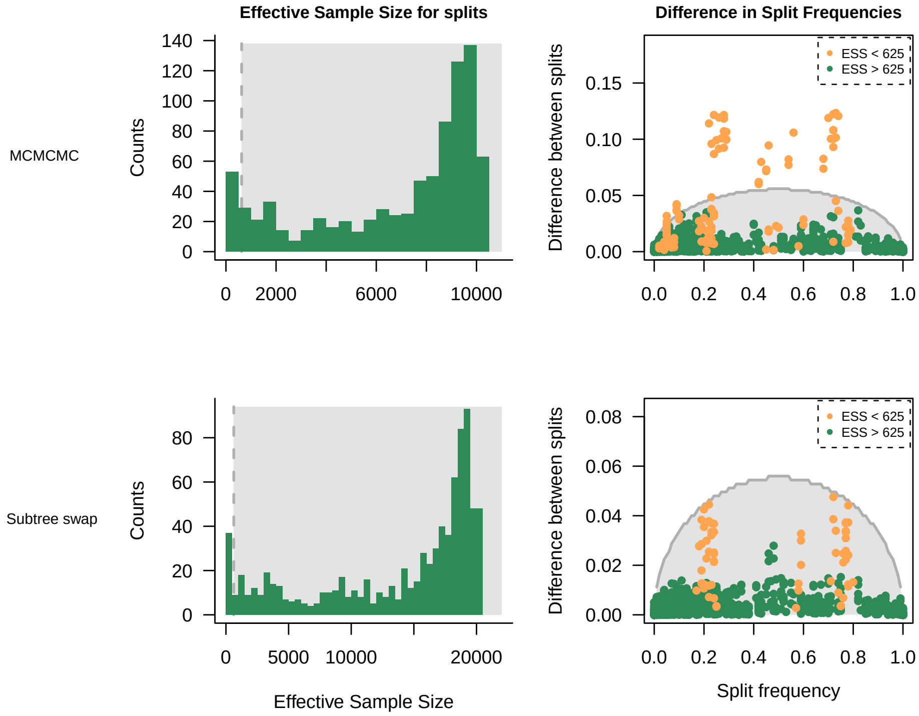


Figure 5.2: Comparison of convergence assessment for the splits between two different MCMC strategies. The first row presents the convergence assessment for the analysis performed with MC^3, while the strategy for the second row was to add the move Subtree Swap.

probability of being invariable. This new strategy increased overall the moves taken on the tree, while decreasing the moves on the continuous parameters. We tested the new MCMC arrangement on 20 new alignments. Eight of which achieved convergence. The rate of success for the new settings is 40%. This shows significant improvement in the rate of convergence with the new MCMC settings. But further investigations are needed to increase more the rate of success.

5.4 Results

5.4.1 Posterior probability of mammalian orders

The posterior probabilities (PP) of the 12 mammalian orders is depicted in Table 5.5. The posterior probabilities under 0.5 are highlighted in red. The sampled taxa for all genes were not the same. Therefore, the genes that had no representatives of a given order have the $-$ sign on the table. The orders Xenarthra, Marsupialia, Carnivora and Scandentia

Table 5.4: Description of the updated move settings during the MCMC. Each row corresponds to a model parameter, the second column shows the move applied to the parameter and the third column shows the weight for the moves. The tree topology moves were the Nearest Neighbor Interchange (NNI) and Subtree Pruning and Re-Grafting (SPR).

Parameter	Move	Weight
Tree	NNI (Nearest Neighbor Interchange)	$\# \, taxa * 2$
	SPR (Subtree Pruning and Re-Grafting)	$\frac{\# \, taxa*2}{10}$
Branch length	Branch length scale	$\# \, taxa * 2$
Equilibrium frequencies	Simplex element scale	2.0
Exchangeability rates	Simplex element scale	3.0
Alpha	Scale	2.0
Proportion of invariant	Beta probability	probability of invariant

have posterior probabilities 0.85 or higher for all genes that these orders were present. All the other orders show at least one gene with low posterior probability. Figure 5.3 shows the histograms of posterior probabilities for each order. Eulipotyphla shows the least agreement among the tested genes, with approximately 44% of the genes with posterior probability lower than 0.9.

We assessed the impact of convergence assessment in the posterior probabilities by summarizing the posterior probabilities for the 12 orders for 29 analysis that did not converge. Figure 5.4 shows the histograms of posterior probabilities for each order. All orders apart from Marsupialia and Scandentia show analysis with posterior probabilities below 0.85. The most prominent changes occur for Primates and Rodentia. For these groups, there is a shift from a high peak in high PP to intermediate and low PP values.

5.4.2 Gene trees with conflicting clades

A total of eleven gene trees presented posterior probability below 0.5 for at least one of the tested orders. Apart from the low posterior probability for the orders, the gene trees also presented conflicting topologies and the placement of groups along the tree varied. We evaluated the topologies for the gene trees with low posterior probability for three orders: Rodentia, Lagomorpha and Eulipotyphla. We used the tree in Figure 5.5 as a reference for the expected species tree. This species tree was constructed by pruning the whole tree provided in [188]. The whole tree in [188] was inferred from a Bayesian molecular-clock dating approach for 4705 mammalian species. We further used this time tree to approximate the divergence time between groups of interest. Finally, we compared the expected species tree with the species tree inferred with the 29 gene trees generated in the present study.

Table 5.5: Posterior probabilities for the 12 mammalian orders for the 29 gene trees that achieved convergence. Values below 0.5 are highlighted in red.

Gene	Xenarthra	Marsupialia	Afrotheria	Artiodactyla	Carnivora	Chiroptera	Eulipotyphla	Perissodactyla	Lagomorpha	Primates	Rodentia	Scandentia
CROT	1	1	1	1	1	0.99	1	1	1	1	0.018	1
AGPAT4	–	0.99	1	1	1	0	–	0.02	1	0.82	1	1
NECTIN1	–	1	1	0.99	0.99	0	0	1	1	0.97	0.88	1
PGK1	–	–	0	0	1	1	0.93	–	–	0.99	0.99	–
ZYG11B	0.99	1	1	1	1	1	0.63	1	0	1	0	1
TMCO4	0.99	1	0.99	0.99	1	0.99	0	1	1	0.91	0.99	1
LYPD8	–	–	1	0.99	0.97	–	–	–	–	0.03	0.98	–
MATK	–	1	1	0.99	0.98	0.99	0	0.99	–	0.97	0.93	–
GP6	–	–	0.99	1	0.99	0.99	–	1	–	0.99	0	–
MAP3K1	1	1	1	1	1	1	0.41	1	1	1	1	1
LARP4	1	1	0.99	0.99	1	1	0.11	0.99	–	0.81	0.99	1
ACSM3	0.99	1	1	1	1	1	1	1	1	1	1	1
ST7L	0.99	1	1	0.99	1	0.99	0.70	0.99	1	0.99	1	1
GPC1	–	1	1	1	1	1	–	1	–	0.82	0.99	1
EIF2AK1	1	1	1	1	1	0.99	0.89	1	1	1	1	1
NME8	0.99	1	0.99	1	1	1	0.77	1	1	1	1	1
TMPRSS11E	1	1	0.99	0.99	1	1	0.99	0.99	–	1	1	–
FAM83D	1	1	1	1	1	1	0.99	1	1	1	0.99	1
METTL4	0.99	1	1	1	1	1	0.93	1	1	0.99	0.92	1
MED24	1	1	0.97	1	1	1	–	1	1	1	1	1
MRVI1	–	1	1	1	1	1	0.80	0.99	1	0.95	1	1
LLGL2	–	1	1	1	1	1	0.99	1	1	1	1	1
DYNC1I2	0.86	1	0.99	1	1	1	0.97	1	1	0.98	0.99	1
BCORL1	–	–	0.98	1	1	1	1	1	1	0.99	1	1
FLT1	0.99	1	1	1	1	1	1	1	1	1	1	1
ROBO3	–	1	1	1	1	1	1	1	1	1	1	–
TCP11X2	–	–	1	1	1	1	–	1	1	0.99	1	–
PDK4	1	1	1	1	1	1	1	1	0.99	0.99	0.99	1
COL14A1	1	1	1	1	1	1	1	1	1	0.99	1	1

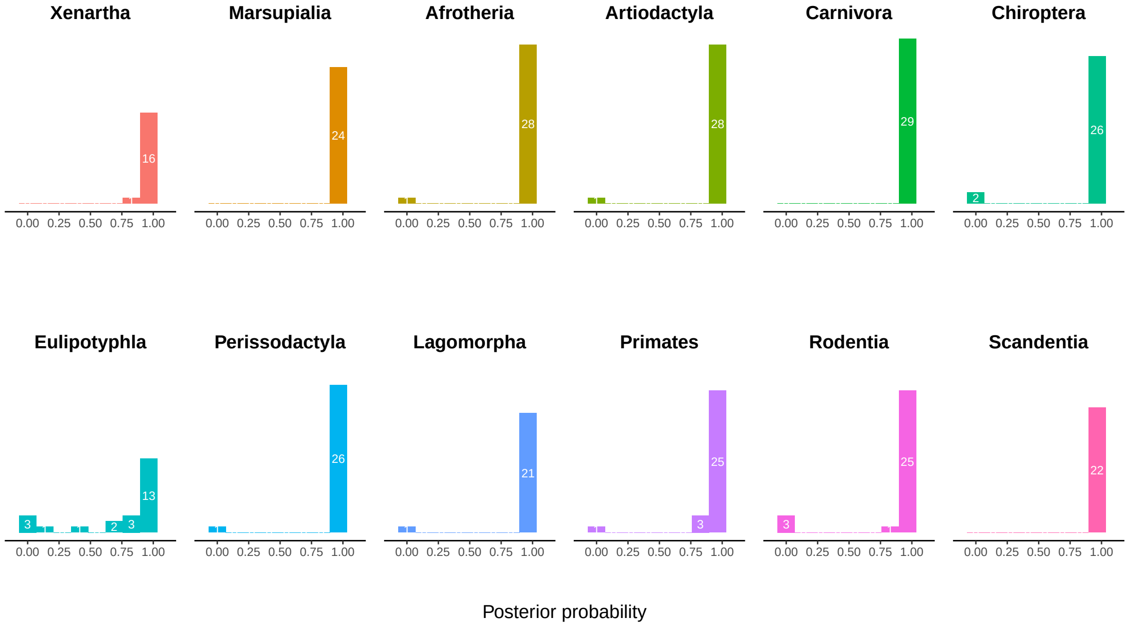


Figure 5.3: Histograms of posterior probabilities for the clades corresponding to the mammalian orders for the 29 analyses that achieved convergence. A probability of one means full support for the clade across the entire posterior distribution of topologies. A probability of zero means no support for the clade. If the histogram is bimodal, this indicates that there is discordance among the gene trees in support of the order.

Rodentia and Lagomorpha Three gene trees inferred very low posterior probabilities for the clade Rodentia, one of which also inferred posterior probability of 0 for the clade Lagomorpha. The maximum a posteriori (MAP) tree for these analysis are shown in Figure 5.6. For these three gene trees, the sampled taxa from Lagomorpha appears within rodents. The gene tree for the gene CROT (Figure 5.6A) displays Lagomorpha as a sister group to the suborder Ctenohystrica. For the gene GP6 (Figure 5.6B) the taxon *Tupaia chinensis* that belongs to the order Scandentia is grouped together with rodents and *Heterocephalus glaber* (Lagomorpha). The gene ZYG11B (Figure 5.6C) shows Lagomorpha and Rodentia as early branchings in the tree and both orders are not monophyletic. The sampled taxa are not equal among the three genes, the gene GP6 has the lowest number of sampled taxa with 92 in total. Besides no representatives of the orders Marsupialia and Eulipotyphla are included. The gene trees for CROT and ZYG11B have 112 and 114 taxa, respectively. In this case, all orders have at least one representative taxon sampled. Apart from the recovered monophyly of the orders, these gene trees are also different for the placement of the groups along the tree. The gene trees for CROT and GP6 are more similar to the expected topology of the species tree (Figure 5.5). While the gene ZYG11B differs more from the expected species tree, for example with Scandentia as a sister group to Primates.

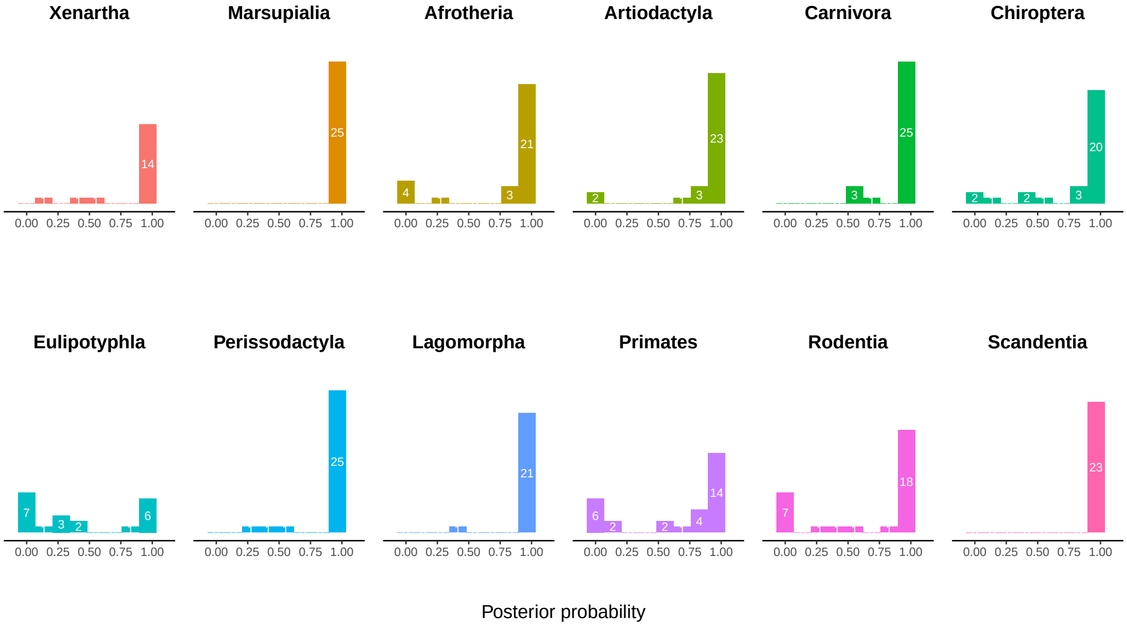


Figure 5.4: Histograms of posterior probabilities for the clades corresponding to the mammalian orders for 29 analysis that did not achieve convergence. A probability of one means full support for the clade across the entire posterior distribution of topologies. A probability of zero means no support for the clade. If the histogram is bimodal, this indicates that there is discordance among the gene trees in support of the order.

Eulipotyphla The order Eulipotyphla has in total three sampled species for the OrthoMam database, namely: *Condylura cristata*, *Erinaceus europaeus* and *Sorex araneus*. Five gene trees presented low posterior probability for the monophyly of the group. Therefore, Eulipotyphla was the tested order with the highest number of gene trees with low support for the monophyly of the group. The MAP trees for these gene trees are shown in Figures 5.7 and 5.8. The gene MATK (Figure 5.7A) recovered *E. europaeus* as sister taxa to the grouping of Lagomorpha and Scandentia. For this gene tree there is only one sampled species for the orders Lagomorpha and Scandentia. The group *E. europaeus*, Lagomorpha and Scandentia is placed as sister to Rodentia. Finally, the species *C. cristata* appears as sister to the whole Euarchontoglires group. The gene MAP3K1 (Figure 5.7B) places the taxa belonging to Eulipotyphla within Laurasiatheria, in accordance with the expected species tree. But *C. cristata* is not grouped together with *E. europaeus* and *S. araneus*. Instead *C. cristata* is a sister branch to Scrotifera and separated from the other Eulipotyphla taxa. The gene NECTIN1 (Figure 5.7C) recovered *S. araneus* grouped with *Dasypus novemcinctus* (Xenarthra) and this clade as sister to Chiroptera. *E. europaeus* and *C. cristata* are grouped together as a clade, but the position within Laurasiatheria is in disagreement with the expected species tree. The gene tree for LARP4 (Figure 5.8A) shows the clade formed by *E. europaeus* and *C. cristata* separated from *S. araneus*. This gene tree recovered Eulipotyphla as an early branching in the tree and closer related to Marsupialia. Finally, the gene tree for TMCO4 (Figure 5.8B) presents *S. araneus* as a sister taxa to Lagomorpha and *C. cristata* within Laurasiatheria.

5.4.3 Coalescent unit times

We estimated the coalescent unit times for the branches leading to Rodents and Eulipotyphla for different possible value of effective population size. For the divergence times we used the time tree from [188]. The approximate estimated time for the branch that leads to the first split within Rodentia is 5.5 million years, and 3.8 million years for Eulipotyphla. We approximated the generation time to be 100 days for Rodentia [121] and 253 days for Eulipotyphla [163]. From Equation 5.1 we estimated the coalescent unit times for both orders as seen in Tables 5.6 and 5.7.

As expected from Equation 5.1, t_{coal} is inversely proportional to N_e. Therefore, larger N_e values produce smaller t_{coal} values. Precisely, t_{coal} above five indicates high probability of lineages having coalesced. Since this is expected after $\sim 5N_e$ [32] and coalescent times are proportional to the effective population size. The estimated t_{coal} for Rodentia indicates that the effective population size for the lineage leading to rodents would require values in the magnitude of 10^7 to support ILS in that branch.

Table 5.6: Estimated coalescent unit times for the order Rodentia for different effective population sizes (N_e).

N_e	Coalescent unit time
10^4	1992.5
5×10^4	398.5
10^5	199.2
10^6	19.9
10^7	1.9

Table 5.7: Estimated coalescent unit times for the order Eulipotyphla for different effective population sizes (N_e).

N_e	Coalescent time units
10^4	542.2
5×10^4	108.4
10^5	54.2
10^6	5.4
10^7	0.54

5.4.4 Model adequacy

We calculated p-values for ten focal statistics and four ancillary statistics as shown in Table 5.8. The p-values below 0.025 or above 0.975 are colored in red. We evaluated the gene trees separately between the ones that showed no small posterior probability for the mammalian orders (well behaved) and the ones that showed at least one posterior probability below 0.5 for one of the orders (poor behaved). Furthermore, we looked at focal and ancillary test statistics separately for these two groups. The histograms of p-values for the focal statistics are shown in Figures 5.9 and 5.10. The histograms for the ancillary statistics are shown in Figures 5.11 and 5.12. Overall, the well-behaved analyses and the poorly behaved analyses show no clear difference in model fit for the tested statistics. Among the focal test statistics, the maximum GC is the only statistic that seems to show better fit for the well-behaved analyses. On the other hand, the multinomial-likelihood showed more intermediate p-values for the poorly behaved analyses. For the ancillary test statistics, the well-behaved analyses showed more intermediate p-values than the poorly behaved analyses for all cases. But the sample size for the analyses was rather small and more replicates are needed to see clearer tendencies. Importantly, not a single analysis showed an appropriate model fit for all tested statistics.

5.5 Discussion

Our results demonstrate the importance of proper convergence assessment in Bayesian phylogenetic tree inference, where the lack of convergence can lead to spurious topologies and false conclusions about evolutionary relationships. We tested different MCMC settings to obtain a good rate of convergence success. Our investigation suggests that arbitrarily increasing the number of MCMC iterations does not produce a better rate of convergence success. Instead, a better approach is to propose more moves at each iteration. The most challenging parameter for convergence purposes, and yet the object of interest in most phylogenetic studies, is the topology. Our newly implemented Subtree Swap move seemed to perform better than the inference with only NNI and SPR and much better than the inference using MC^3. However further investigations of MCMC moves for topologies are necessary to grant optimal settings for convergence with real data sets.

After testing miscellaneous MCMC settings, we ended up with 29 gene trees that achieved convergence according to the criteria in Fabreti and Höhna [39]. For these 29 analyses we evaluated the posterior probability of mammalian orders. Our results show that eleven gene trees present very low support for the monophyly of at least one of the tested mammalian orders. We further investigate the coalescent unit times for the branches leading to two orders: Rodentia and Eulipotyphla. Incomplete lineage sorting is perhaps the most prominent hypothesis for explaining conflict among gene trees. For incomplete lineage-sorting alone to explain the conflicts in our gene trees, we would need to have effective population sizes on the order of 10^7. Estimates of effective population size for rodents show results in the order or 10^5 [168, 183]. Although there are no reliable estimates

Table 5.8: Posterior predictive mid-point p-values for different test statistics. Each row corresponds to a gene tree analysis that achieved convergence. p-values below 0.025 and above 0.975 are highlighted in red.

Gene	Max GC	Min GC	Mean GC	# Invariant Sites	Max Pairwise Diff	Min Pairwise Diff	Theta	Tajima-D	Tajima-Pi	Multinomial-Likelihood	Var GC	Max Invariant Block Length	Max Variable Block Length	# Invariable Block
	Focal statistics										Ancillary statistics			
CROT	0	0	0	0	0.91	1	0.26	1	0	0.15	1	0.9	0	0.79
AGPAT4	0	0.34	0	0.22	0	0.97	0.06	1	0	0.32	1	0.9	1	0
NECTIN1	1	0	1	0.36	0.32	0.97	0.5	1	0.5	0.99	0.5	0.96	1	0
PGK1	1	0	1	0.34	1	1	0.28	1	0.06	1	0.01	0.96	0.73	0.5
ZYG11B	0.96	0	0	0.38	1	1	0.97	0.96	1	0.5	0	0	1	0.13
TMCO4	0.82	0.34	1	0	0.5	0	0.96	0	0.5	0.2	0	0	0.47	0.5
LYPD8	0	1	0	0	0.02	0	1	0.91	0.8	0	0.5	1	1	0.91
MATK	0	0.66	0.72	1	0	1	0.99	1	0.87	0	0.01	1	0	0
GP6	1	0.95	0	1	0	1	0	1	0.96	0.5	1	1	0.77	0
MAP3K1	1	0.99	0	0	0.5	1	0	0.5	0.5	0.2	0.5	1	1	0
LARP4	0.02	0.99	1	0	0.02	1	1	0.06	0.99	0.99	0.99	0.99	1	0
ACSM3	0.5	0.01	0.1	0.86	1	0.03	0.9	0.99	0	0	1	0.26	0	0.5
ST7L	0.03	1	0	0.86	1	0	0.94	0.15	0.91	0	0.71	0.21	0.53	0.04
GPC1	1	1	0	0.01	0.95	0	0.91	0.98	0.19	0	0.01	0.04	0.5	1
EIF2AK1	1	0.4	0.97	0.01	0.27	0.01	0.41	1	0.98	1	0	0.02	1	0
NME8	1	0	0.97	0.01	0	1	0.07	1	0	1	0.53	0.11	1	0.95
TMPRSS11E	0.76	0.65	0	0.01	0	0.57	0.77	0.28	0	1	0.67	0.11	0.07	0.43
FAM83D	0.88	0.01	0	0.52	0.95	0.55	0.72	0.15	0.35	1	0.98	0.13	0.5	0.86
METTL4	0.68	1	0	1	0.27	0.01	1	0.01	0.24	0	0.95	0.13	0.26	0
MED24	0.94	0.5	0	0.85	1	0.91	1	0.01	0.98	1	0.91	0	0.5	0.57
MRVI1	0	0.32	0	0.73	0.72	0.96	1	0.28	1	1	0.28	0	0.02	0
LLGL2	0.27	0	1	0.01	0.92	1	1	0.15	0	1	0.41	1	0.02	0.43
DYNC1I2	0.9	0	0.84	0.06	0	0	0	1	0	0.5	0.99	1	0.03	1
BCORL1	1	0.5	0.99	1	0.99	0.05	0	0.99	0.44	0.03	1	0.94	1	1
FLT1	0.27	0.32	0	1	0.69	0	1	1	0.44	0	0.89	1	1	0.02
ROBO3	0.01	1	0.54	0.5	0.96	0	1	0.87	0	0	0.89	0.84	0	0
TCP11X2	0.32	1	0.34	0.03	0.01	0	1	1	0	0.5	1	1	0	0.26
PDK4	0.4	0.95	1	0	0	0.05	1	0.33	0.03	0.03	1	0	0	0.12
COL14A1	0.79	0.39	1	0	0.53	1	1	0.04	0.03	1	0.28	0.98	0	0

for Eulipotyphla, we have no reason to believe that past populations presented such large effective population sizes. Since the necessary effective population size for ILS presented improbable values, we conclude that the conflict in the gene trees cannot be explained due to ILS alone. The lack of ILS support in older splits within mammals had been previously shown in Scornavacca and Galtier [155], where ILS was found to be only a minor contribution of the conflicting phylogenetic signal in mammals. Furthermore, Scornavacca and Galtier [155] compared the gene tree discordance for exons within the same gene and exons from different genes. They observed that the amount of discordance was similar for the two groups. This finding could be explained by recombination (biological explanation) or error in the estimation of the gene trees (methodological explanation).

We verified the second hypothesis by exploring the model adequacy of the substitution model to the empirical data. We observed that not a single gene showed proper model fit. Both gene trees with conflicting posterior probability for one of the mammalian orders and the gene trees with no conflict showed poor model fit. Our results for the posterior predictive p-values presented extreme values for both focal and ancillary test statistics. The extreme values for the focal statistics should be considered as a stronger signal of model inadequacy, since the distributions of these p-values under the null hypothesis have been shown to be non-uniform (*Chapter 3*).

Overall, the results here presented suggest that a considerable amount of gene tree discordance is due to methodological grounds. The current commonly used sequence evolution models do not adequately reflect the heterogeneity in real data, and therefore future research should focus on the aspects of the model that are failing to capture the complexity of real data.

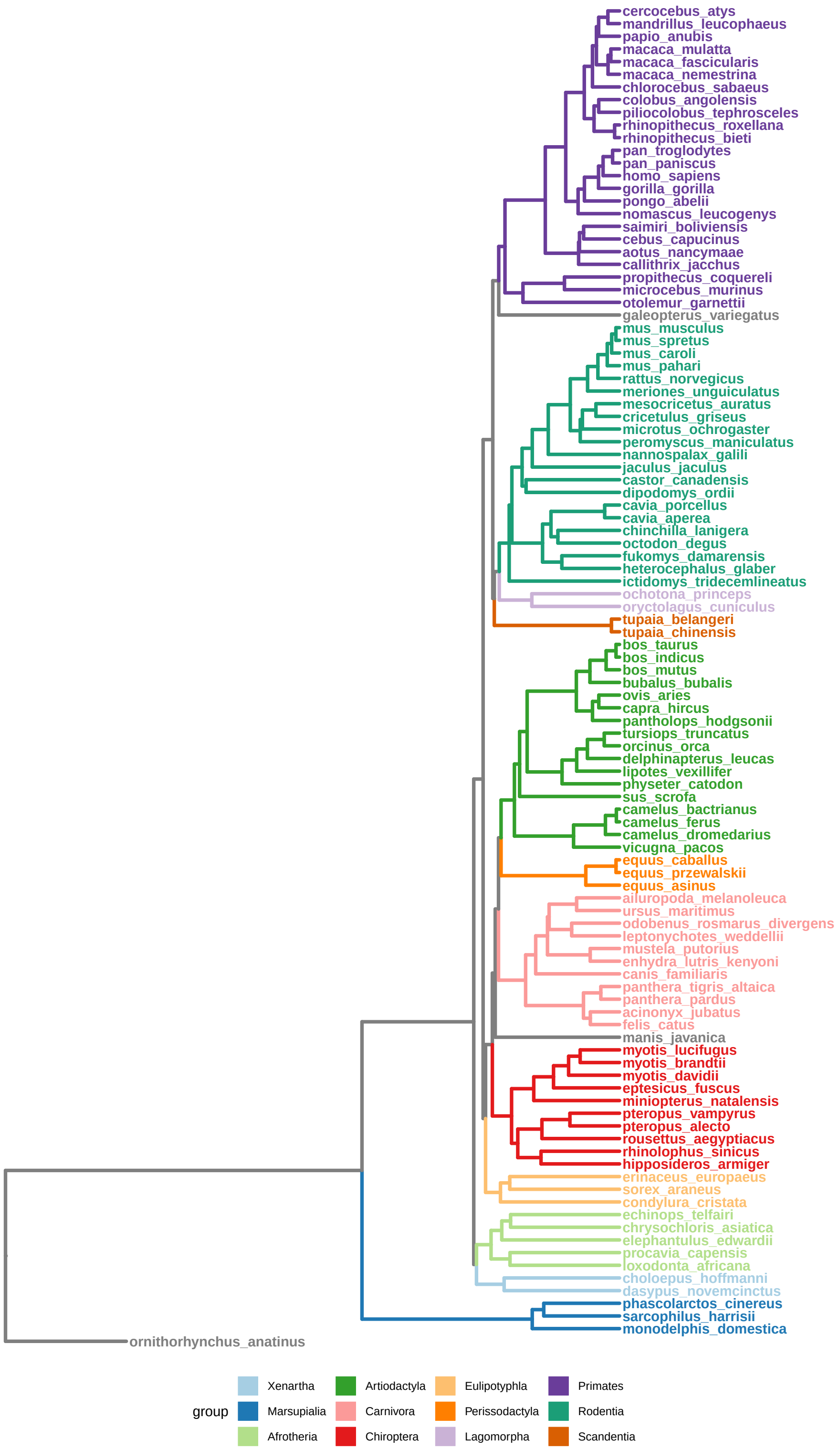


Figure 5.5: Species tree reconstructed based on the analysis performed in [188] for the 116 species evaluated in the present study. The different colors represent the 12 mammalian orders relevant for this study.

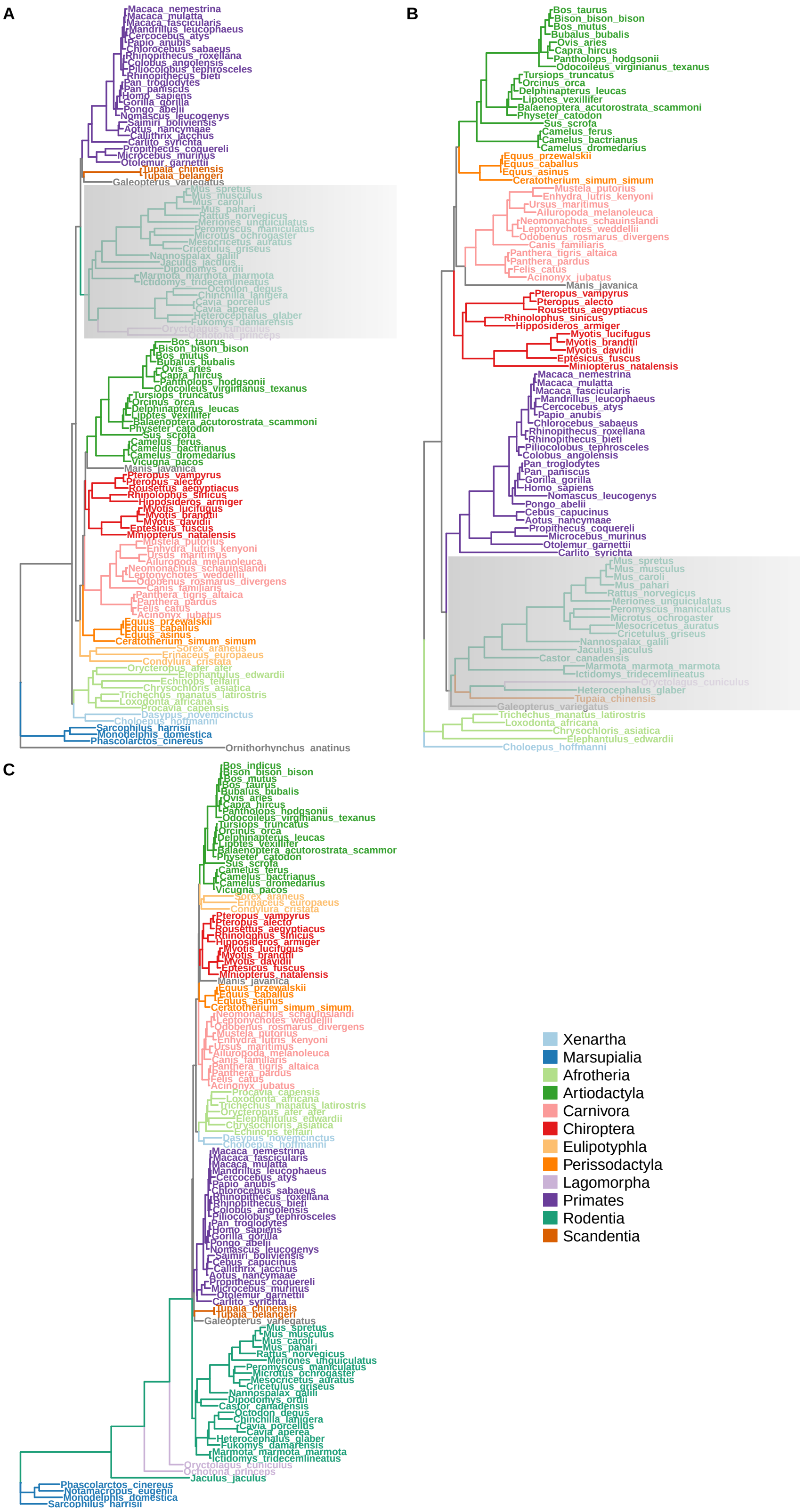


Figure 5.6: Maximum a posteriori gene trees for the genes CROT(A), GP6(B) and ZYG11B(C).

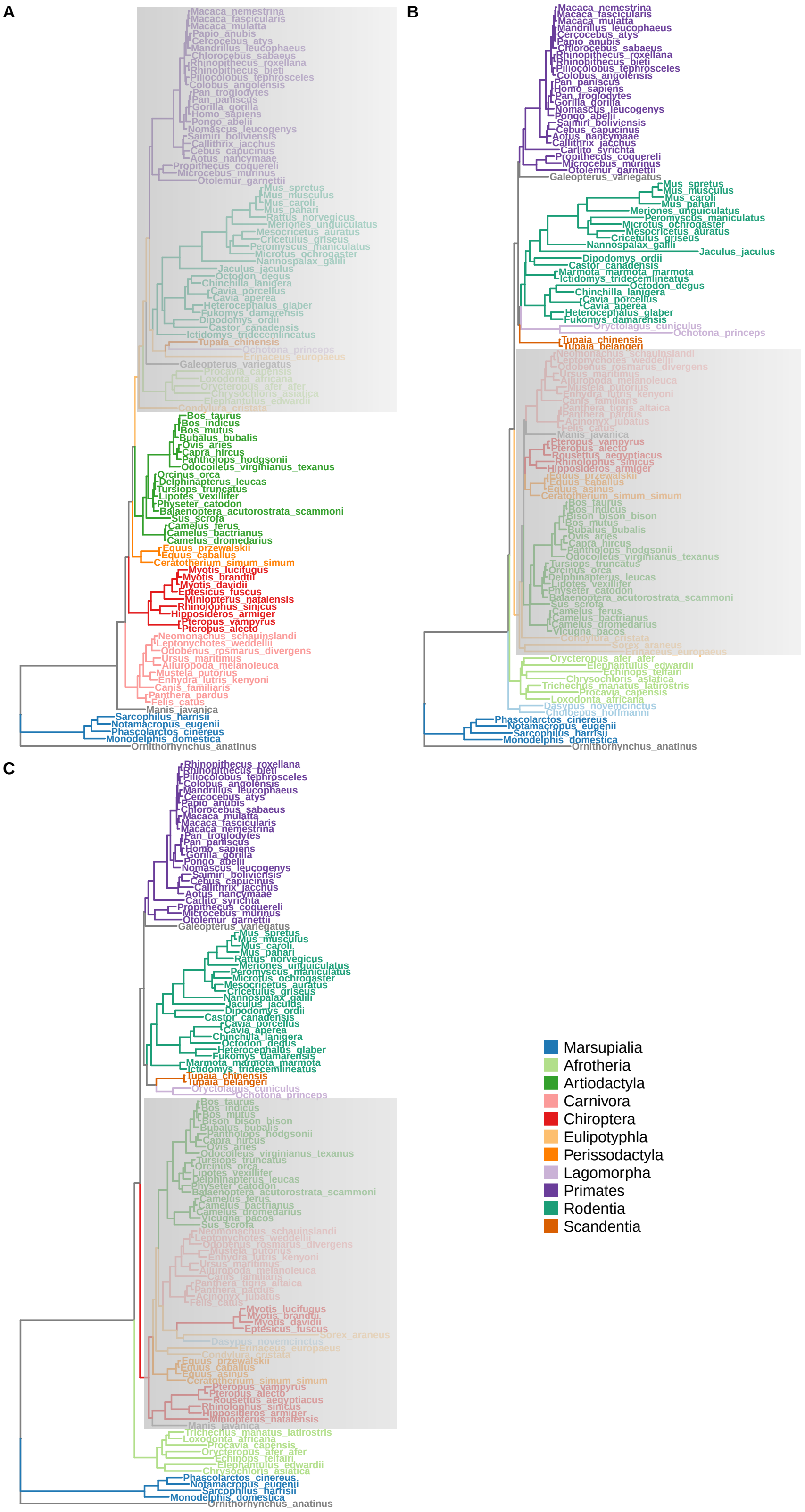


Figure 5.7: Maximum a posteriori gene trees for the genes MATK(A), MAP3K1(B) and NECTIN1(C).

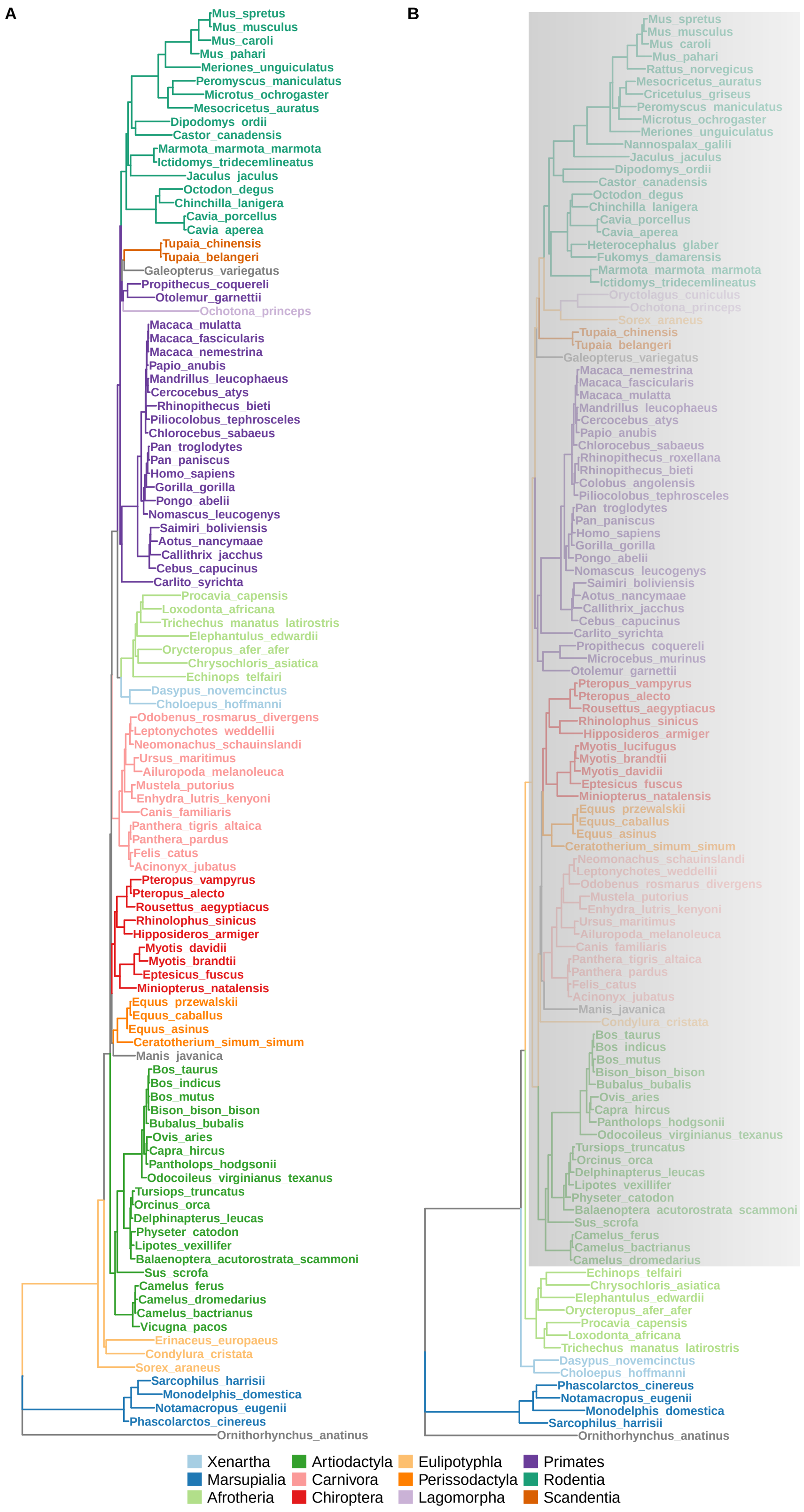


Figure 5.8: Maximum a posteriori gene trees for the genes LARP4(A) and TMCO4(B).

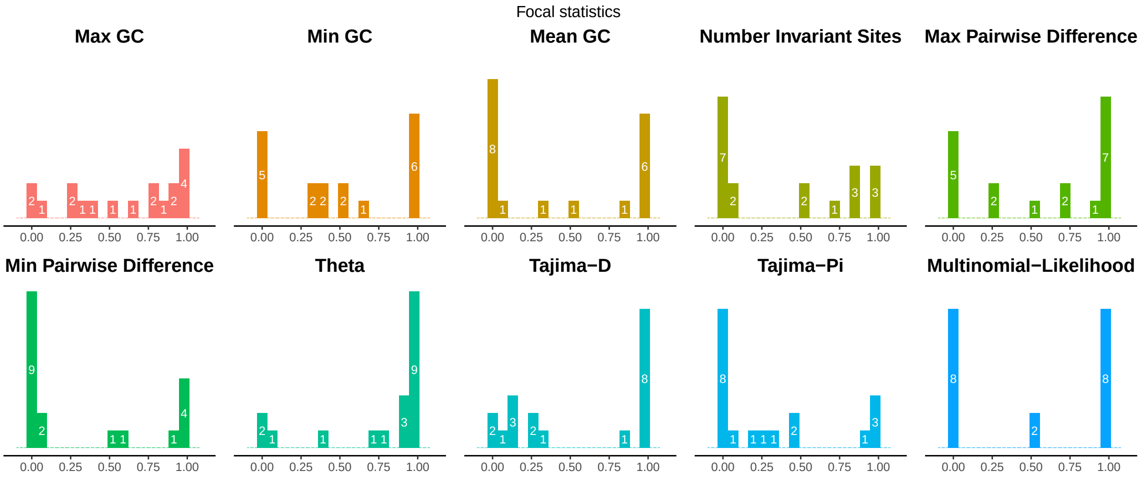


Figure 5.9: Histograms of posterior predictive mid-point p-values for the focal test statistics for the well behaving gene trees.

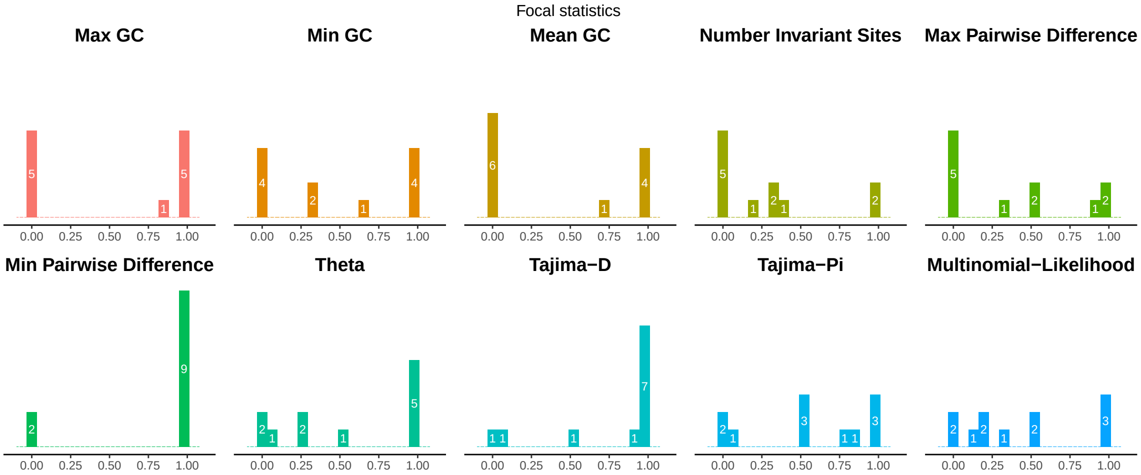


Figure 5.10: Histograms of posterior predictive mid-point p-values for the focal test statistics for the poor behaving gene trees.

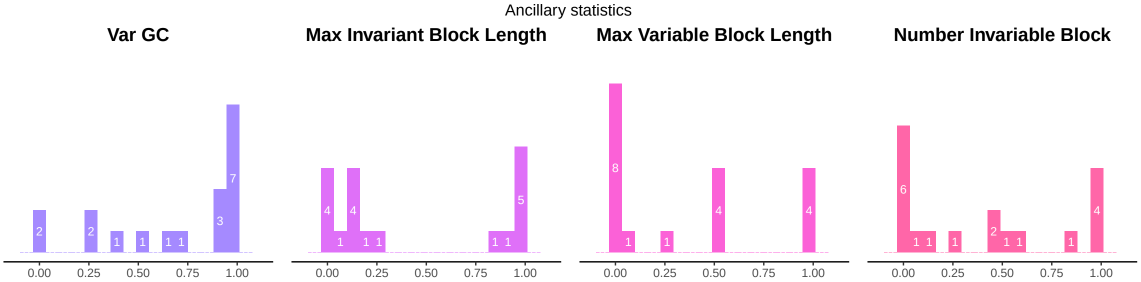


Figure 5.11: Histograms of posterior predictive mid-point p-values for the ancillary test statistics for the well behaving gene trees.

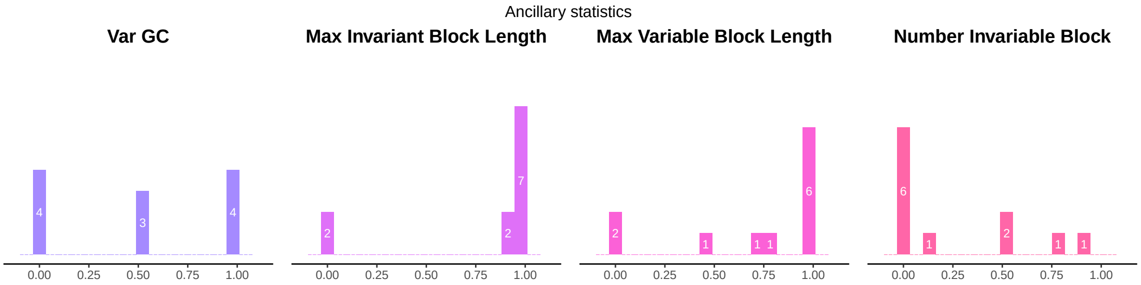


Figure 5.12: Histograms of posterior predictive mid-point p-values for the ancillary test statistics for the poor behaving gene trees.

Appendix A

Supplementary Information Chapter 2

A.1 Precision of an estimator to assess sufficiently many samples

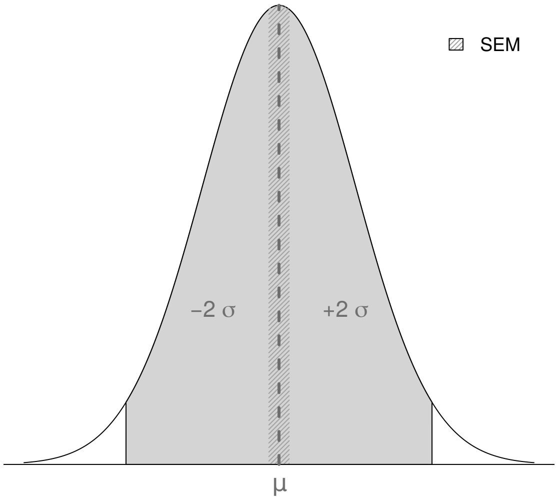


Figure A.1: Schematic of a posterior distribution with mean estimate μ and 95% credible interval. Here, we chose a normal distribution to represent the posterior distribution. The shaded area shows the 95% credible interval, which has a width of ± 2 σ. The dashed line represents the mean (μ) of the distribution. The darker shaded area represents the standard error of the mean. If the SEM is smaller or to 1% of the 95% credible interval size, then we accept the mean estimate as sufficiently precise.

A.2 ESS Estimates for Independent Monte Carlo Samples

We assessed the ability of `CODA`, `MCMCSE` and `Tracer` to estimate the true Effective Sample Size (ESS) from an independent sample. This represents the case where we thinned our MCMC samples and all values are virtually independent. Furthermore, in this first test case we were interested in whether the shape of the distribution had an impact on the efficiency of the method. Thus, we used a normal distribution with mean 0 and standard deviation 1, a lognormal distribution with mean 0 and standard deviation 1 (both on the log scale) and an exponential distribution with rate 1. For each distribution we simulated 1,000 replicates of $N = \{100, 200, 300, 400, 500, 625, 800, 1000\}$ independent and identically distributed (*iid*) samples (i.e., N is equal to the true ESS). Then, we estimated the ESS of these simulated samples using `CODA`, `MCMCSE` and `Tracer`.

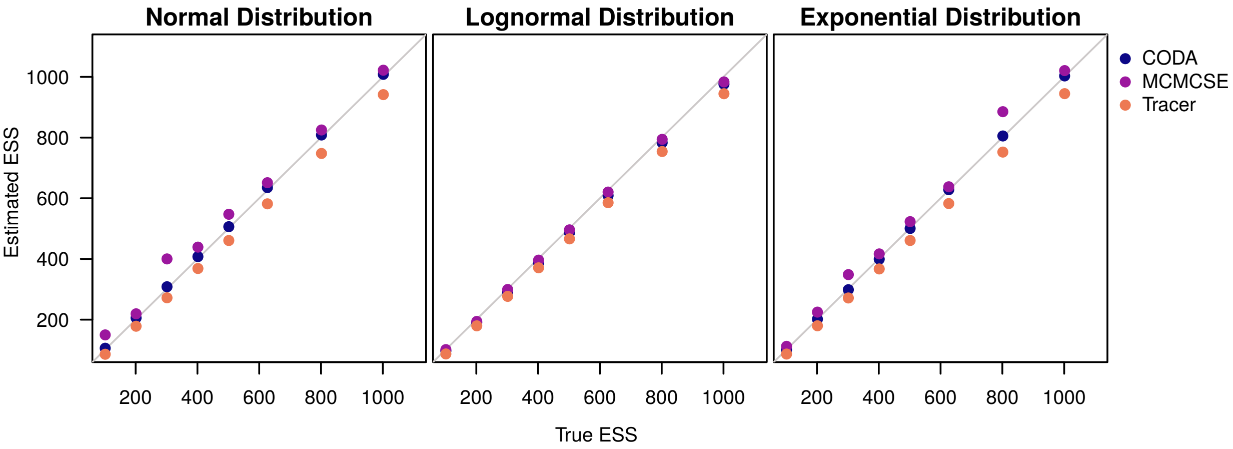


Figure A.2: Estimated Effective Sample Sizes (ESS) for independent samples from different continuous distributions. We used `CODA`, `MCMCSE` and `Tracer` to estimate the ESS and evaluate their accuracy. The x-axis is the true ESS used to generate the sample, the y-axis is the average estimated ESS. The samples were simulated under a normal, lognormal and exponential distribution (from left to right).

All three methods perform comparably well and appear sufficiently precise and robust (Figure A.2). Neither the number of samples nor the choice of the shape of the underlying distribution impacted the accuracy. Thus, we will only use the normal distribution in the following experiments. `Tracer` always had the lowest estimated ESS and thus is the most conservative of the three methods. `CODA`, on the other hand, had the highest overall precision.

A.3 ESS Estimation with MCMCSE

Figure 1 (central panel) shows that for an autocorrelation time (ACT) of $\tau = 50$, MCMCSE overestimates the ESS. To better understand this unusual behavior, we explored new values of τ closer to 50. The simulation was done in the same fashion as for Figure 1. We simulated 1000 replicates for ESS values between $N = \{100, 200, 300, 400, 500, 625, 800, 1000\}$ and ACT varying between $\tau = \{10, 20, 30, 40, 45, 50, 60, 65, 70\}$. For $\tau = \{50, 60, 65\}$ and $N > 200$ the ESS is overestimated (Figure A.3).

`MCMCSE` uses as default the batch means approach to estimate the autocorrelation time and, consequently, the ESS. Such approach has been shown to not consistently estimate the variance when the number of batches is fixed [58, 57, 44]. This could explain why we observed overestimation of the ESS for some ACT values. Further investigation of this behavior should be performed. For now, we advise users to avoid the batch mean approach to estimate the ESS.

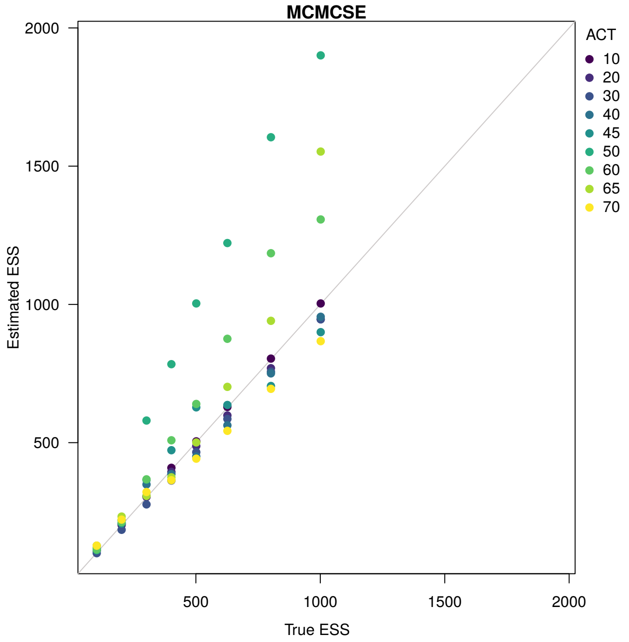


Figure A.3: Estimated ESS from `MCMCSE` for independent samples from a normal distribution with a known ACT. The autocorrelation varied between $\tau = \{10, 20, 30, 40, 45, 50, 60, 65, 70\}$ and is represented by the colored dots. The x-axis is the true ESS used to generate the sample, the y-axis is the average estimated ESS.

A.4 ESS Estimation of Samples from a Binomial Distribution

Figure A.2 shows the estimation of the ESS for a normal, lognormal and exponential distribution. The same tests were performed for the binomial distribution for different p and the results are shown in Figure A.4. For each value of p and sample size, we replicated 1000 test to calculate the mean estimated ESS. The estimation of the ESS is not influenced by the probability of success(p) from the binomial distribution, since all plots in Figure A.4 exhibit the same tendency.

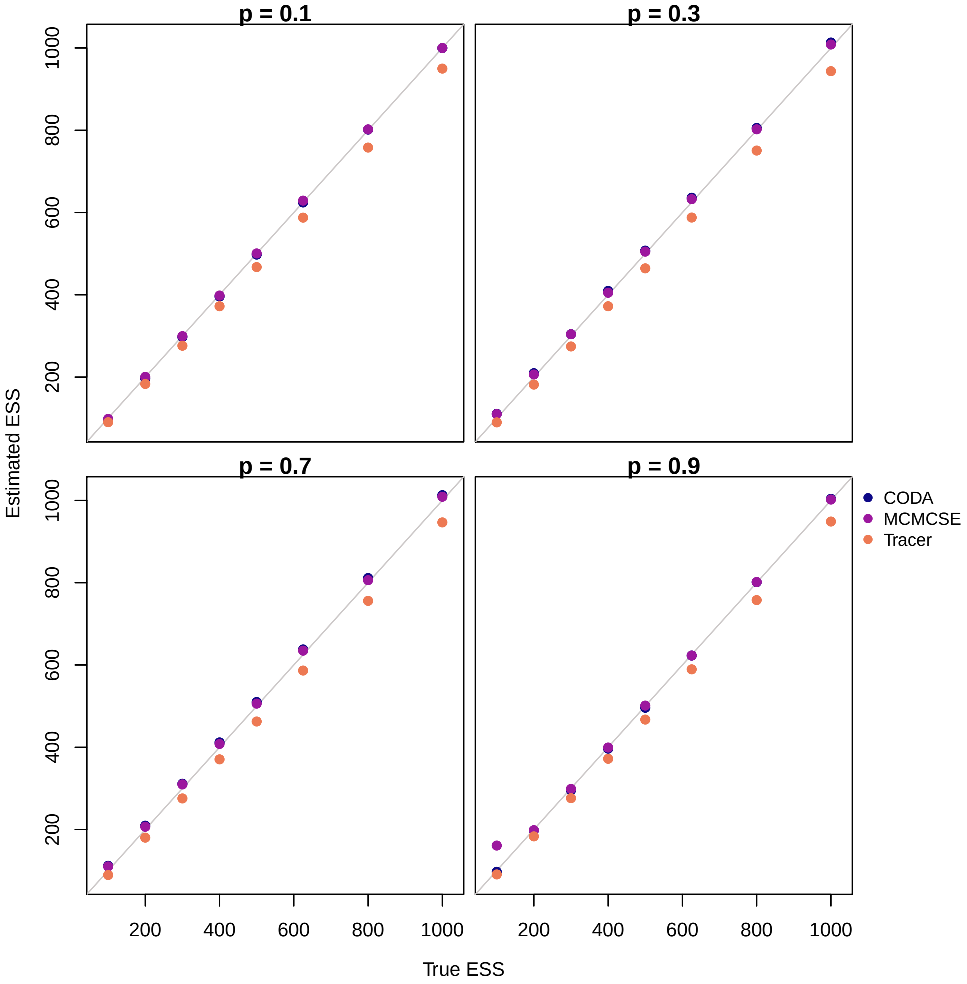


Figure A.4: The estimated Effective Sample Size (ESS) for samples drawn from a binomial distribution with different probabilities of success (p). The p is indicated on top of each plot. For each plot we varied the size of the sample drawn (true ESS) from 100 to 10,000. We then calculated the ESS with three methods: CODA, MCMCSE and Tracer. The blue dots correspond to CODA, the purple dots to MCMCSE and the orange dots to Tracer.

A.5 ESS of splits from samples of trees

Phylogenetic MCMC approaches sample phylogenetic trees and thus splits only indirectly. Different splits are not independent of another as some splits are mutually exclusive and cannot co-occur. Here we investigated if the independence assumption of splits is problematic. Instead of simulating the splits with a know ESS as in the main text, we simulated draws of trees with known ESS. We used unrooted trees with 4, 5 or 6 taxa. We assumed an equal probability of $\frac{1}{n}$ for each tree. Note that this does not imply that each split had equal probability because uneven splits with one large and one small clade are consistent with more trees.

As before, we simulated samples with true ESS values of $N = \{100, 200, 300, 400, 500, 625, 800, 1000\}$ for the trees. Additionally, we sampled auto correlated traces using Algorithm 1 with ACT values of $\tau = \{1, 5, 10, 20, 50\}$. We simulated 100 replicates for each combination of N and τ. For each trace of trees we constructed all possible traces of splits. Then, we calculated the mean ESS over all splits and replicates as our main objective here was to evaluate the bias of estimating ESS values.

Figure A.5 shows the mean ESS value for different combination of N and τ. The results match our observations of our previous tests on autocorrelated continuous traces (Figure 1). The mean estimate is slightly underestimated across the different number of taxa and thus a slightly conservative estimate. Reassuringly, the ESS estimates of splits are a good proxy for the ESS of trees.

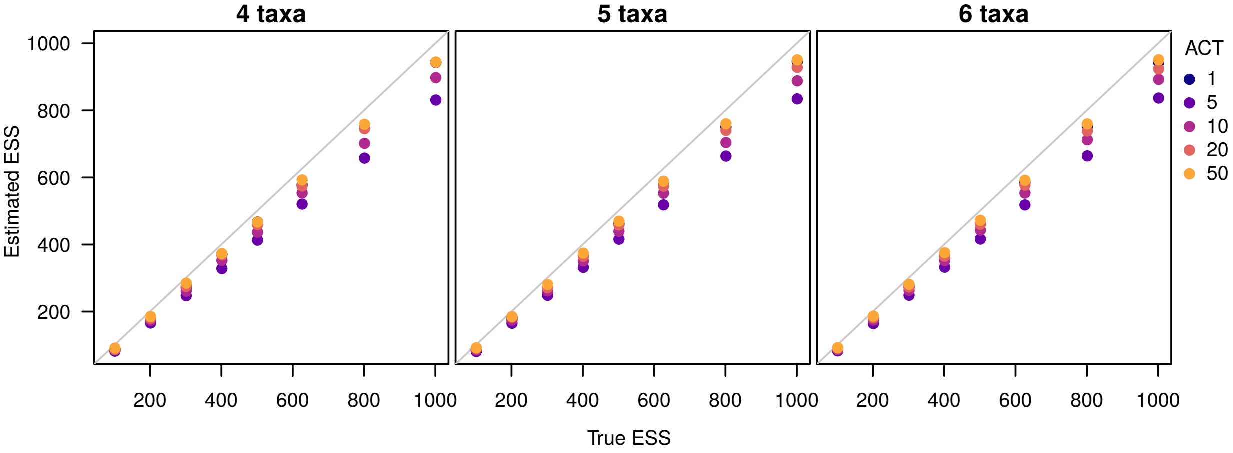


Figure A.5: The mean estimated ESS for all splits over 100 replicates of simulated trace of trees. The true ESS corresponds to $N = \{100, 200, 300, 400, 500, 625, 800, 1000\}$. The autocorrelation assumed values of $\tau = \{1, 5, 10, 20, 50\}$. The left panel displays the results for the case of 4 taxa, the middle panel for 5 taxa and the right panel for 6 taxa.

A.6 Potential Scale Reduction Factor (PSRF)

In phylogenetics, the only (or at least most widely) used convergence assessment method for continuous parameters between replicated MCMC runs is the potential scale reduction factor [PSRF, 53]. We tested the PSRF under different common statistical distributions (normal distribution, gamma distribution and lognormal distribution). In practice, we never know the true shape of the posterior distribution when using empirical data. It is possible that the posterior distribution is well approximated by a normal distribution, but it is similarly possible that the posterior distribution is skewed and thus better approximated by a lognormal or gamma distribution.

We generated two sets of values, according to the same distribution, to compare the performance of evaluating the variance within each set and between sets. Each set had 1×10^5 values drawn from the specified distribution. The distributions used were the normal, lognormal, exponential and gamma. For each distribution, we used different values of variance (1.0, 4.0, 25.0, 100.0) to compare the behavior of the PSRF when the variance of the set of values changes. For the normal and exponential distributions the PSRF values varied from 1.001 to 1.019 as we increased the size of the samples. The values for both distribution had a maximum difference of 0.01 and increasing the variance of the samples did not affect the estimated PSRF values. In the tests for the lognormal and gamma distribution, the PSRF values varied according to the variance of the sample, as it is shown in Figure A.6

Surprisingly, we observed that the PSRF does not converge towards 1.0 with increasing sample size if the distribution is heavily skewed (Figure A.6). For example, when we simulated values from a gamma distribution with $\theta = 10$ and varied κ, PSRF values were comparably high (PSRF > 1.05). Similarly, when we used simulated values from a lognormal distribution, PSRF values never converged towards 1.0 and the asymptotic PSRF increased with higher variances. Thus, we conclude that the PSRF is not universally applicable in phylogenetics and other approaches, such as the Kolmogorov-Smirnov test, are superior.

Figure A.6 shows the PSRF values calculated for 3 different distributions: normal, lognormal and exponential. The mean PSRF was calculated by comparing 1×10^5 times two sets of values drawn from the same distribution with equal parameters. We used 4 different variance values (1.0, 2.0, 5.0 and 10.0) for each distribution to compare how they affect the PSRF estimate.

A.7 Estimating burn-in length

Most recorded samples from an MCMC algorithm in phylogenetics do not start with a random draw from the posterior distribution. Instead, most MCMC algorithms are either initialized with fixed starting values [151, 16] or random values drawn from the prior distribution [74]. It is therefore necessary to remove the first $X\%$ of samples as burn-in to obtain an unbiased approximation from the posterior distribution. In phylogenetics,

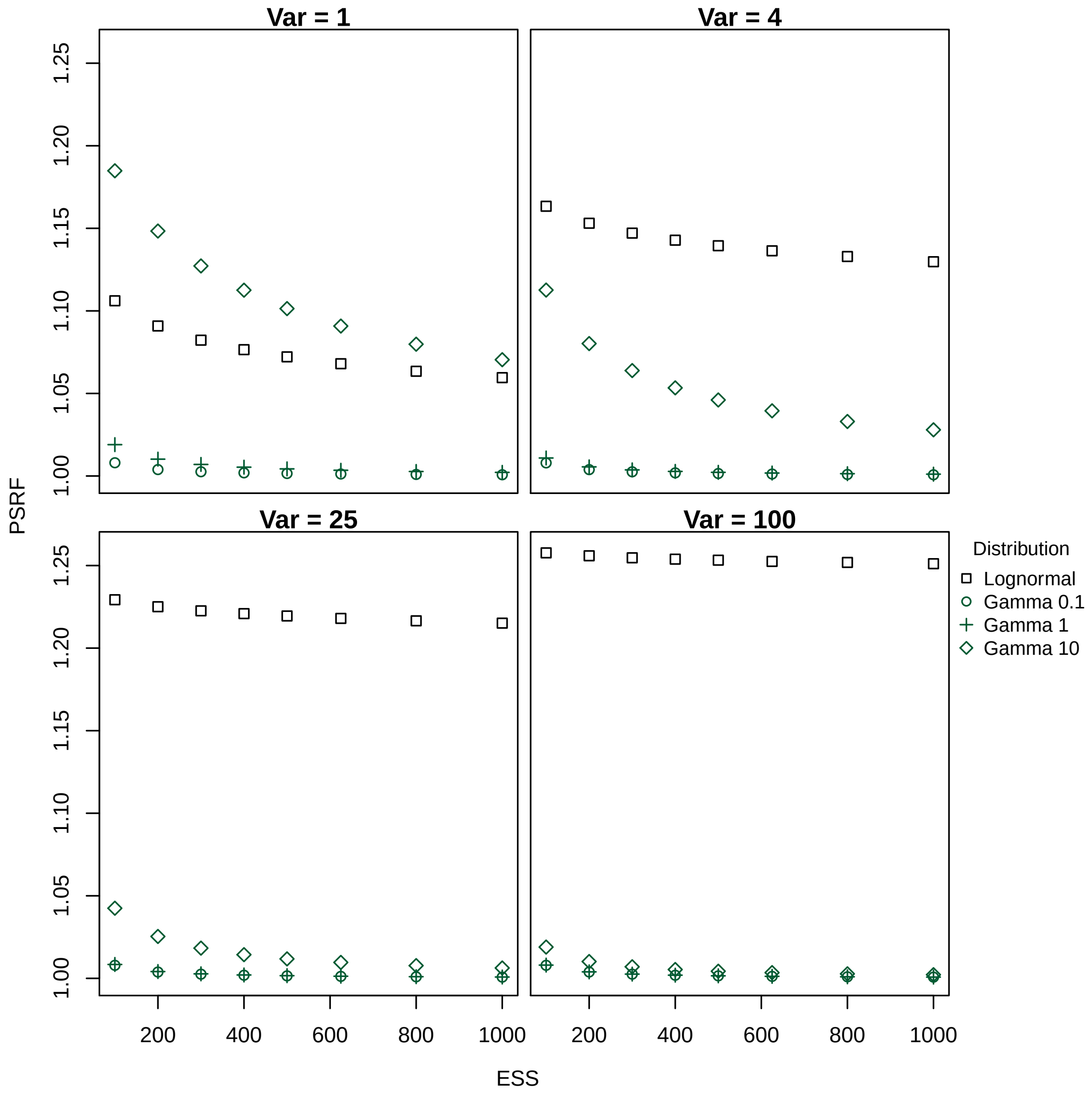


Figure A.6: The potential scale reduction factor (PSRF) values calculated for two samples drawn under the same distribution. The variance of the distributions assumed values of 1, 4, 25 and 100, as is shown on top of each plot. The plotted values correspond to the mean of 1×10^5 replicates. The black squares represent the values calculated from a lognormal distribution. The green figures correspond to the Gamma distribution. The circle, the cross and the diamond correspond to scales parameters of 0.1, 1 and 10, respectively.

a common burn-in is either 10% or 25%, depending on the arbitrary preference of the software developer.

To make the burn-in selection slightly less arbitrary, we developed the following procedure. We search for the optimum burn-in value defined as the lowest burn-in that passes the convergence tests. We start with no burn-in and increase it by 10% up to a maximum of 50% of the chain. If more than 50% of the chain has to be discarded, the MCMC spent too much time outside of the stationary distribution and the whole analysis should be redone and/or run longer.

Algorithm 4 Estimating the burn-in length.

```
Inputs:
        $X$: the samples.

Initialize:
          $n \leftarrow$ length($X$)                    // the total number of correlated samples

for $i$ in $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ do              // generate independent chains
    $Y \leftarrow X[(i*n):n]$                             // retrieve the post-burnin samples
    if convergence($Y$) == pass then                          // check for convergence
        break                                                          // stop
    end if
end for

return $i$
```

Appendix B

Supplementary Information Chapter 4

B.1 Variable sites in simulated data sets

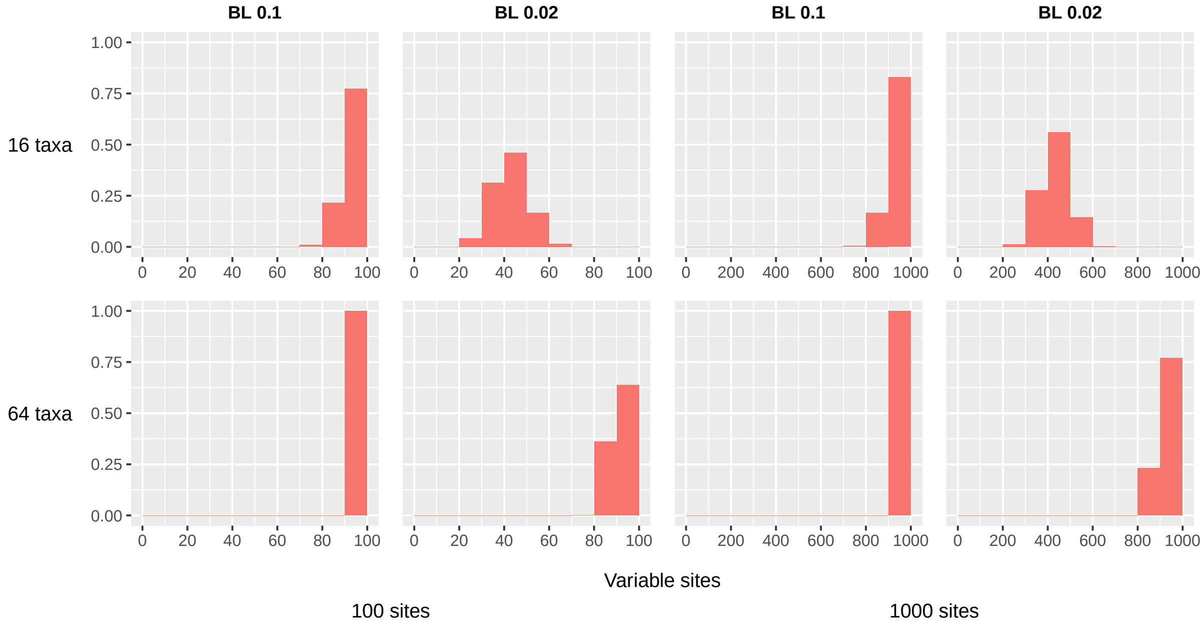


Figure B.1: Histogram of relative number of variable sites in the data sets for each simulation scenario. The data were simulated in `RevBayes` [74]. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

B.2 Credible interval for the Tree Length

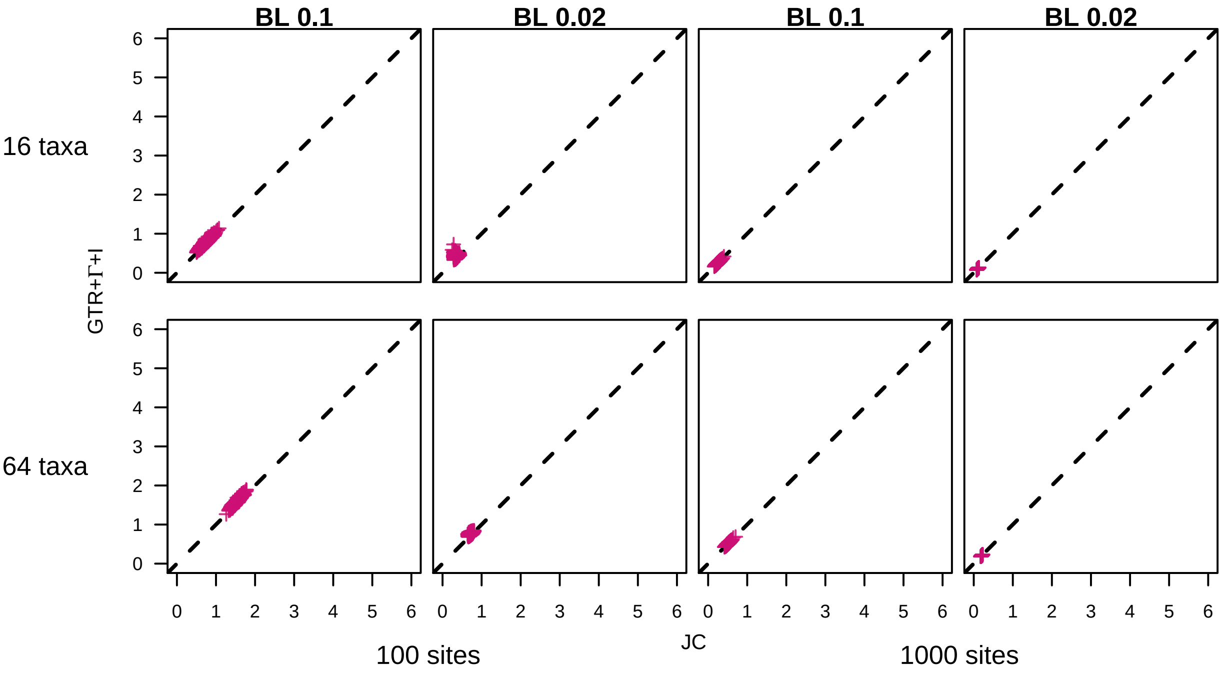


Figure B.2: The 95% credible interval for the tree length for the inference with Jukes-Cantor [JC, 86] against the inference with GTR+Γ+I [165, 177, 62] under the `Tame` prior. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-parametrized substitution model. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

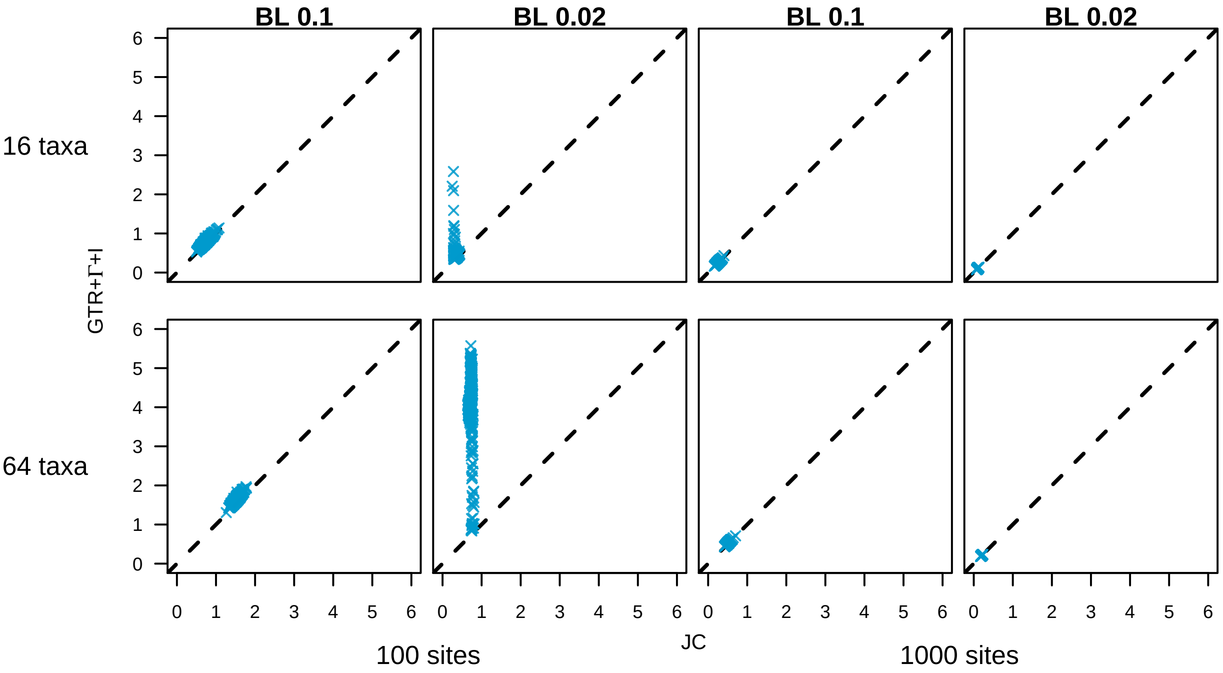


Figure B.3: The 95% credible interval for the tree length for the inference with Jukes-Cantor [JC, 86] against the inference with GTR+Γ+I [165, 177, 62] under the `MrBayes` prior [151]. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-parametrized substitution model. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

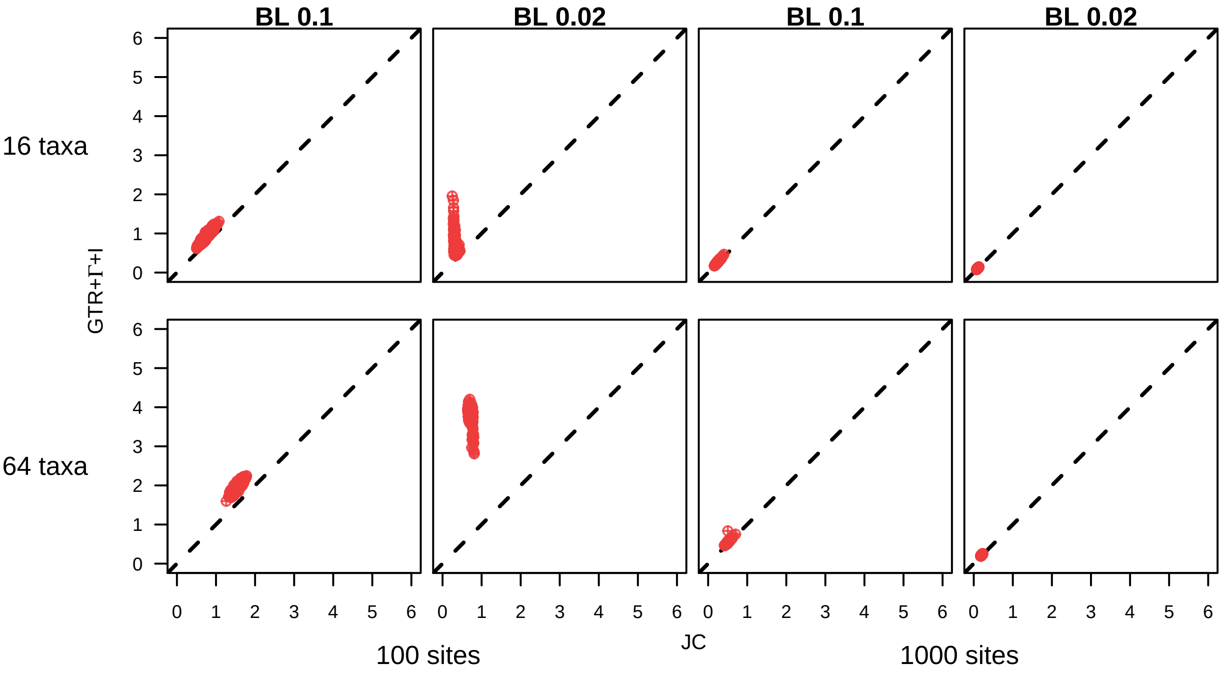


Figure B.4: The 95% credible interval for the tree length for the inference with Jukes-Cantor [JC, 86] against the inference with GTR+Γ+I [165, 177, 62] under the `RevBayes` prior [73]. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-parametrized substitution model. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

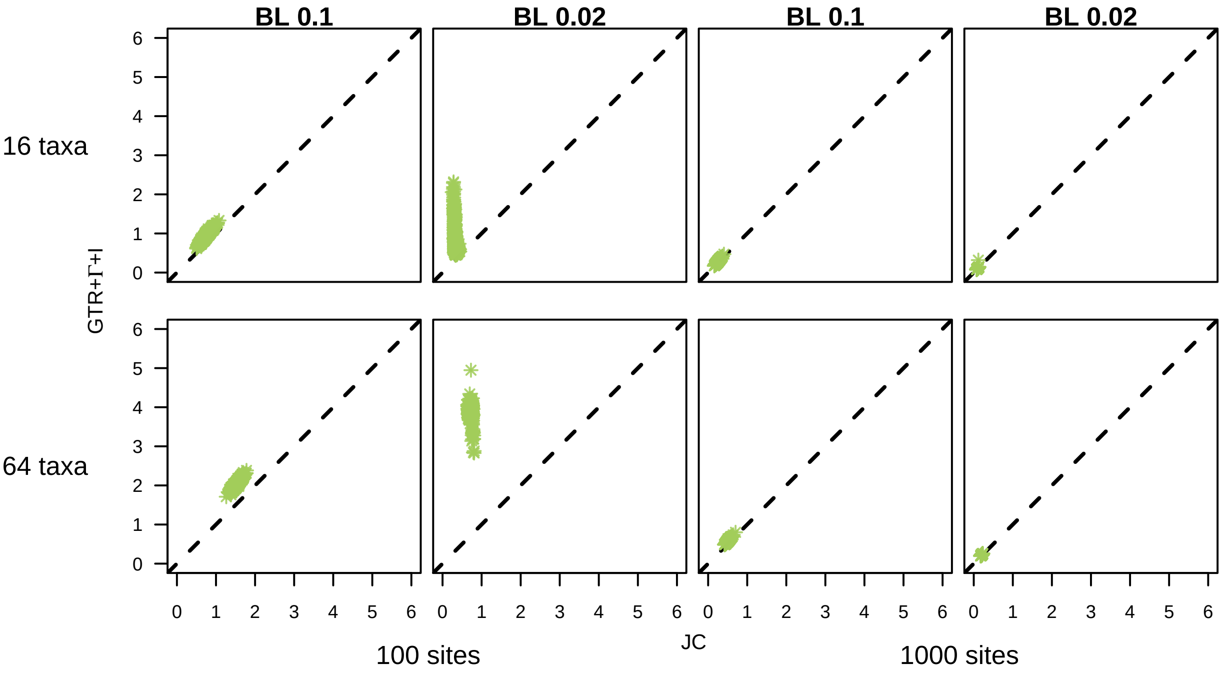


Figure B.5: The 95% credible interval for the tree length for the inference with Jukes-Cantor [JC, 86] against the inference with GTR+Γ+I [165, 177, 62] under the `BEAST` prior [17]. On the x-axis we show the estimated 95% credible interval size for the JC substitution model (true model). On the y-axis we plot the estimated 95% credible interval for the over-parametrized substitution model. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

B.3 Mean Tree Length for 100 sites and short branch lengths

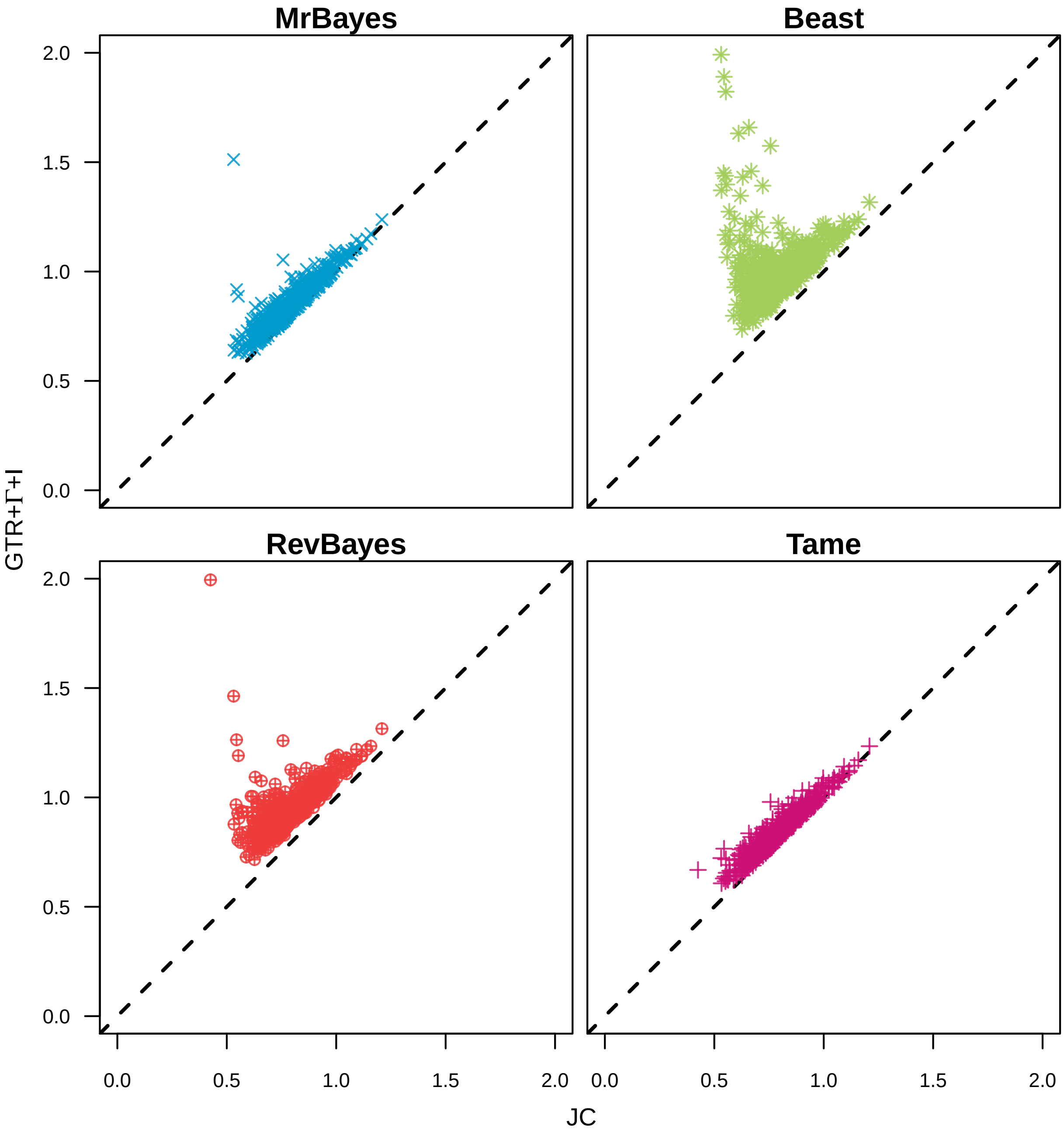


Figure B.6: Comparison of mean tree length between [JC, 86] and GTR+Γ+I [165, 177, 62] for the data sets with 16 taxa, 100 sites and mean branch length 0.02. The x-axis represents the mean tree length for the inference under JC, while the y-axis represents the mean tree length for the inference under GTR+Γ+I. Each plot displays the means for the inference with the four different prior schemes.

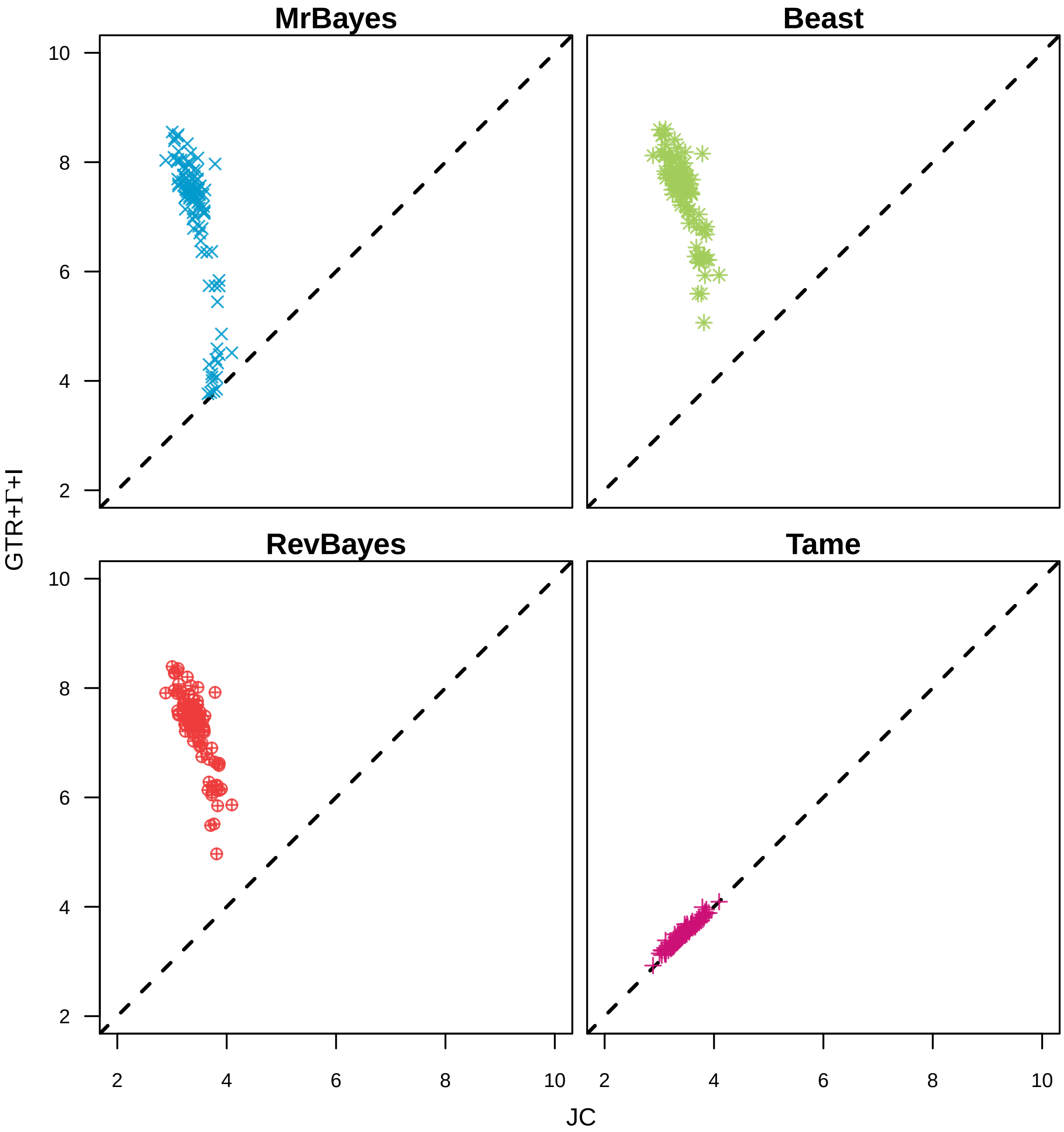


Figure B.7: Comparison of mean tree length between [JC, 86] and GTR+Γ+I [165, 177, 62] for the data sets with 64 taxa, 100 sites and mean branch length 0.02. The x-axis represents the mean tree length for the inference under JC, while the y-axis represents the mean tree length for the inference under GTR+Γ+I. Each plot displays the means for the inference with the four different prior schemes.

B.4 Further exploration of model combinations

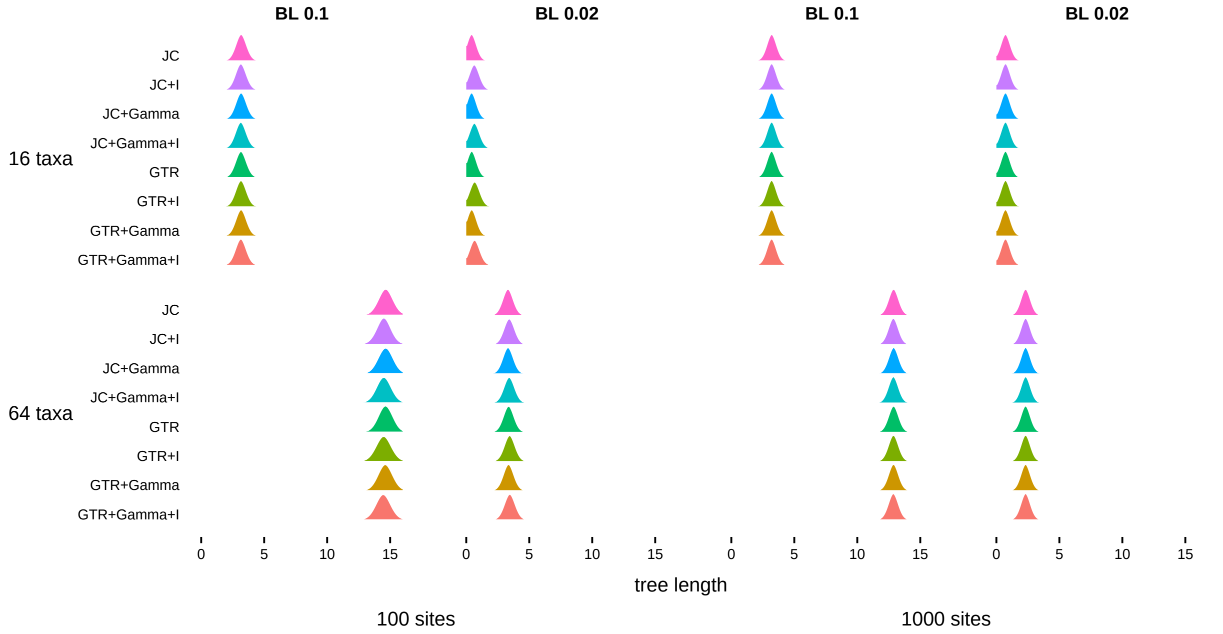


Figure B.8: Posterior distributions for tree length for one example of each data set. Each line corresponds to a different model (JC, JC+I, JC+Γ, JC+Γ+I, GTR, GTR+I, GTR+Γ, GTR+Γ+I; [86, 165, 177, 62]). The GTR models followed the `Tame` prior setting. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

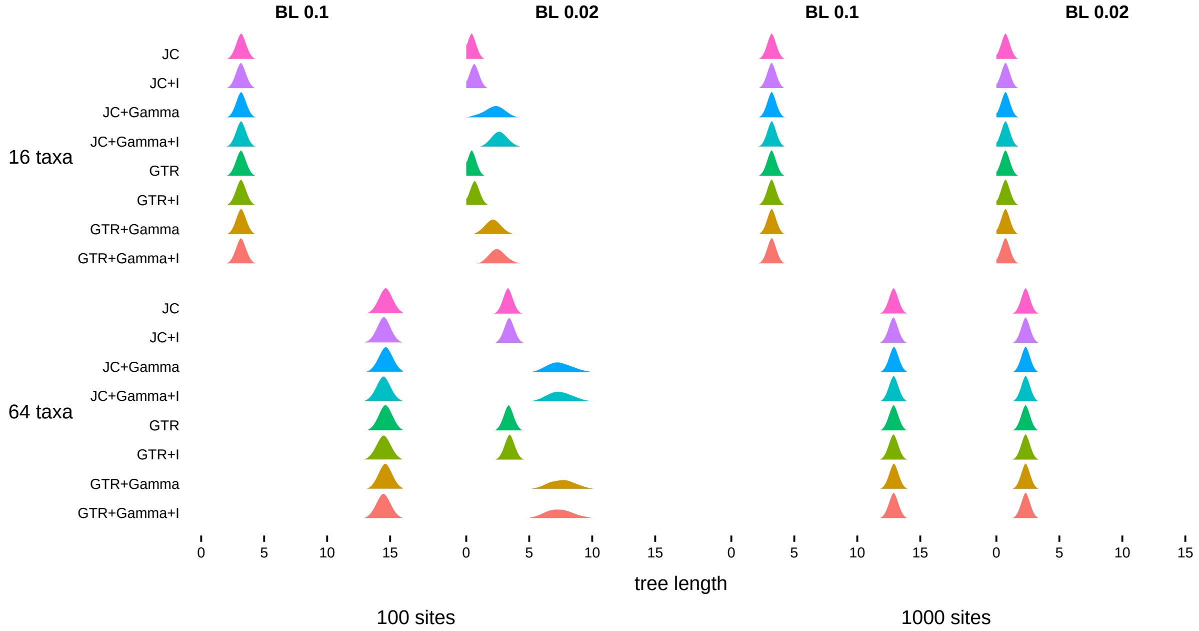


Figure B.9: Posterior distributions for tree length for one example of each data set. Each line corresponds to a different model (JC, JC+I, JC+Γ, JC+Γ+I, GTR, GTR+I, GTR+Γ, GTR+Γ+I; [86, 165, 177, 62]). The GTR models followed the `MrBayes` prior setting. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

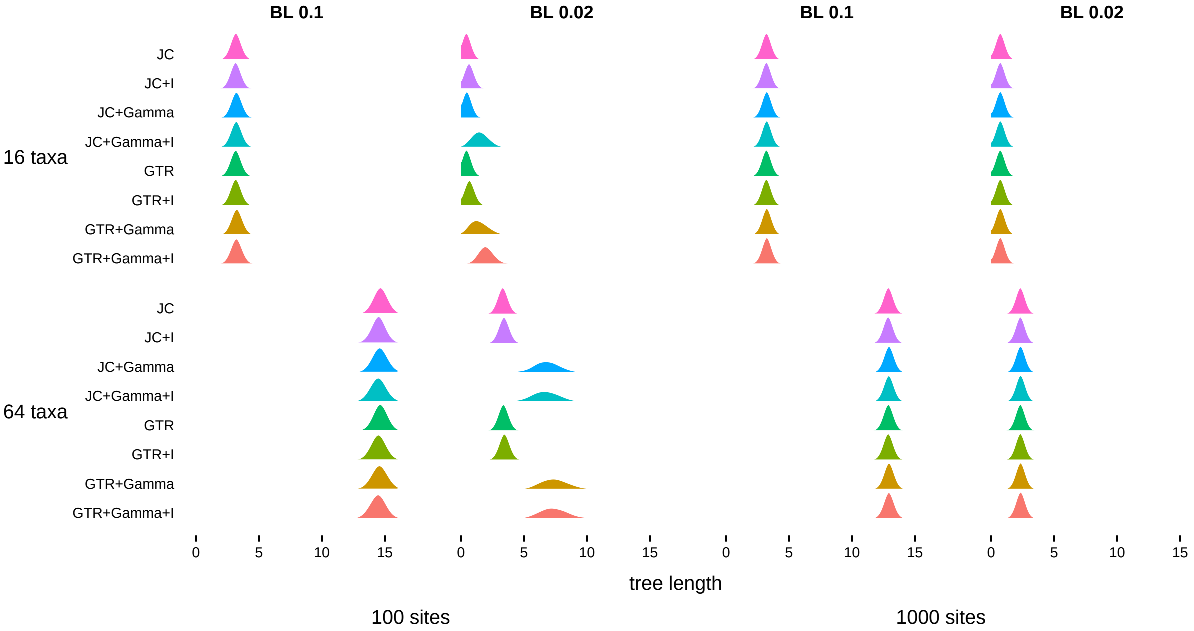


Figure B.10: Posterior distributions for tree length for one example of each data set. Each line corresponds to a different model (JC, JC+I, JC+Γ, JC+Γ+I, GTR, GTR+I, GTR+Γ, GTR+Γ+I; [86, 165, 177, 62]). The GTR models followed the `RevBayes` [73] prior setting. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

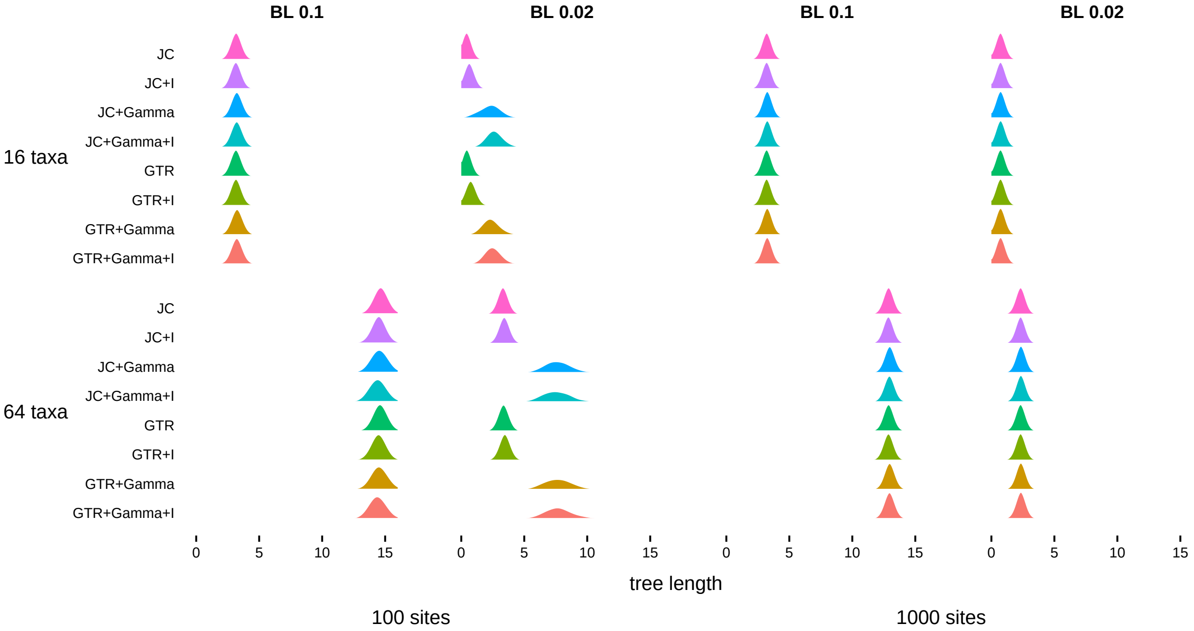


Figure B.11: Posterior distributions for tree length for one example of each data set. Each line corresponds to a different model (JC, JC+I, JC+Γ, JC+Γ+I, GTR, GTR+I, GTR+Γ, GTR+Γ+I; [86, 165, 177, 62]). The GTR models followed the `BEAST` [17] prior setting. The first row corresponds to the simulated trees with 16 taxa, while the second row corresponds to the simulated trees with 64 taxa. The mean branch lengths (BL) for the data sets are on top of each column. The two first columns display the data sets with 100 sites, the two other columns show the data sets with 1000 taxa.

B.5 Convergence assessment example

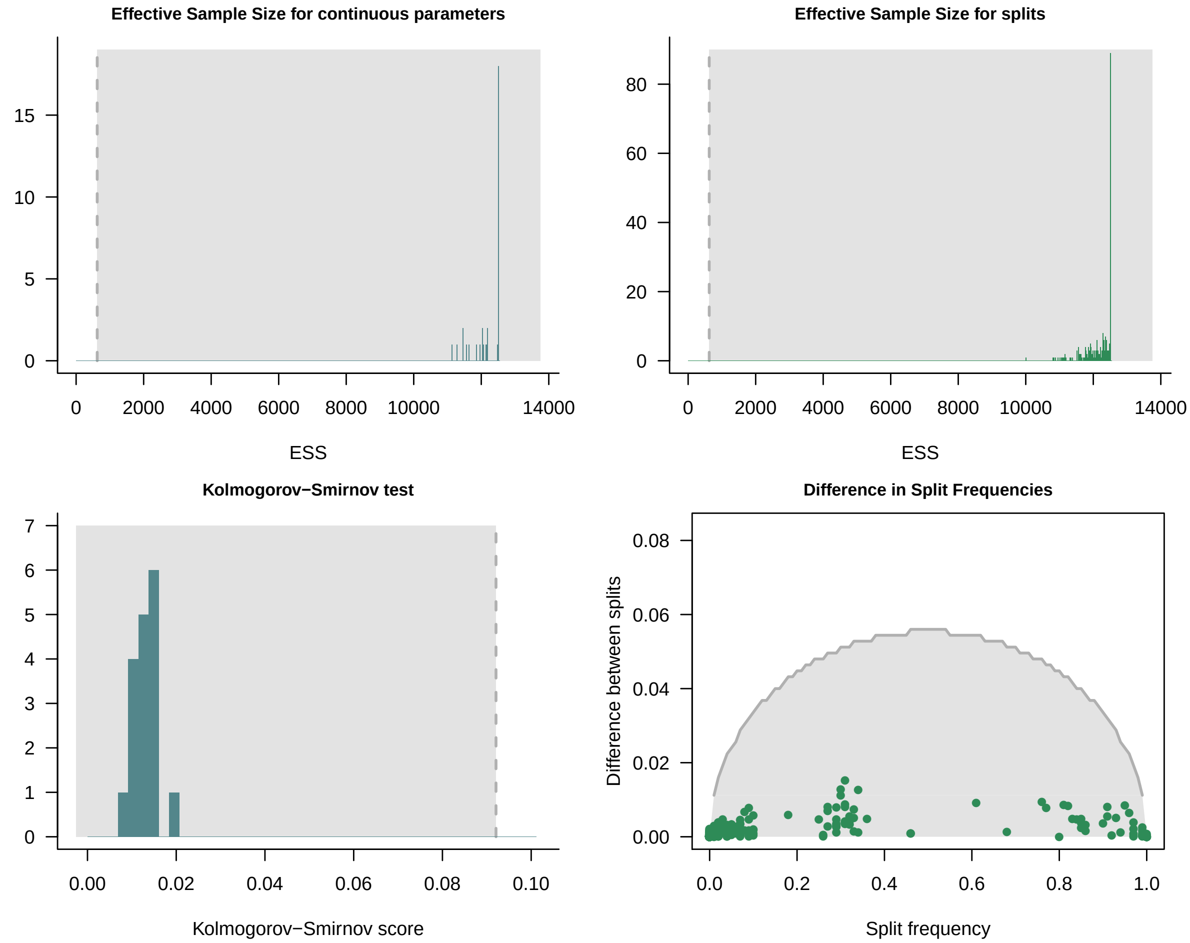


Figure B.12: Convergence assessment plots using the package `Convenience` [38] for one example MCMC from Figure 7. The first column shows the plots for the assessment of convergence for the continuous parameters. The second column displays the assessment of the splits in the tree. In all plots the values are within the gray shaded area, which indicates that convergence was achieved.

Appendix C

Lessons Learned from Organizing and Teaching Virtual Phylogenetics Workshops

C.1 Abstract

In 2020 and 2021, the COVID-19 pandemic led to an abrupt overhaul of many academic practices, including the transition of scientific events, such as workshops, to a fully virtual format. We describe our experiences organizing and teaching online-only statistical phylogenetics workshops and the lessons we learned along the way. We found that online workshops present some specific challenges, but format choices and rigorous planning can alleviate many of the concerns typically associated with a virtual medium. In addition, online workshops have unique advantages such as the flexibility they offer to participants and instructors and their accessibility to non-traditional and underprivileged audiences. We hope that our experience will encourage workshop organizers to consider online-only events as an integral part of potential training opportunities rather than simply a stop-gap solution for unusual circumstances. In addition, we hope to prompt broader discussion about integrating aspects of online workshops into traditional in-person courses to make training opportunities more flexible and inclusive.

C.2 Introduction

Phylogenetic analysis of biological data often requires a high level of expertise not only in the statistical framework underlying applied models and approaches, but also in the specific software implementations and their wide range of available options. This, in turn, leads to a high barrier to entry for researchers interested in using phylogenetic programs and packages. As a result, developer teams spend considerable effort creating materials and opportunities for new users to learn how to use complex software tools so that they can apply phylogenetic methods to their own data. Workshops are perhaps the most common mechanism used by scientific software developers to expand their user base and provide expert training to empiricists. These events are an opportunity for scientists to directly interact with the developers and obtain deeper insight into the software. At the same time, these short courses also enable developers to learn more about the needs of users working with empirical data. Moreover, many software developers gain valuable experience in teaching and pedagogy as instructors in hands-on workshops. Participants and instructors recognize the value these experiences can have in improving software, building the knowledge base of scientists at all levels, and creating opportunities for networking that often lead to fruitful collaborations.

This work focuses on workshops dedicated to RevBayes [76], a broadly used Bayesian phylogenetic software tool that enables inference of evolutionary parameters under complex, hierarchical models. The RevBayes developer team provides extensive, publicly available documentation and user tutorials for a wide range of analyses and applications via the project website[1]. Since 2013, RevBayes has been featured in over 40 workshops[2], either as standalone events, or part of more general courses, such as the Woods Hole Workshop on Molecular Evolution[3] and the Bodega Bay Workshop in Applied Phylogenetics[4].

In early 2020, the onset of the COVID-19 pandemic required instructors to cancel in-person workshops and innovate ways to deliver training materials to practitioners [116, 142, 6]. The majority of workshop participants are early career researchers, many of whom attend workshops to deliberately meet planned professional goals, such as attaining skills to complete dissertation research or seeking out postdoctoral research opportunities. Thus, a year without workshop opportunities may be a significant setback to many scientists early in their training. Rather than canceling all of our planned workshops, the RevBayes team opted to transition to fully online events, and we have recently completed two so-called "Stay-at-Home RevBayes" workshops. Our experiences and the feedback from participants have been very positive, and we believe that this format has unique advantages and a few challenges when compared to traditional, in-person workshops.

This paper describes our experience organizing the Stay-at-Home RevBayes online courses, explains the rationale behind some of our choices, and provides suggestions for future workshop organizers. Our goal is to share our experience organizing and teaching a technical software workshop in an online format, as well as demonstrate some of the advantages and challenges of such a course. In particular, we believe that online-only events remain relevant beyond the specific context of the pandemic, and that they should not be dismissed in a rush to get back to previous practices. Furthermore, as we transition back to planning in-person activities, we hope to stimulate discussions among the developers of phylogenetic methods on new approaches for enhancing workshop experiences and inclusivity, while creating broadly accessible learning opportunities.

[1] The RevBayes Project Website: `http://revbayes.com`

[2] RevBayes Workshops: `http://revbayes.com/workshops`

[3] Workshop on Molecular Evolution, Woods Hole, MA, USA: `https://molevolworkshop.github.io`

[4] Workshop in Applied Phylogenetics, Bodega Bay, CA USA: `http://treethinkers.org`

C.3 The Stay-at-Home RevBayes Workshops

The primary goal of all RevBayes workshops is to provide participants with a solid foundation in the theory and application of phylogenetic methods—as well as practical knowledge of the software implementation—so that they will be able to analyze their own data using complex models and Bayesian statistics. To achieve this goal, the RevBayes team has developed a rich library of tutorials[5] providing extensive details about various phylogenetic analyses. When presenting this material in an in-person setting, we are often constrained by time and only able to spend a couple of hours on each topic during a five- to seven-day workshop. However, a virtual course offers the opportunity to spread the material over several weeks, enabling participants to work at their own pace and review what they have learned before moving on to the next tutorial. Thus, the format of the Stay-at-Home RevBayes Workshops included a mix of synchronous meetings (using the Zoom video-conference service), detailed tutorials and pre-recorded videos, and real-time discussions via Slack (an online instant messaging platform), all spread out over five to six weeks (we discuss the communication tools used in more detail in Sections C.3.2 and C.4.3). An overview of the core workshop components is provided in Box 1.

[frametitle=Box 1: Overview of the main components of the Stay-at-Home RevBayes Workshops, skipabove=, skipbelow=, roundcorner=5pt, linewidth=0.5pt, frametitlerule=true, frametitlebackgroundcolor=gray!30]

- *Course website*[67]: The workshop description, application link, schedule, and materials are provided on a public website for each course.
- *Introductory synchronous session (Zoom)*: Participants and instructors introduce themselves, then instructors give an orientation on the workshop format and procedures, offer an overview of RevBayes and the Rev language, and check that all participants succeeded in installing the required software.
- *Introductory lectures*: Participants work through previously published videos providing background on the theory of Bayesian phylogenetics.
- *Asynchronous completion of RevBayes tutorials*: Participants work at their own pace to learn a curated set of methods and analyses in RevBayes (Fig. C.1). Each lesson includes:
  - *Detailed online tutorial*: Each online tutorial provides the theory and background for a specific model or statistical method and a step-by-step explanation of how the corresponding analysis is performed in RevBayes.
  - *Video guide*: Each online tutorial links to a series of videos (hosted on YouTube) created by a RevBayes instructor walking the viewer through each section of the lesson and providing additional details.
- *Communication*: Instructors are available to answer participants' questions and engage in group discussions via the course messaging tool (Slack) and regular office hours (on Zoom).

[5] RevBayes Tutorial Library: `http://revbayes.com/tutorials`

[6] Stay-at-Home RevBayes Workshop Summer 2020: `http://revbayes.com/workshops/online2020.html`

[7] Stay-at-Home RevBayes Workshop Spring 2021: `http://revbayes.com/workshops/online2021.html`

- *Final group synchronous session (Zoom)*: Participants and instructors discuss the course materials, common issues faced during the workshop, and future directions for new methods or applications in Bayesian phylogenetics.
- *One-on-one meetings*: Each participant is paired with an instructor to meet via Zoom and discuss the participant's plan for applying RevBayes to their own data.

C.3.1 Workshop Content

We created a syllabus that included four introductory lectures and eight detailed tutorials. At the start of the workshop, participants learned about the course format, timeline, and content in a synchronous meeting. Additionally, during the first synchronous session, we included a background lecture on RevBayes and the Rev language. Clearly outlining the structure, tools, and course expectations early helps build participant trust and comfort [187], which is key when in an online format or using new tools. It was important to include lectures on basic probability theory and Bayesian phylogenetics—as background knowledge on these topics is required to correctly assess models and inference output in RevBayes—and thus it is fortunate that this material was already available online. In 2018, Paul Lewis recorded a series of lectures entitled "Phylogenetics 101" (or Primer on Phylogenetics)[8] for *Phyloseminar*, an online seminar on phylogenetics topics created by Frederick Matsen in 2009[9]. These lectures begin with topics as fundamental as the definition of conditional probability, and, by building upon that foundation, culminate in the construction of complex phylogenetic models and the assessment of their statistical properties. For the RevBayes virtual workshop, these lectures provided participants with an accessible introduction to (or review of) the core theory in Bayesian phylogenetics.

After completing the introductory material and installing RevBayes, the workshop participants were assigned a series of tutorials. The lessons began with an introduction to Markov chain Monte Carlo (MCMC) in RevBayes and then increased in complexity to include analyses of datasets combining fossil and extant taxa [48, 11], polymorphism aware phylogenetic methods [30, 31, 15], and posterior predictive analysis [71] (Fig. C.1). For each tutorial, we created a video guide (hosted on YouTube) that walked through each step and concept. The videos were time-stamped or recorded in segments so that video links could be placed at each section heading of the online tutorials. For example, in the "Introduction to Posterior Prediction" tutorial[10], each section links to a YouTube video where the tutorial author describes the contents of that section. The video guides emulate how we often walk participants through a tutorial during an in-person workshop, with features like "pause" and "replay" that are not really possible in a synchronous class. During these demonstrations, we are often able to insert practical tips and other topics that might not fit naturally into the written tutorial and thus enhance the content. For instance, we can

[8] Primer on Phylogenetics (YouTube Playlist): `https://www.youtube.com/playlist?list=PLztACvN0g42vSxiQ4tM0sQTddMx-V40LE`

[9] Phyloseminar: `http://phyloseminar.org`

[10] Introduction to Posterior Prediction: `http://revbayes.com/tutorials/intro_posterior_prediction`

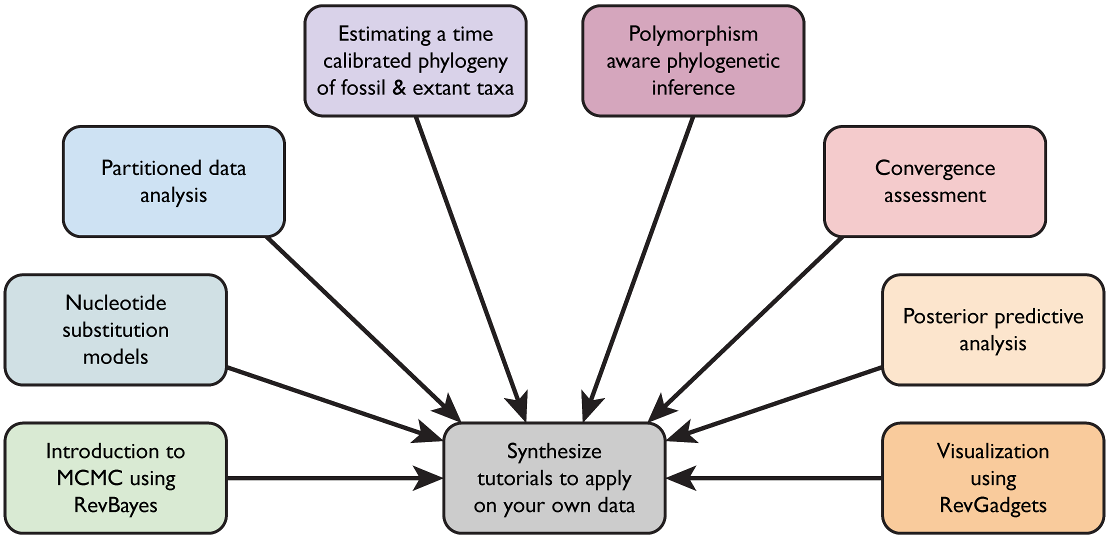

Estimating a time calibrated phylogeny of fossil & extant taxa
Polymorphism aware phylogenetic inference
Partitioned data analysis
Convergence assessment
Nucleotide substitution models
Posterior predictive analysis
Introduction to MCMC using RevBayes
Synthesize tutorials to apply on your own data
Visualization using RevGadgets

Figure C.1: The Stay-at-Home RevBayes Workshop focused on eight core topics, each with a detailed tutorial and accompanying video guide. The goal of the course is to provide enough time for participants to complete the tutorials while considering how the methods will be applicable to their own data and research questions.

remind the audience of the difference between stochastic (*i.e.,* estimated) and constant (*i.e.,* fixed) parameters, which use a different syntax in the Rev language and can be confusing to inexperienced users. The extensive details included in each tutorial may also be somewhat intimidating to new users and the video guides serve as a way to ease learners into the material. Participants were provided with a suggested timeline for completing each component of the course. After completing the set of tutorials curated for the online course, workshop participants were then given time to explore the other tutorials on the RevBayes site or to start analyzing their own data.

The core content created for the Stay-at-Home RevBayes Workshops is accessible to anyone at any time. Thus, researchers are able to work through the tutorials and videos even if they are not part of a workshop. Nevertheless, registering and committing to a course—online or in-person—provides a timeline and structure, as well as access to experts in the field for guidance, and these facilitate completion of learning goals.

C.3.2 Workshop Interactions

Phylogenetics workshops offer participants the unique opportunity to learn methods and software directly from experts and developers. Moreover, these kinds of courses enable researchers from diverse fields and backgrounds to build connections that can often lead to exciting new collaborations. While online workshops do allow attendees to interact via text chats, such spontaneous interactions may not come as easily in a virtual medium—particularly across multiple time zones—as they would when meeting in person. Traditional activities amenable to, or even fostering, spontaneous discussions, such as breaks or meals, must be rethought and deliberately executed. We therefore used a variety of activities and tools (described in detail in this section) to provide participants direct access to instructors and create ways to engage and network with one another.

Prior to the start of the workshop, all participants and instructors were asked to create an introduction slide that was then shown during our first synchronous session (Fig. C.2). All synchronous meetings were held on Zoom[11] and the introductory session provided space for participants and instructors to get to know one another. We used break-out rooms in Zoom to hold small group discussions to enable more casual conversations among participants and instructors. These interactions were also included to help reduce participants' hesitancy to ask questions or request help during the course.

The first synchronous meeting provided a detailed overview of the workshop format and introduced participants to our primary communication tool: Slack[12]. The workshop Slack space included a separate channel for each tutorial, as well as channels for participants to discuss general questions on phylogenetics and Bayesian theory, technical issues (*e.g.,* software installation problems), and the RevBayes interpreted language. Importantly, Slack offered a private communication platform that helped participants feel more comfortable asking questions and a mechanism for sharing links to synchronous Zoom meetings and other course materials. In addition, after the conclusion of each workshop, the associated Slack space remained open for several months, providing the opportunity for participants to refer back to previous answers and discussions, as well as ask follow-up questions.

While the participants worked through the material on their own time, we held regular "office hours" via Zoom (each scheduled for one hour), where they were invited to raise issues and ask questions about the workshop content. In the first edition of the workshop, these meetings were held every week. In the second workshop, synchronous sessions were mirrored because of less time-zone overlap, thus office hours were reduced to every two weeks to avoid overloading instructors.

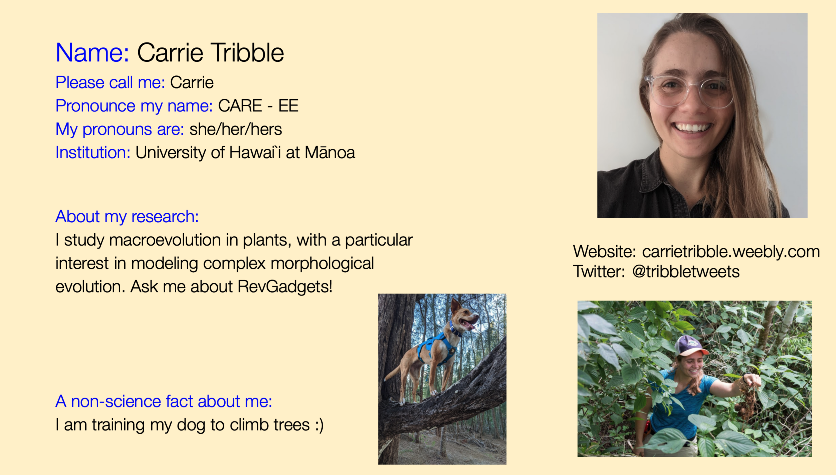
At the conclusion of the multi-week Stay-at-Home RevBayes course, we held a final synchronous session to address remaining questions about the tutorials and discuss RevBayes and Bayesian phylogenetic inference in general. In the first edition of the workshop, this final session was held over several days. Based on feedback from the participants, this session was reduced to two hours in the second workshop.

We then arranged a one-on-one meeting between each participant and an instructor selected based on the participant's specific interests and dataset. The one-on-one meetings allowed participants to troubleshoot analyses applied to their own data under the guidance of a workshop instructor and collaborate to devise creative solutions to unique biological problems. Both participants and instructors found these meetings to be one of the most valuable interactions in the workshop.

In summary, we held scheduled sessions and optional office hours on Zoom and created a Slack space for communication throughout the duration of the course. Additionally, each participant met in a one-on-one meeting with an instructor at the end of the workshop.

[11] Zoom: `https://zoom.us`

[12] Slack: `https://slack.com`



Name: Carrie Tribble

Please call me: Carrie

Pronounce my name: CARE - EE

My pronouns are: she/her/hers

Institution: University of Hawai`i at Mānoa

About my research:

I study macroevolution in plants, with a particular interest in modeling complex morphological evolution. Ask me about RevGadgets!

A non-science fact about me:

I am training my dog to climb trees :)

Website: carrietribble.weebly.com

Twitter: @tribbletweets

Figure C.2: An example of an introduction slide by workshop instructor Carrie Tribble. All instructors and course participants used the same slide template. In the first meeting on Zoom, everyone was able to introduce themselves using their slide.

We believe that all of these elements have important and non-overlapping roles. In our experience, questions raised on the Slack forum tended to be shorter and more narrowly focused on the workshop material, such as technical issues or specific analysis choices in the tutorials. Synchronous sessions attracted broader, more open-ended questions and provided an opportunity for instructors to discuss general guidelines, best practices, or exciting future directions for methods development. Finally, the one-on-one meetings ensured that all participants left the workshop with actionable advice on how to apply the teachings on their own datasets, even if they did not feel comfortable raising questions in front of the whole group.

C.3.3 Flipping the Workshop Format

In our experience, the intense schedule of most in-person workshops is very tiring for both instructors and participants, making it difficult for some participants to complete all the activities and tutorials. Even when all activities are completed, an extremely heavy schedule can lead to lower understanding and long-term retention of important concepts. Since online workshops are not constrained by the physical presence of participants at the venue, it was easier to extend the workshop schedule to run over several weeks and develop material amenable to a flipped-workshop format.

A flipped-classroom format [92, 97, 130]—where lectures and tutorials are pre-recorded and synchronous sessions can be used for questions and discussion—was an optimal approach for several reasons. First, it is widely acknowledged that online meetings require more focus and are more tiring than in-person meetings [leading to so-called "Zoom fatigue"; 8]. Therefore, we limited synchronous sessions to material that could not be covered in other ways. In addition, recording video tutorials and lectures creates a bank of teaching materials that can easily be reused for future workshops, whether virtual or in-person, and made freely accessible online to both participants and non-participants. This ensures that time and effort invested by the instructors has a lasting impact beyond the participants of the current workshop, making it much easier to organize subsequent events, even if the original instructors are unavailable. Finally, a flipped format allows participants to make their own choices about the proposed material, spending more time on topics they find relevant, interesting, or challenging and skipping topics they have already mastered or that do not apply to their research. In turn, this means that instructors are free to offer a wider range of topics, since they need not be relevant to all participants.

Since the flipped format used synchronous meetings for discussion, we encouraged participants to form study groups and work through the material together, much like what might happen at a traditional in-person workshop; however, this rarely happened in our experience. It is possible that such groups connected through other communication channels that were not visible to us, or that participants simply preferred to work through the material with their own local colleagues, whose research interests are closer to their own. This lack of group work likely also reflects limitations intrinsic to online-only, asynchronous communication. Online events may thus be less likely to foster close relationships between participants, although we could not assess whether this impacted the learning process.

Participant engagement can take three forms: learner-to-learner, learner-to-instructor, and learner-to-content; students value all three forms and broad engagement is critical for learning [122]. In general, participant engagement during the Stay-at-Home RevBayes Workshops was somewhat varied. This manifested as a core group of learners active on open Slack channels and asking questions during synchronous meetings, a subset of participants communicating primarily via direct messages to instructors and in the one-on-one meeting, and a small number of participants who were unable to fully participate because of unexpected changes to their local circumstances. Aside from the last group, similar patterns happen in on-site workshops. Although we believe the online format was not hugely detrimental to engagement, an online format provides overall less opportunity for participation than an on-site workshop, making it vital that interactions are engaging and meaningful.

In order to remain flexible, we only required attendance at the first and last sessions. Participants were made aware of this requirement before the event and attendance was very good (only 2 or 3 participants were unable to join). While office hours were not mandatory, we saw consistent attendance from many of the participants: the usual participation was around 10 participants (out of 20) in the first workshop, and around 4 for each of the two sessions (out of 25) in the second workshop. Overall, we found that having a formal round of introductions at the start of the workshop, as well as encouraging everyone to keep their camera on if possible during synchronous sessions, helped both participants and instructors to engage in the event.

C.4 Practical Considerations When Organizing a Virtual Workshop

Although the logistics involved in organizing an online workshop are reduced compared to an on-site event, there are still some key elements that must be considered to ensure that a workshop is accessible and successful.

C.4.1 Time Zones

At first glance, online events seem extremely accessible no matter where in the world interested participants are located. However, the diversity of participants' and instructors' locations means that holding synchronous activities in an online setting requires working to identify times that work for everyone. Thus, paying careful attention to overlap among the participants' and instructors' time zones is critical for promoting communication and engagement.

Figure C.3 shows the geographic distribution of the workshop participants and instructors. All the time zones are described in reference to Coordinated Universal Time (UTC). While the first iteration of the Stay-at-Home RevBayes Workshop attracted applications from all over the world, we restricted our participant selection to applicants residing in a specific time-zone range (from UTC-7 to UTC+3). Since most of the instructors also reside in those time zones, we were able to schedule synchronous meetings during a time that worked well for everyone involved. Because time zones prevented us from including a wider distribution of participants in the first course, the second iteration of the Stay-at-Home RevBayes Workshop specifically targeted applications from researchers based from UTC+4 to UTC+14 (including UTC-10).

In general, the set of time zones involved in the workshop will determine whether a synchronous session can accommodate everyone involved, or if replicate sessions must be offered at different times. For instance, it became clear early on that it would not be possible to find a single time for synchronous meetings during our workshop for participants in Asia and the Pacific, since our instructor team is based in Europe and North America. Thus, we held duplicate sessions that involved different combinations of instructors and participants. In order to ensure continuity across these duplicate sessions, we recorded the sessions or took notes to share the discussion with participants not in attendance.

Ultimately, confusion is difficult to avoid when holding events spanning time zones. To mitigate scheduling complications, we announced session times using UTC and provided links to online time-zone conversion services (*e.g.,* World Time Buddy[13]). Whether single or replicate sessions are chosen, announcing meeting times well in advance is critical, so that participants can plan their attendance around other commitments they may have. Additionally, it is also useful to send a notification about the synchronous session via Slack 30 minutes or an hour ahead of time to ensure that everyone is aware of the upcoming meeting, even if they accidentally miscalculated the time-zone adjustment.

[13] World Time Buddy: `https://www.worldtimebuddy.com`

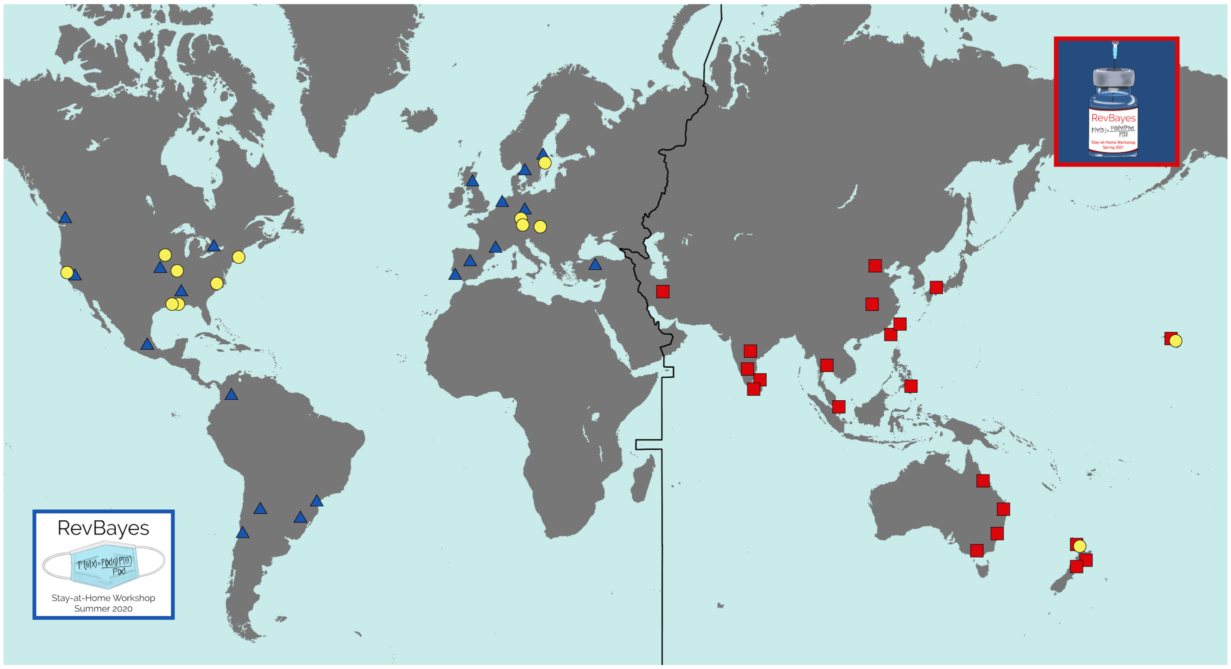


Figure C.3: Locations of participants and instructors from both Stay-at-Home RevBayes Workshops. Instructors (yellow circles) primarily reside in the United States and Europe. Participants from the Summer 2020 workshop (blue triangles) were based in North America, South America, and Europe. Participants from the Spring 2021 workshop (red squares) attended from Asia, Australia, New Zealand, and Hawaii. The black line dividing the map approximately delineates the boundary between UTC+3 and UTC+4 time zones, which determined the selection of participants in the two workshops. We designed logos (shown in the bottom-left and top-right corners) for each workshop that were inspired by current events.

C.4.2 Participant Recruitment and Selection

We created an application form using the online service Qualtrics[14]. Using this form, we asked applicants to rate their previous knowledge of Bayesian phylogenetics theory and applications and describe their learning goals, research questions, and datasets. Applicants were also required to indicate the time zone in which they would be residing during the workshop. Examples of the application form and participant confirmation form can be found in the Supplementary Materials.

We advertised the workshops using Twitter and the Evolution Directory[15]. For the first Stay-at-Home RevBayes Workshop, we advertised generally and this resulted in over 300 applications from all over the globe. When soliciting applications for the second virtual course, we contacted applicants from the first round who resided in our targeted time zones

[14] Qualtrics: `https://www.qualtrics.com`

[15] The Evolution Directory: `https://evol.mcmaster.ca/evoldir.html`

(UTC+4 to UTC+14) and encouraged them to reapply. Additionally, our advertisements specified that preference would be given to applicants from Asia and Pacific time zones and we received just over 100 applications in the second round. Applicants' responses indicated that they all felt comfortable with the prospect of participating in an online course, which likely contributed to the success of our workshops.

When organizing an online or in-person workshop, the number of participants and instructors involved is an important consideration. Adding instructors to the team comes at a very low cost for an online event, and we found that having a broad team of instructors, both in terms of geographical location and expertise, was very helpful in spreading the amount of work and ensuring that instructors would be responsive to questions. Since there is similarly little additional cost in adding participants, it can be tempting to expand the number of participants well beyond the usual attendance of on-site workshops. However, we decided to keep the number of participants low (20-25 participants) to guarantee that synchronous sessions could remain interactive and personal. Thus, we chose to provide the materials created for this workshop freely online, to ensure that unselected applicants and future students could still benefit from our efforts.

Selecting just 20-25 participants from the large pools of applications was difficult. We created a list of selected participants that maximized the geographic and institutional representation within the time-zone range for each workshop. Our hope is that by working with researchers from a wide array of institutions, they will be equipped with the knowledge to communicate what they learn to their colleagues and local communities. Although we selected participants at a variety of career stages (graduate students, postdocs, professors), we primarily focused on early career scientists, since they are usually more closely involved in setting up and running analyses and would, in our opinion, benefit the most from getting hands-on experience with the software. Since our workshops focused on learning to apply phylogenetic methods in RevBayes, we also prioritized applicants with datasets ready (or soon-to-be ready) for analysis. Finally, although we provided the *Phyloseminar* lectures for background on phylogenetic theory, our workshop did not focus heavily on this topic. As such, we preferred applicants who already had some knowledge of phylogenetic methods. In general, the specific goals and aims of the workshop should guide the participant selection process.

C.4.3 Technical Tools

For many university researchers and educators, the sudden switch to virtual learning and collaboration in the spring of 2020 was essentially a crash course on various tools for online communication. Because of our experiences teaching and collaborating remotely, we felt equipped to host a virtual workshop with participants from all around the world. We were fortunate to have access to institutional licenses for Zoom and Qualtrics, otherwise we would have had to opt for alternative services or purchase licenses specifically for the course. The global shutdown in response to the spread of COVID-19 additionally made Zoom a familiar tool for all workshop participants. Thus, this was the ideal service for our synchronous meetings.

In addition to Zoom, we relied heavily on Slack for communication among instructors and participants during the course. This service enables real-time chat that can be organized by topic and is much better suited to a virtual workshop format than email. Our workshop Slack space was created using the free version, which limits access to only the 10,000 most recent messages. Thus, participants and instructors must be made aware that not all of the messages will be accessible and they may have to save discussions they would like to view again.

We used several other tools and services for generating content for these virtual workshops including Google Docs for organizing information and sharing documents, YouTube for hosting recorded videos, and Open Broadcaster Software (OBS) for recording video tutorial guides. Open Broadcaster Software[16], in particular, is an extremely useful and flexible program for recording (and streaming) technical videos demonstrating software usage. This open-source and free tool is frequently used by video-game enthusiasts to live stream or record screencasts of game play, thus it is ideally suited for creating video tutorials on phylogenetic applications that require interacting with different platforms (*e.g.,* RevBayes, R, text editors, etc.).

C.4.4 Inclusivity and Accessibility

Online courses have the potential to enable participation from a much larger and diverse pool of scientists than most face-to-face workshops. However, it is important to develop a course timeline and format that enables flexibility and to carefully consider factors that may limit access to materials and communication. There are ways we can improve future virtual courses to make them more inclusive and accessible, however, we gained some key insights that are unique to the online-workshop format.

When recruiting participation from a global audience, it is important that efforts to make a workshop inclusive and accessible are mindful of the availability of required tools and software. This consideration is not limited to scientific software, but also any tool or service used for communication and coordination. For instance, Google services (Docs, Forms, YouTube) are blocked in China, requiring alternative tools or work-arounds to connect participants to materials hosted on Google sites. Announcing the required tools before the start of the workshop is essential so the participants can make the necessary arrangements or contact the organizers if there are issues.

There can be substantial monetary costs associated with in-person workshops that are significantly reduced in a virtual setting. These costs (*e.g.,* renting the venue and audiovisual equipment) are often, in turn, passed on to participants if the workshop organizers do not have access to funding or resources on site. Furthermore, an online format does not require travel and lodging (sometimes totaling several thousand dollars), reducing potentially prohibitive participant costs, particularly for researchers from countries with lower cost of living. Both Stay-at-Home RevBayes Workshops were offered free-of-charge because the instructor team is supported by grants and other sources of funding for which

[16] Open Broadcaster Software: `https://obsproject.com`

delivering workshops is a stated goal. Additionally, the size of the instructor team and online flipped-workshop format significantly reduced the workload, requiring a lower time commitment from instructors and organizers. For everyone involved, a virtual course additionally eliminates administrative and geographical burdens associated with traveling internationally (obtaining visas can be difficult or impossible depending on an individual's citizenship and the location of the workshop), making it much easier to reach scientists from regions where international travel is heavily restricted.

Ultimately, an online and flipped-format course can operate with much more scheduling flexibility than on-site workshops. Our choice to use a flipped-workshop format in combination with a limited number of synchronous sessions was designed to take advantage of this flexibility and allow both instructors and participants to easily combine workshop attendance and other professional or personal responsibilities. This created an opportunity to include both instructors and participants who might not have been able to leave at-home duties (*e.g.,* caregiving, teaching) for an in-person course. Because of this, our synchronous Zoom meetings occasionally welcomed cameos from small children and other family members.

When delivering content to people in their homes (or local offices or cafes) across multiple continents over several weeks, it should be expected that real-life issues will interfere and take some participants or instructors away from the course. For example, on August 10, 2020, during the first Stay-at-Home RevBayes Workshop, a severe thunderstorm (called a "derecho") hit the Midwestern United States. The storm swept through Iowa in the middle of one of the workshop's synchronous meetings and four workshop instructors lost power to their homes for over 72 hours. In other instances, participants faced unexpected changes to their work responsibilities, family emergencies, or pandemic-related effects in their regions. During our introductory sessions, we discussed the possibility of unplanned issues, letting the participants know that we would work to adapt to such interruptions and make sure all participants were able to meet their learning goals.

Workshop Code of Conduct

In recent years, workshop organizers and venues have worked to develop policies and procedures to ensure that in-person courses are safe and welcoming to all participants. It is critical that these efforts are not neglected for a virtual workshop. For the Stay-at-Home RevBayes courses, we developed a code of conduct[17] that provided a clear policy on harassment and discrimination (the code of conduct is also provided in the Supplementary Materials). This was adapted from the Safe Evolution[18] policies developed by the Society of Systematic Biologists, the American Society of Naturalists, and the Society for the Study of Evolution for virtual and in-person activities. This code applied to all interactions during the workshop, including synchronous sessions, but also the Slack forum as well as private messages between participants and/or instructors. Upon acceptance to the

[17] RevBayes Virtual Workshop Code of Conduct: `http://revbayes.com/workshops/code_of_conduct/virtual_coc`

[18] Safe Evolution: `https://www.evolutionmeetings.org/safe-evolution.html`

workshop, participants were required to agree to the policies stated in the code of conduct via the attendance confirmation form (see Supplementary Materials). Then, during our introductory meeting, we reintroduced the policies, discussed the procedures for reporting any discriminatory behavior or harassment, and stated that repeated violations of the code would lead to removal from the workshop. A clearly stated code of conduct communicates to participants that they will be treated respectfully during the workshop, creates a more inclusive culture [46, 40], and helps to reduce participants' hesitancy to post questions or start discussions during our meetings or on Slack.

C.5 Perspectives

In total, we received over 400 applications for the Stay-at-Home RevBayes Workshops and it is clear that there is a world-wide demand for accessible training in phylogenetic methods. Assessing the overall success of workshops, whether online or on-site, is generally tricky, particularly since some benefits of the training may not be apparent to participants until they are more advanced in their research projects. However, feedback from our workshop participants (via a formal survey and informal comments during meetings and on Slack) indicated that many workshop attendees felt that they gained a deeper understanding of applications in Bayesian phylogenetics and RevBayes, and that they would recommend attending future editions of the virtual workshop to colleagues (see the example workshop feedback form in the Supplementary Materials). Furthermore, our instructor team also appreciated the increased flexibility and the lower intensity of the format. All of the instructors from the 2020 team were interested in teaching an online workshop in the future and all who were available returned for the second offering.

While we feel that many of the choices we made in organizing two virtual RevBayes workshops led to successful outcomes, we recognize that there are unique challenges associated with an online setting and several ways we can improve future courses. For example, we plan on expanding the bank of recorded materials to cover more topics so that we can meet the needs of a broader audience of researchers. It will additionally be important to ensure that the videos and tutorials are kept up-to-date as RevBayes is under continued development.

Another area of improvement is apparent from the map in Figure C.3. Although we had participation from 24 different countries throughout the two workshops, there are distinct parts of the world that are not represented among our workshop participants. We must do more work to reach scientists residing in Africa, parts of Central and South America, and Asia, to ensure that residents of these regions interested in learning about RevBayes are connected to workshop opportunities. For instance, we need to broaden our approach to advertising future workshops by posting to mailing lists or communication platforms popular in these areas and by directly contacting local scientists and organizations. Moreover, our instructor team is primarily based in Europe and the US, reflecting the composition of the developer team involved in the RevBayes project. This ultimately created scheduling difficulties and limited synchronous interactions during the Asia/Pacific workshop. In the future, expanding the RevBayes developer community will improve these issues and may also help reach participants from currently underrepresented regions.

We also hope to improve on how we assess learning outcomes and facilitate participant engagement, which can be difficult for online courses. Providing a practical education and hands-on assistance is a common challenge for online teaching [115, 130]. In an in-person workshop, instructors and teaching assistants are able to walk around the room as participants are working through the material and assess progress or answer questions on the spot; this is not possible in an online format. However, it may be possible to encourage more engagement by actively following-up with participants, or implementing lightweight asynchronous follow-up activities such as journaling [26] after each section of the material. Through Slack, instructors could lead discussions, checking that participants were successful with the activities and encouraging discussion about the analyses. Additionally, we could facilitate participant engagement by integrating more discussion questions into the tutorial activities and encouraging participants to report and interpret their analysis results.

Although we encouraged participants to work in groups, the format and geographic distribution likely prevented this from occurring. These types of groups regularly form at in-person workshops, aiding in both material comprehension and community building. It is possible that participants will be more receptive to forming groups if this is facilitated by the workshop's structure and instructor team. Thus, in the future, we are interested in developing ways to help participants form collaborations early on in the course. Lastly, as a result of increased online instruction, there are many innovative strategies and techniques, such as HyFlex learning or utilizing cloud computing resources, that could be implemented in future workshops [see 67, 116].

As vaccination efforts reach more and more parts of the globe, there is an understandable desire to return to the old "normal" and to put everything associated with the pandemic behind us, including online teaching. However, although in-person workshops offer opportunities for networking and interactions that are difficult to facilitate in an online setting, they also tend to select participants with specific characteristics: the ability to pay for the event and the travel expenses, the ability to travel internationally without a heavy administrative burden, and no medical needs or personal responsibilities requiring their presence at home. Online workshops can reach beyond these traditional audiences and offer training to more diverse populations of scientists with less access to such courses locally.

Online events also help limit carbon-emitting air travel and thus lower the contribution of our scientific community to the climate crisis [85, 153]. A geographically dispersed audience for an in-person workshop leads to excessive carbon emissions from travel. Locally based workshops with an emphasis on land-based travel can have a lower environmental impact, but such events are limited to areas with a high concentration of researchers, creating inequality in access to training. Additionally, regional workshops may still require considerable air travel if instructors are not all based in the same area. Thus, online or hybrid workshops have the greatest potential to reduce the carbon footprint of phylogenetics workshops.

The complexity and difficulty of statistical phylogenetics software continues to increase and workshops will remain an extremely important mechanism for researchers to learn how to use analysis tools. In this paper, we have focused on the distinct benefits and challenges of virtual workshops, but it is important to note that no learning format is effective for all people, as can be evidenced by the numerous formats that arose in the evolutionary biology community during the COVID-19 pandemic. The formats range from completely synchronous workshops over that take place over a few days (*e.g.,* Taming the BEAST Online [19] or the Sydney Phylogenetics Workshop [20]) to completely asynchronous where the provided materials are accessed by the participants on their own timelines (*e.g.,* SLiM Workshop [21]). The RevBayes workshop sits between these two extremes by offering both synchronous and asynchronous portions. Any choice of format comes with its own logistical requirements, pedagogical considerations, and impacts the level of accessibility, thus the format should be tailored to the overall goals of each workshop. We felt that the hybrid format provided a balance of independence and autonomy while also giving adequate access to research experts for guidance through the material. Nevertheless, the value of in-person learning and networking is undeniable. Thus, the RevBayes developer community plans to offer both in-person and virtual workshops in the future to strengthen our connections with scientists using statistical phylogenetics to answer biological questions. Many lessons learned from our virtual workshop can be extended to in-person settings. A flipped classroom format allows participants to engage with the material beforehand and seek deeper understanding during synchronous sessions with instructors. We believe this format can help participants achieve learning outcomes and could be adopted for in-person workshops. Additionally, having recorded content creates a bank of reference material for both participants and non-participants long after any workshop concludes. The materials developed for online courses thus present exciting opportunities for organizers of in-person workshops to consider alternative pedagogical practices that may enhance learning in a face-to-face course. By diversifying the formats of the workshops we offer, we not only open educational opportunities to a broader range of learners, but we can also improve how we teach concepts and methods across all courses.

In conclusion, we believe that virtual courses on phylogenetic analyses and approaches are more than a workaround for the current circumstances and offer numerous unique advantages. We hope that our experiences will inspire other methods developers in our community to explore this format further and that online workshops will become an integral part of scientific training in the future.

[19] Taming the BEAST Online: `https://bsse.ethz.ch/cevo/taming-the-beast/overview-2021.html`

[20] Sydney Phylogenetics Workshop: `https://meep.sydney.edu.au/workshops`

[21] SLiM Workshop: `http://benhaller.com/workshops/workshops.html`

Concluding Discussion

Bayesian phylogenetic inference is a widespread method in evolutionary biology. Its success among researchers can be attributed to the robustness of the method and the application in different fields of evolutionary research. Despite great advancements in the field over the past two decades, some methodological questions remained unresolved. Here we addressed some of these questions, namely convergence assessment of Markov chain Monte Carlo (MCMC) algorithms in phylogenetics, posterior predictive tests of substitution model adequacy and its interpretation in phylogenetic inference, substitution model over-parameterization and the impact of prior probability distributions, and the application of such results in the gene tree inference of an empirical dataset. The results presented in this dissertation contribute to the advance of the robustness of Bayesian phylogenetic methods.

The conundrum of convergence assessment in phylogenetics has been dealt with rather vague methods. Such methods relied on visual inspection and arbitrary thresholds [134, 144, 173]. These practices imposed a challenge for reliability and reproduction of phylogenetic studies. Besides these problems, visual inspection becomes a hurdle for the analysis of multiple inferences in the phylogenomic era. Here, we tackled this problem by developing a novel method for convergence assessment with clear thresholds and automation. We tested commonly used methods to estimate the effective sample size (ESS) for different MCMC algorithms and found that `Tracer` performed better for all tested scenarios. Additionally, we proposed the Kolmogorov-Smirnov [94, 156] test for the reproducibility test of continuous parameters of multiple MCMC chains. Furthermore, we implemented the transformation of tree topologies into traces of presence/absence of splits to facilitate the convergence assessment of these difficult discrete parameters. This transformation made it possible to address the reproducibility of MCMC chains with the newly implemented expected difference of split frequencies (EDSF). All these methods were implemented in an easy-to-use R package called `Convenience` that can facilitate the testing for convergence for output from different standard phylogenetic inference software.

Our next contribution was regarding the testing of model adequacy in Bayesian phylogenetics. We characterized the expected behavior of the distributions of posterior predictive p-values for different simulation scenarios. This was achieved by simulating data under different conditions, performing the phylogenetic inference and then, performing the posterior predictive analysis. We observed that the distribution of p-values for the focal test statistics are mostly concentrated around 0.5. While the distribution of p-values for the ancillary test statistics are uniformly distributed. The conclusion from these findings is that when performing posterior predictive analysis, the finding of outlier p-values is a very strong signal that the model failed to capture a feature from the data.

The simulation study presented in Chapter 4 had the goal of answering whether an over-parameterized substitution model biases Bayesian phylogenetic inference. Moreover, if selecting a substitution model is a necessary step in phylogenetic inference. Previous studies had demonstrated that under-parameterization is a problem for tree topology and branch length inference [109, 160, 27, 89, 159]. But the problem of over-parameterization received rather little attention, with two studies partially addressing the question [81, 111]. Additionally to the over-parameterized setting for performing phylogenetic inference, we explored different prior distribution schemes. Our results showed that substitution model over-parameterization does not bias Bayesian phylogenetic inference under the `Tame` prior scheme, with all other prior schemes resulting in biased tree length estimates. The findings in this chapter corroborate the idea that substitution model selection is not necessary for Bayesian phylogenetic inference [1]. Instead of wasting time and resources with the model selection step, researchers should use the most complex substitution model with the proper prior scheme.

The last chapter of this dissertation incorporated the findings of the previous chapters in the exploration of Bayesian phylogenetic gene tree discordance for empirical data. The results show that proper convergence assessment is an essential step of phylogenetic inference. When convergence is not properly assessed, results can lead to false evolutionary conclusions and the amount of gene tree discordance can be overestimated. After the proper analysis of convergence of the phylogenetic gene trees, we evaluated the amount of gene tree discordance among the data and the possible explanations for such discordance. We observed no support of incomplete lineage sorting in the analyzed data and further investigated the model adequacy by performing posterior predictive tests. Our results show that the model failed to capture the features of all investigated genes. We concluded that future research should focus on improving the models of sequence evolution to better capture the heterogeneity present in real data.

The results presented in this dissertation contribute to further development of robust Bayesian phylogenetic inference. We tackled methodological gaps in convergence assessment, substitution model adequacy and choice of substitution model. After applying the proposed advancements in real data, we provided insights for future research.

Bibliography

[1] S. Abadi, D. Azouri, T. Pupko, and I. Mayrose. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature communications*, 10(1):1–11, 2019.

[2] J. Adachi and M. Hasegawa. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *Journal of Molecular Evolution*, 40(6):622–628, June 1995.

[3] J. Adachi and M. Hasegawa. Improved dating of the human/chimpanzee separation in the mitochondrial dna tree: Heterogeneity among amino acid sites. *J. Mol. Evol.*, 40:622–628, 1995.

[4] M. E. Alfaro and M. T. Holder. The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):19–42, 2006.

[5] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415, 2004.

[6] L. M. Andrade de Oliveira, E. Cordeiro-Spinetti, F. P. G. Neves, P. S. Sujii, R. L. Ribeiro, S. S. de Lyra, T. C. A. Pinto, and M. L. Bonatelli. Going online in pandemic time: A divulgamicro workshop experience. *Journal of microbiology & biology education*, 22(1):ev22i1–2493, 2021.

[7] G. Baele, M. S. Gill, P. Bastide, P. Lemey, and M. A. Suchard. Markov-modulated continuous-time Markov chains to identify site-and branch-specific evolutionary variation in BEAST. *Systematic Biology*, 70(1):181–189, 2021.

[8] J. N. Bailenson. Nonverbal overload: A theoretical argument for the causes of zoom fatigue. *Technology, Mind, and Behavior*, 2(1), 2 2021. https://tmb.apaopen.org/pub/nonverbal-overload.

[9] K. M. Banner, K. M. Irvine, and T. J. Rodhouse. The use of Bayesian priors in Ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889, 2020.

[10] J. Barido-Sottani, J. A. Justison, A. M. Wright, R. C. Warnock, W. Pett, and T. A. Heath. Estimating a time-calibrated phylogeny of fossil and extant taxa using RevBayes. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 5.2:1–5.2:23. No commercial publisher — Authors open access book, 2020.

[11] J. Barido-Sottani, J. A. Justison, A. M. Wright, R. C. Warnock, W. Pett, and T. A. Heath. Estimating a time-calibrated phylogeny of fossil and extant taxa using RevBayes. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 5.2:1–5.2:23. No commercial publisher | Authors open access book, 2020.

[12] M. Bayes and M. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763. Publisher: The Royal Society.

[13] L. Bofkin and N. Goldman. Variation in Evolutionary Processes at Different Codon Positions. *Molecular Biology and Evolution*, 24(2):513–521, 11 2006.

[14] J. P. Bollback. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19:1171–1180, 2002.

[15] R. Borges, G. J. Szöllősi, and C. Kosiol. Quantifying gc-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, 212(4):1321–1336, 2019.

[16] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537, 2014.

[17] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4):e1006650, 2019.

[18] R. R. Bouckaert and A. J. Drummond. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC evolutionary biology*, 17(1):1–11, 2017.

[19] G. E. Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):383–404, 1980.

[20] S. P. Brooks, P. Giudici, and A. Philippe. Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, 12(1):1–22, 2003.

[21] J. M. Brown. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Syst. Biol.*, 63(3):334–348, May 2014.

[22] J. M. Brown, S. M. Hedtke, A. R. Lemmon, and E. M. Lemmon. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. *Systematic Biology*, 59(2):145–161, 2010.

[23] J. M. Brown and A. R. Lemmon. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology*, 56(4):643–655, 2007.

[24] J. M. Brown and R. C. Thomson. Evaluating model performance in evolutionary biology. *Annu. Rev. Ecol. Evol. Syst.*, Nov. 2018.

[25] M. W. Cadotte, B. J. Cardinale, and T. H. Oakley. Evolutionary history and the effect of biodiversity on plant productivity. *Proceedings of the National Academy of Sciences*, 105(44):17012–17017, 2008.

[26] E. K. Camfield, L. Beaster-Jones, A. D. Miller, and K. M. Land. Using writing in science class to understand and activate student engagement and self-efficacy. In *Active learning in College science*, pages 89–105. Springer, 2020.

[27] C. W. Cunningham, H. Zhu, and D. M. Hillis. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. *Evolution*, 52(4):978–987, 1998.

[28] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada. jmodeltest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8):772, 2012.

[29] C. Darwin. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

[30] N. De Maio, C. Schlötterer, and C. Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular Biology and Evolution*, 30(10):2249–2262, 2013.

[31] N. De Maio, D. Schrempf, and C. Kosiol. PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031, 2015.

[32] J. H. Degnan and N. A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology Evolution*, 24(6):332–340, 2009.

[33] V. P. Doyle, R. E. Young, G. J. P. Naylor, and J. M. Brown. Can we identify genes with increased phylogenetic reliability? *Syst. Biol.*, 64(5):824–837, Sept. 2015.

[34] V. P. Doyle, R. E. Young, G. J. P. Naylor, and J. M. Brown. Can we identify genes with increased phylogenetic reliability? *Systematic Biology*, 64(5):824–837, 2015.

[35] A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3):1307–1320, 07 2002.

[36] S. Duchêne, F. Di Giallonardo, and E. Holmes. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Mol. Biol. Evol.*, 33:255–267, 2016.

[37] S. Ekman and R. Blaalid. The devil in the details: interactions between the branch-length prior and likelihood model affect node support and branch lengths in the phylogeny of the Psoraceae. *Systematic Biology*, 60(4):541–561, 2011.

[38] L. G. Fabreti and S. Höhna. Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *Methods in Ecology and Evolution*, 13(1):77–90, 2022.

[39] L. G. Fabreti and S. Höhna. Convergence assessment for bayesian phylogenetic analysis using mcmc simulation. *Methods in Ecology and Evolution*, 13(1):77–90, 2022.

[40] B. Favaro, S. Oester, J. A. Cigliano, L. A. Cornick, E. J. Hind, E. Parsons, and T. J. Woodbury. Your science conference should have a code of conduct. *Frontiers in Marine Science*, 3:103, 2016.

[41] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[42] J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

[43] J. M. Flegal, J. Hughes, D. Vats, and N. Dai. *mcmcse: Monte Carlo Standard Errors for MCMC.* Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN, 2020.

[44] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

[45] P. Foster. Modeling compositional heterogeneity. *Syst. Biol.*, 53:485–495, 2004.

[46] A. J. Foxx, R. S. Barak, T. M. Lichtenberger, L. K. Richardson, A. J. Rodgers, and E. W. Williams. Evaluating the prevalence and quality of conference codes of conduct. *Proceedings of the National Academy of Sciences*, 116(30):14931–14936, 2019.

[47] P. B. Frandsen, B. Calcott, C. Mayer, and R. Lanfear. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evolutionary Biology*, 15(1):1–17, 2015.

[48] A. Gavryushkina, T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1):57–73, 2017.

[49] E. Gaya, B. D. Redelings, P. Navarro-Rosinés, X. Llimona, M. De Cáceres, and F. Lutzoni. Align or not to align? Resolving species complexes within the Caloplaca saxicola group as a case study. *Mycologia*, 103(2):361–378, 2011. Publisher: Mycological Society of America.

[50] A. Gelman. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electron. J. Stat.*, 7:2595–2602, 2013.

[51] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, 2014.

[52] A. Gelman and C. Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):967–1033, 2017.

[53] A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. Publisher: Institute of Mathematical Statistics.

[54] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, 1991.

[55] C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, Nov. 1992. Publisher: Institute of Mathematical Statistics.

[56] C. J. Geyer. Introduction to Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. Num Pages: 46.

[57] P. W. Glynn. ESTIMATING THE ASYMPTOTIC VARIANCE WITH BATCH MEANS. *Operations Research Letters*, 10:431–435, 1991.

[58] P. W. Glynn and D. L. Iglehart. Simulation Output Analysis Using Standardized Time Series. *Mathematics of Operations Research*, 15(1):1–16, Feb. 1990. Publisher: INFORMS.

[59] N. Goldman. Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36(2):182–198, 1993.

[60] L. Gong and J. M. Flegal. A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25(3):684–700, 2016.

[61] X. Gu, Y. Fu, and W. Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, 12:546–557, 1995.

[62] X. Gu, Y.-X. Fu, and W.-H. Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. 12(4):546–557, 1995.

[63] H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–396, 1999.

[64] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[65] J. D. Hamilton. *Time series analysis*, volume 10. Cambridge Univ Press, 1994.

[66] S. M. Harrington, V. Wishingrad, and R. C. Thomson. Properties of Markov Chain Monte Carlo Performance across Many Empirical Alignments. *Molecular Biology and Evolution*, 2021.

[67] B. N. Harris, P. C. McCarthy, A. M. Wright, H. Schutz, K. S. Boersma, S. L. Shepherd, L. A. Manning, J. L. Malisch, and R. M. Ellington. From panic to pedagogy: Using online active learning to promote inclusive instruction in ecology and evolutionary biology courses and beyond. *Ecology and evolution*, 10(22):12581–12612, 2020.

[68] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[69] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[70] S. Höhna, L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown. P3: Phylogenetic posterior prediction in RevBayes. *Mol. Biol. Evol.*, 35(4):1028–1034, Apr. 2018.

[71] S. Höhna, L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown. P^3: Phylogenetic Posterior Prediction in RevBayes. *Molecular biology and evolution*, 35(4):1028–1034, Apr. 2018.

[72] S. Höhna and A. J. Drummond. Guided Tree Topology Proposals for Bayesian Phylogenetic Inference. *Systematic Biology*, 61(1):1–11, 2012.

[73] S. Höhna, M. J. Landis, and T. A. Heath. Phylogenetic Inference Using RevBayes. *Current protocols in bioinformatics*, 57:6–16, 2017.

[74] S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, 65(4):726–736, 2016.

[75] S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive Model-Specification language. *Syst. Biol.*, 65(4):726–736, July 2016.

[76] S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736, 05 2016.

[77] S. Höhna, M. J. Landis, and J. P. Huelsenbeck. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. *PeerJ*, 9:e12438, 2021.

[78] S. Höhna, M. R. May, and B. 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.

[79] J. Huelsenbeck, B. Larget, and M. Alfaro. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21(6):1123, 2004.

[80] J. P. Huelsenbeck and B. Rannala. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Systematic Biology*, 53(6):904–913, 12 2004.

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

[82] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, Aug. 2001.

[83] S. Höhna, L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown. P3: Phylogenetic posterior prediction in RevBayes. *Mol. Biol. Evol.*, 35(4):1028–1034, 2018.

[84] S. Höhna, M. R. May, and B. R. Moore. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics*, 32(5):789–791, 11 2015.

[85] S. Jäckle. Reducing the carbon footprint of academic conferences by online participation: The case of the 2020 virtual european consortium for political research general conference. *PS: Political Science & Politics*, pages 1–6, 2021.

[86] T. Jukes and C. Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132, 1969.

[87] T. H. Jukes and C. R. Cantor. CHAPTER 24 - evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

[88] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589, 2017.

[89] C. R. Kelsey, K. A. Crandall, and A. F. Voevodin. Different Models, Different Trees: The Geographic Origin of PTLV-I. *Molecular Phylogenetics and Evolution*, 13(2):336–347, Nov. 1999.

[90] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[91] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458, 1981.

[92] A. King. From sage on the stage to guide on the side. *College Teaching*, 41(1):30–35, 1993.

[93] L. King, J. Wakeley, and S. Carmi. A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. *Theor. Popul. Biol.*, 122:22–29, July 2018.

[94] A. L. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91, 1933.

[95] L. S. Kubatko and J. H. Degnan. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*, 56(1):17–24, 02 2007.

[96] D. Kühnert, C.-H. Wu, and A. J. Drummond. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, 11(8):1825–1841, 2011.

[97] M. J. Lage, G. J. Platt, and M. Treglia. Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, 31(1):30–43, 2000.

[98] C. Lakner, P. van der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology*, 57(1):86–103, 2008.

[99] R. Lanfear, B. Calcott, S. Y. W. Ho, and S. Guindon. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29(6):1695–1701, 2012.

[100] R. Lanfear, P. B. Frandsen, A. M. Wright, T. Senfeld, and B. Calcott. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, 34(3):772–773, 2017.

[101] R. Lanfear, X. Hua, and D. L. Warren. Estimating the Effective Sample Size of Tree Topologies from Bayesian Phylogenetic Analyses. *Genome Biology and Evolution*, 8(8):2319–2332, 2016.

[102] P. S. Laplace. *Essai philosophique sur les probabilités*. Cambridge Library Collection - Mathematics. Cambridge University Press, Cambridge, 5 edition, 2009.

[103] B. Larget and D. Simon. Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Molecular Biology and Evolution*, 16(6):750–750, Jan. 1999.

[104] N. Lartillot. The Bayesian Approach to Molecular Phylogeny. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, pages 1.4:1–1.4:17. No commercial publisher — Authors open access book, 2020.

[105] N. Lartillot, H. Brinkmann, and H. Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology*, 7(Suppl 1):S4, 2007.

[106] N. Lartillot, H. Brinkmann, and H. Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.*, 7:S4, 2007.

[107] N. Lartillot, T. Lepage, and S. Blanquart. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286, 2009.

[108] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109, 2004.

[109] T. Leitner, S. Kumar, and J. Albert. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *Journal of Virology*, 71(6):4761–4770, June 1997.

[110] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard. Bayesian phylogeography finds its roots. *PLOS Computational Biology*, 5(9):1–16, 09 2009.

[111] A. R. Lemmon and E. C. Moriarty. The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology*, 53(2):265–277, 2004.

[112] N. P. Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7):912–928, 2019.

[113] S. Li, D. K. Pearl, and H. Doss. Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 95(450):493–508, 2000. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[114] D. V. Lindley. The Use of Prior Probability Distributions in Statistical Inference and Decisions. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 4.1:453–469, Jan. 1961. Publisher: University of California Press.

[115] J. M. Long, M. A. Joordens, and G. Littlefair. Engineering distance education at deakin university australia. In *Proceedings of the IACEE 14th World Conference on Continuing Engineering Education*, pages 1–13. International Association for Continuing Engineering Education, 2014.

[116] P. Lowenthal, J. Borup, R. West, and L. Archambault. Thinking beyond zoom: Using asynchronous video to maintain connection and engagement during the covid-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):383–391, 2020.

[117] A. N. Lukashev, Y. A. Vakulenko, N. A. Turbabina, A. A. Deviatkin, and J. F. Drexler. Molecular epidemiology and phylogenetics of human enteroviruses: Is there a forest behind the trees? *Reviews in Medical Virology*, 28(6):e2002, 2018. e2002 RMV-2018-031.R1.

[118] W. P. Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, 09 1997.

[119] J. Marin, G. Achaz, A. Crombach, and A. Lambert. The genomic view of diversification. *Journal of Evolutionary Biology*, 33(10):1387–1404, 2020.

[120] D. C. Marshall. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Systematic Biology*, 59(1):108–117, 2010.

[121] A. P. Martin and S. R. Palumbi. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences*, 90(9):4087–4091, 1993.

[122] F. Martin and D. U. Bolliger. Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning*, 22(1):205–222, 2018.

[123] B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55(1):1–12, Mar. 1999.

[124] X.-L. Meng. Posterior predictive p-values. *Ann. Stat.*, 22:1142–1160, 1994.

[125] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[126] D. F. Morales-Briones, G. Kadereit, D. T. Tefarikis, M. J. Moore, S. A. Smith, S. F. Brockington, A. Timoneda, W. C. Yim, J. C. Cushman, and Y. Yang. Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l. *Systematic Biology*, 70(2):219–235, 06 2020.

[127] W. K. Morris, P. A. Vesk, M. A. McCarthy, S. Bunyavejchewin, and P. J. Baker. The neglected tool in the Bayesian ecologist's shed: a case study testing informative priors' effect on model accuracy. *Ecology and Evolution*, 5(1):102–108, 2015.

[128] N. F. Müller and R. R. Bouckaert. Adaptive Metropolis-coupled MCMC for BEAST 2. *PeerJ*, 8:e9473, 2020.

[129] N. F. Müller, K. E. Kistler, and T. Bedford. A Bayesian approach to infer recombination patterns in coronaviruses. *Nature Communications*, 13(1):4186, July 2022. Number: 1 Publisher: Nature Publishing Group.

[130] K. Nahar, R. Chowdhury, et al. Effectiveness of flipped classroom model in distance learning. In *30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate, Motivate*, page 453. Engineers Australia, 2019.

[131] F. F. Nascimento, M. dos Reis, and Z. Yang. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1(10):1446–1454, 2017.

[132] F. F. Nascimento, M. dos Reis, and Z. Yang. A biologist's guide to Bayesian phylogenetic analysis. *Nature ecology & evolution*, 1(10):1446–1454, Oct. 2017.

[133] R. Nielsen and M. Slatkin. *An Introduction to Population Genetics: Theory and Applications*. Sinauer, July 2013. Google-Books-ID: Iy08kgEACAAJ.

[134] J. Nylander, J. Wilgenbusch, D. Warren, and D. Swofford. Awty (are we there yet?): a system for graphical exploration of mcmc convergence in bayesian phylogenetics. *Bioinformatics*, 24(4):581, 2008.

[135] J. A. A. Nylander, F. Ronquist, J. P. Huelsenbeck, and J. Nieves-Aldrey. Bayesian phylogenetic analysis of combined data. *Systematic Biology*, 53(1):47–67, 2004.

[136] J. E. O'Reilly, M. dos Reis, and P. C. Donoghue. Dating tips for divergence-time estimation. *Trends in Genetics*, 31(11):637–650, 2015.

[137] J. F. Parham, P. C. J. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. L. Patané, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. Müller, K. T. Smith, J. M. Theodor, R. C. M. Warnock, and M. J. Benton. Best practices for justifying fossil calibrations. *Systematic Biology*, 61(2):346–359, 2012.

[138] M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11, Mar. 2006.

[139] D. Posada. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25(7):1253–1256, 2008.

[140] D. Posada and K. A. Crandall. MODELTEST: testing the model of DNA substitution. *Bioinformatics (Oxford, England)*, 14(9):817–818, 1998.

[141] D. Posada and K. A. Crandall. Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4):580–601, 2001.

[142] N. Prasad, S. Fernando, S. Willey, K. Davey, F. Kent, A. Malhotra, and A. Kumar. Online interprofessional simulation for undergraduate health professional students during the covid-19 pandemic. *Journal of Interprofessional Care*, 34(5):706–710, 2020.

[143] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[144] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5):901, 2018.

[145] B. Rannala, T. Zhu, and Z. Yang. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution*, 29(1):325–335, 2012.

[146] E. Richards, J. Brown, A. Barley, R. Chong, and R. Thomson. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? *Syst. Biol.*, 67:847–860, 2018.

[147] E. J. Richards, J. M. Brown, A. J. Barley, R. A. Chong, and R. C. Thomson. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? *Systematic Biology*, 67(5):847–860, 2018.

[148] C. P. Robert and G. Casella. *The Metropolis—Hastings Algorithm*, pages 231–283. Springer New York, New York, NY, 1999.

[149] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

[150] S. Roch and M. Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, 2015.

[151] F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

[152] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.*, 12:1151–1172, 1984.

[153] S. Sarabipour, A. Khan, Y. F. S. Seah, A. D. Mwakilili, F. N. Mumoki, P. J. Sáez, B. Schwessinger, H. J. Debat, and T. Mestrovic. Changing scientific meetings for the better. *Nature Human Behaviour*, 5(3):296–300, 2021.

[154] C. Scornavacca, K. Belkhir, J. Lopez, R. Dernat, F. Delsuc, E. J. P. Douzery, and V. Ranwez. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Molecular Biology and Evolution*, 36(4):861–862, 01 2019.

[155] C. Scornavacca and N. Galtier. Incomplete Lineage Sorting in Mammalian Phylogenomics. *Systematic Biology*, 66(1):112–120, 09 2016.

[156] N. V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

[157] D. L. Stern and V. Orgogozo. The loci of evolution: How predictable is genetic evolution? *Evolution*, 62(9):2155–2177, 2008.

[158] T. P. Straatsma, H. J. C. Berendsen, and A. J. Stam. Estimation of statistical errors in molecular simulation calculations. *Molecular Physics*, 57(1):89–95, 1986.

[159] J. Sullivan and P. Joyce. Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 36:445–466, 2005.

[160] J. Sullivan and D. L. Swofford. Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution*, 4(2):77–86, 1997.

[161] E. Y. Suárez-Villota, C. A. Quercia, L. M. Díaz, V. Vera-Sovier, and J. J. Nuñez. Speciation in a biodiversity hotspot: Phylogenetic relationships, species delimitation, and divergence times of patagonian ground frogs from the eupsophus roseus group (alsodidae). *PLOS ONE*, 13(12):1–19, 12 2018.

[162] G. J. Szöllősi, E. Tannier, V. Daubin, and B. Boussau. The Inference of Gene Trees with Species Trees. *Systematic Biology*, 64(1):e42–e62, 07 2014.

[163] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. Diana, G. Lehmann, D. Toren, J. Wang, V. E. Fraifeld, and J. P. de Magalhães. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090, 11 2017.

[164] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[165] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *In: Some Mathematical Questions in Biology—DNA Sequence Analysis, Miura RM (Ed.), American Mathematical Society, Providence (RI)*, 17:57–86, 1986.

[166] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

[167] M. B. Thompson. A Comparison of Methods for Computing Autocorrelation Time. Technical report, Department of Statistics, University of Toronto, 2010.

[168] K. K. Ullrich and D. Tautz. *Population Genomics of the House Mouse and the Brown Rat*, pages 435–452. Springer US, New York, NY, 2020.

[169] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[170] R. Van Noorden, B. Maher, and R. Nuzzo. The top 100 papers. *Nature News*, 514(7524):550, 2014.

[171] D. Vats and C. Knudson. Revisiting the Gelman-Rubin Diagnostic. *arXiv:1812.09384 [stat]*, Sept. 2020. arXiv: 1812.09384.

[172] R. C. M. Warnock, Z. Yang, and P. C. J. Donoghue. Exploring uncertainty in the calibration of the molecular clock. *Biology letters*, 8(1):156–159, 2012.

[173] D. L. Warren, A. J. Geneva, and R. 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.

[174] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7(2):256–276, 1975.

[175] C. Whidden and F. A. Matsen. Quantifying MCMC exploration of phylogenetic tree space. *Systematic Biology*, page syv006, 2015.

[176] Y. Xing and R. H. Ree. Uplift-driven diversification in the hengduan mountains, a temperate biodiversity hotspot. *Proceedings of the National Academy of Sciences*, 114(17):E3444–E3451, 2017.

[177] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. 39(3):306–314, 1994.

[178] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, Sept. 1994.

[179] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology Evolution*, 11(9):367–372, 1996.

[180] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, 11(9):367–372, Sept. 1996.

[181] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7):717–724, July 1997.

[182] Z. Yang and C. E. Rodríguez. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proceedings of the National Academy of Sciences*, 110(48):19307–19312, 2013.

[183] H. Yu, A. Jamieson, A. Hulme-Beaman, C. J. Conroy, B. Knight, C. Speller, H. Al-Jarah, H. Eager, A. Trinks, G. Adikari, H. Baron, B. Böhlendorf-Arslan, W. Bohingamuwa, A. Crowther, T. Cucchi, K. Esser, J. Fleisher, L. Gidney, E. Gladilina, P. Gol'din, S. M. Goodman, S. Hamilton-Dyer, R. Helm, J. C. Hillman, N. Kallala, H. Kivikero, Z. E. Kovács, G. K. Kunst, R. Kyselý, A. Linderholm, B. Maraoui-Telmini, N. Marković, A. Morales-Muñiz, M. Nabais, T. O'Connor, T. Oueslati, E. M. Quintana Morales, K. Pasda, J. Perera, N. Perera, S. Radbauer, J. Ramon, E. Rannamäe, J. Sanmartí Grego, E. Treasure, S. Valenzuela-Lamas, I. van der Jagt, W. Van Neer, J.-D. Vigne, T. Walker, S. Wynne-Jones, J. Zeiler, K. Dobney, N. Boivin, J. B. Searle, B. Krause-Kyora, J. Krause, G. Larson, and D. Orton. Palaeogenomic analysis of black rat (Rattus rattus) reveals multiple European

introductions associated with human economic history. *Nature Communications*, 13(1):2399, May 2022. Number: 1 Publisher: Nature Publishing Group.

[184] C. Zhang, B. Rannala, and Z. Yang. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. *Systematic Biology*, 61(5):779–784, 2012.

[185] Y. Zhou, H. Brinkmann, N. Rodrigue, N. Lartillot, and H. Philippe. A dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Mol. Biol. Evol.*, 27:371–384, 2010.

[186] D. J. Zwickl and M. T. Holder. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Systematic Biology*, 53(6):877–888, 2004.

[187] J. M. Zydney, Z. Warner, and L. Angelone. Learning through experience: Using design based research to redesign protocols for blended synchronous learning environments. *Computers & Education*, 143:103678, 2020.

[188] S. Álvarez Carretero, A. U. Tamuri, M. Battini, F. F. Nascimento, E. Carlisle, R. J. Asher, Z. Yang, P. C. J. Donoghue, and M. Dos Reis. A species-level timeline of mammal evolution integrating phylogenomic data. *Nature*, 602(7896):263–267, Feb. 2022.