

Evaluating, Accelerating and Extending the Multispecies Coalescent Model of Evolution

HUW ALEXANDER OGILVIE
DECEMBER 2017

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
OF THE AUSTRALIAN NATIONAL UNIVERSITY.

DECLARATION

The length of this thesis is 30,924 words, exclusive of footnotes, tables, figures, maps, bibliographies and appendices. The abstract, introduction and conclusion of the thesis are my own original work, with input and advice from my chair supervisor, Professor Craig Moritz. Each chapter was based on research conducted jointly with others, and all chapters have been published in or submitted to peer-reviewed journals.

Chapter 1 was published in 2016 in *Systematic Biology* as “Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods” (volume 65, issue 3, pages 381–396). The authors are Huw A. Ogilvie, Joseph Heled, Dong Xie and Alexei J. Drummond. DX and AJD designed, conducted, and analysed the results of Experiment 1. JH and HAO analysed the results of Experiment 2, which was designed and conducted by JH. HAO designed, conducted, and analysed the results of Experiment 3. All authors contributed to the writing of the manuscript.

Chapter 2 was published in 2017 in *Molecular Biology and Evolution* as “StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates” (volume 34, issue 8, pages 2101–2114). The authors are Huw A. Ogilvie, Remco R. Bouckaert and Alexei J. Drummond. HAO and RRB wrote the StarBEAST2 software. HAO designed, conducted and analysed the results of all experiments with advice and input from AJD. The manuscript was written by HAO with advice and input from AJD.

Chapter 3 was published online in 2017 in *Molecular Biology and Evolution* as “Bayesian Inference of Species Networks from Multilocus Sequence Data”. The authors are Chi Zhang, Huw A. Ogilvie, Alexei J. Drummond and Tanja Stadler. The SpeciesNetwork software was written by HAO and CZ with input from AJD and TS. The birth-hybridization process was derived by TS and CZ. CZ designed, conducted and analysed the results of all experiments with input from HAO and AJD. The manuscript was written by CZ and HAO with input from TS and AJD.

Chapter 4 has been submitted to *Systematic Biology* as “Inferring Species Trees Using Integrative Models of Species Evolution”. The authors are Huw A. Ogilvie, Timothy G. Vaughan, Nicholas J. Matzke, Graham J. Slater, Tanja Stadler, David Welch and Alexei J. Drummond. The integrative model of species evolution was formulated by AJD, TS and HAO. Software to support the model was developed by HAO with input from AJD. TGV designed, conducted and analysed the results of experiments to test the correctness of the implementation, with input from HAO. HAO designed, conducted and analysed the results of all other experiments with input from AJD, NJM and GJS. The manuscript with written by HAO with input from all other authors, and parts of the introduction were based on a grant application authored by and awarded to AJD, NJM, TS, TGV and DW.

Huw Alexander Ogilvie, December 15th 2017

Acknowledgments

THIS THESIS IS LARGE AND CONTAINS MULTITUDES. By that I mean although I am the nominal author and wrote much of the text and code, it is the product of everyone who has supported me. Of course I first have to thank my doctoral committee of Craig Moritz, Alexei Drummond and Jason Bragg. Each of them believed in me, and I hope I have absorbed some of their copious knowledge and wisdom in the process of completing my PhD. Thanks also to Tanja Stadler who, although not an official member of my committee, oversaw a good chunk of my PhD research. Prior to my PhD my honours supervisors Michael Djordjevic and Nijat Imin, and my honours examiners John Rathjen and Ulrike Mathesius, taught me much about research and academia — preparing me for what was to come.

For the last five years I have been part of a community of honours students, PhD students, and postdocs, and it was understood by everyone that we are all in this together. In alphabetical order, the following people from this community made life and work much easier for me: Adam, Alex P, Alex G, Ana, Ariel, Armand, Bec, Bo, Chi, Christiana, Connie, Damien, Dan, David, Duncan, Eli, Fábio, Ian, Izzy, Jack, Jared, Jessica, Jessie, Jo, Joëlle, John, Jordan, Josh, Jūlija, Karen, Kelly, Kevin, Lauren, Leo, Liam, Louisa, Marta, Meenu, Megan, Michael T, Mitzy, Mozes, Nadia, Nick, Nicola, Ollie, Paul, Regina, Sally, Sasha, Stephen, Susan, Tim, Vero, Veronica, Veronika, Viri, and Xia. If anyone is missing from that list, it is because my memory is temporarily out-of-order as a side effect of writing a thesis.

My parents, Teresa and Ian, my late grandparents, Graeme and Pix, and my brother Martin made me who I am today. My grandfather was a research scientist at CSIRO, and my admiration for him certainly predisposed me towards a career in science.

In addition to my family and academic buddies, thanks to everyone I know in a non-science, non-blood-relative capacity (although one of them has a PhD, it's in evolutionary computation and not computational evolution). Especially Rico, Shan Shan, He Chuan, Robbie, John, Ben, Nick, Dzung, Darryl, Yin, Damien, Ron and Miri. Benji, the best dog in the world, helped too.

I am legally required to acknowledge that “This research is supported by an Australian Government Research Training Program (RTP) Scholarship”

Abstract

So much research builds on evolutionary histories of species and genes. They are used in genomics to infer synteny, in ecology to describe and predict biodiversity, and in molecular biology to transfer knowledge acquired in model organisms to humans and crops. Beyond downstream applications, expanding our knowledge of life on Earth is important in its own right. From *Naturalis Historia* to *On the Origin of Species*, the acquisition of this knowledge has been a part of human development.

Evolutionary histories are commonly represented as trees, where a common ancestor progressively splits into descendant species or alleles. Time trees add more information by using height to represent genetic distance or elapsed time. Species and gene trees can be inferred from molecular sequences using methods which are explicitly model-based, or implicitly assume or are statistically consistent with a particular model of evolution. One such model, the multispecies coalescent (MSC), is the topic of my thesis. Under this model, separate trees are inferred for the species history and for each gene's history. Gene trees are embedded within the species tree according to a coalescent process.

Researchers often avoid the MSC when reconstructing time trees because of claims that available implementations are too computationally demanding. Instead, the species history is inferred using a single tree by concatenating the sequences from each gene. I began my thesis research by evaluating the effect of this approximation. In a realistic simulation based on parameters inferred from empirical data, concatenation was grossly inaccurate, especially when estimating recent species divergence times. In a later simulation study I demonstrated that when using concatenation, credible intervals often excluded the true values.

To address reluctance towards using the MSC, I developed a faster implementation of the model. “StarBEAST2” is a Markov chain Monte Carlo (MCMC) method, meaning it characterizes the probability distribution over trees by randomly walking the parameter space. I improved computational performance by developing more efficient proposals used to traverse the space, and reducing the number of parameters in the model through analytical integration of population sizes.

Despite its sophistication, the MSC has theoretical limitations. One is that the substitution rate is assumed to stay constant, or uncorrelated between lineages of different genes. However substitution rates do vary and are associated with species traits like body size. I addressed this assumption in StarBEAST2 by extending the MSC to estimate substitution rates for each species. Another assumption is that genetic material cannot be transferred “horizontally”, but a more general model called the multispecies network coalescent (MSNC) permits introgression of alleles across species boundaries. My collaborators and I have developed and evaluated an MCMC implementation of the the MSNC.

My final thesis project was to combine the MSC with the fossilized birth-death (FBD) process, which models how species are fossilized and sampled through time. To demonstrate the utility of the FBD-MSNC model, I used it to reconstruct the evolutionary history of Caninae (dogs and foxes) using fossil data and molecular sequences.

Contents

0	INTRODUCTION	1
0.1	Inferring trees from data	2
0.2	Inferring species trees	4
0.3	Evaluating and accelerating the multispecies coalescent	6
0.4	Extending the multispecies coalescent	8
1	COMPUTATIONAL PERFORMANCE AND STATISTICAL ACCURACY OF *BEAST AND COMPARISONS WITH OTHER METHODS	13
1.1	Introduction	14
1.2	Methods	17
1.3	Results	25
1.4	Discussion and conclusions	38
1.5	Supplementary information and data	46
1.6	Acknowledgements	46
2	STARBEAST2 BRINGS FASTER SPECIES TREE INFERENCE AND ACCURATE ESTIMATES OF SUBSTITUTION RATES	47
2.1	Introduction	48
2.2	New approaches	52

2.3	Results and discussion	57
2.4	Conclusions	70
2.5	Materials and methods	71
2.6	Supplementary material	78
2.7	Acknowledgments	78
3	BAYESIAN INFERENCE OF SPECIES NETWORKS FROM MULTILOCUS SEQUENCE DATA	79
3.1	Introduction	80
3.2	New Approaches	81
3.3	Simulations	93
3.4	Analysis of biological data	99
3.5	Discussion	104
3.6	Supplementary material	108
3.7	Acknowledgments	108
4	INFERRING SPECIES TREES USING INTEGRATIVE MODELS OF SPECIES EVOLUTION	109
4.1	Introduction	110
4.2	Methods	115
4.3	Results	119
4.4	Discussion	127
4.5	Acknowledgments	131
5	CONCLUSION	133
5.1	Beyond Markov chain Monte Carlo	134
5.2	Modelling morphological evolution	135

5.3 Final remarks	136
APPENDIX A ALGORITHMS USED IN CHAPTER 3	137
APPENDIX B SUPPLEMENTARY FIGURES FOR CHAPTER 4	143
BIBLIOGRAPHY	149

O

Introduction

Phylogenetic trees are one of the fundamental structures used to understand biology. They can represent the evolutionary history of species and genes, where they are dubbed “species trees” and “gene trees” respectively. Trees can even be used to model the evolution of epidemics (Stadler *et al.*, 2012) and languages (Bouckaert *et al.*, 2012). The unifying theme of my thesis is the inference of species trees, and the more general case of species networks, using a probabilistic model called the multispecies coalescent or the multispecies network coalescent respectively.

Each species tree is part of the “tree of life” that connects all life on Earth. In the case of humans, species trees place us within great apes, great apes within primates, primates within mammals, and so on through to the origin of life on Earth (Sibley and Ahlquist, 1984; Goodman *et al.*, 1998; Murphy *et al.*, 2001).

Knowledge of our origins, and those of other organisms, is something worth striving for in and of itself. Charles Darwin travelled around South America and the world for five years to acquire this knowledge (Darwin, 1839). Evolutionary history is the subject of successful popular science books such as *The Monkey's Voyage* (de Queiroz, 2014), and film and television documentaries such as *Life on Earth* (BBC and Warner Bros., 1979). The coverage of and interest in evolutionary findings outside academia reveals a deep desire know where species come from.

Accurate species trees underpin comparative biology and genomics, and inform value judgments in conservation biology. One measure of biodiversity is phylogenetic diversity (PD), which uses the distance between species in a tree to quantify evolutionary diversity. PD can be used to map biodiversity across a landscape, which identifies geographic areas most in need of conservation protection (Rosauer *et al.*, 2017).

Another practical application is to inform research into agronomically important traits. Species trees reveal that the predisposition to evolving nitrogen-fixing nodulation symbiosis is limited to the nitrogen-fixing clade of angiosperms (Doyle, 2011). Phylogenomic differences between this clade and other angiosperms, and between sister species within this clade where only one species is capable of nodulation symbiosis, may reveal its molecular basis (Geurts *et al.*, 2012).

0.1 INFERRING TREES FROM DATA

Trees can be inferred from multiple sequence alignments (MSAs) using probabilistic or non-probabilistic methods. In the past non-probabilistic methods were more common as they require less computational power or were easier to implement and understand. Maximum parsimony identifies the tree or trees which minimise the number of mutations necessary to explain the site patterns observed in an MSA (Fitch, 1971). Neighbor-joining builds a tree

heuristically from the tips upwards based on the genetic distances between pairs of taxa in an MSA (Saitou and Nei, 1987).

Phylogenetic likelihood methods are probabilistic methods based on the likelihood of observing an MSA, conditional on a phylogenetic tree and a substitution model (Felsenstein, 1981). Maximum likelihood methods use a heuristic algorithm to identify the single most likely tree, and Bayesian likelihood methods characterise the posterior distribution of trees. Ideally the posterior distribution will include all plausible clades.

All of these methods can be used to estimate the topology of a tree, which for a species tree corresponds to the taxonomic relationships. Maximum parsimony is a statistically inconsistent estimator of the species tree topology (Felsenstein, 1978), whereas in practice neighbor-joining appears to be a good estimator of the species tree topology (Rusinko and McPartlon, 2017).

Likelihood methods are able to estimate time trees by adding one or more clock rate parameters. The branch lengths, and the distance from a node to the tips of the tree, of a time tree are in units of time, making likelihood methods indispensable for inference where absolute or relative times are important (Drummond *et al.*, 2006). This is the case for many applications, including molecular epidemiology, historical biogeography and studies of speciation.

The characters used for any of these methods can be phenotypic or molecular. Molecular characters are extremely popular for phylogenetic inference because the cost of sequencing DNA has fallen super-exponentially over the past 15 years (Hayden, 2014). It is now possible for approximately 5,000 USD to sequence and produce a largely complete *de novo* assembly of a 722 million base pair genome (Paajanen *et al.*, 2017). In contrast phenotypic data sets are typically limited to tens or hundreds of characters. Besides their low cost and massive scale, molecular sequences are a good source of data for phylogenetic inference because time trees with clock models have been shown to fit empirical molecular data sets (Drummond *et al.*,

2006).

0.2 INFERRING SPECIES TREES

Inferring a tree from a single genomic locus (a gene tree), or from mitochondrial DNA (mtDNA) alone, is relatively straightforward. We assume that recombination is rare enough within short loci that all sites will share a common genealogy, and there is evidence to support at least a correlation in genealogy within individual exons (Scornavacca and Galtier, 2017). Likewise mtDNA recombination is thought to be only occasional in plants, and rare to non-existent in animals (Barr *et al.*, 2005). Therefore the inference of a single tree is likely an appropriate model for single genes or mitochondrial genomes. Note that protein and mRNA sequences may include multiple exons spanning megabases of genome, so a single genealogy cannot be assumed in those cases (Springer and Gatesy, 2016).

Species trees are typically inferred from multiple genomic loci, and one approach is to concatenate the MSAs from each locus to create a supermatrix, and inferring a single tree using phylogenetic likelihood or other methods. However genomic loci recombine during meiosis, so for any clade with sexually reproducing species, each genomic locus will have a distinct genealogy. Therefore concatenating MSAs and inferring a single tree, or “concatenation” for short, is only an approximation of the truth.

Concatenation is a statistically inconsistent estimator of the species tree topology because of discordance between the gene and species trees (Mendes and Hahn, 2017). Concatenation also overestimates species divergence times in proportion to effective population sizes (Arbogast *et al.*, 2002).

A more sophisticated model is the multispecies coalescent (MSC), where separate gene trees are estimated for each locus. Each gene tree must be compatible with a proposed species tree by following the assumption that gene flow cannot occur after speciation, so any coalescent

event between two individuals of different species must be older than the divergence time of those two species.

The species tree is estimated from the gene trees using the MSC likelihood. The coalescent likelihood is the probability of an observed distribution of coalescent times given an initial number of gene lineages and effective population size (Kingman, 1982). Under the MSC, each species tree branch has an initial number of gene lineages and a distribution of coalescent times, and the MSC likelihood is simply the coalescent likelihoods for each branch multiplied together (Degnan and Rosenberg, 2009).

Some Bayesian MSC methods like *BEAST (Heled and Drummond, 2010), StarBEAST2 (Chapter 2) and BPP (Yang, 2015; Rannala and Yang, 2017) jointly estimate gene trees from MSAs together with the species tree. I will refer to them as “fully Bayesian” methods. The fully Bayesian species tree probability $P(S|D)$ can be expressed as:

$$P(S|D) = \prod_i (P[D_i|G_i] \cdot P[G_i|S]) \cdot P(S|\theta). \quad (1)$$

The likelihood of a gene is the phylogenetic likelihood $P(D_i|G_i)$ where D_i is the MSA for the i th gene tree G_i . The MSC likelihood for that gene tree is $P(G_i|S)$ where S is the species tree. The prior probability of the species tree is $P(S|\theta)$, where θ is a vector of parameters (for example the speciation and extinction rates under a birth-death model).

Other methods have been developed which do not calculate phylogenetic or MSC likelihoods in order to quickly analyse large data sets. MP-EST (Liu *et al.*, 2010) and ASTRAL (Mirarab *et al.*, 2014a; Mirarab and Warnow, 2015; Zhang *et al.*, 2017) take frequency counts of gene tree topologies as input instead of MSAs, discarding much of the information contained in the molecular sequences. Another method, SVDquartets (Chifman and Kubatko, 2014), combines the theory of phylogenetic invariants (Cavender and Felsenstein, 1987; Lake, 1987) with a method of quartet reconciliation (e.g. Reaz *et al.* 2014) to infer species tree topolo-

gies from single nucleotide polymorphism (SNP) matrices or concatenated MSAs.

All these methods are motivated by coalescent theory and are statistically consistent estimators of the species tree topology. However they cannot estimate branch lengths in units of substitutions, and hence cannot estimate branch lengths or node heights in units of time either.

0.3 EVALUATING AND ACCELERATING THE MULTISPECIES COALESCENT

The rapid improvement in cost and capabilities of next generation sequencing means it is now possible to sequence thousands of loci from representative individuals (or a handful of loci from thousands of individuals) for phylogenetic studies. Researchers are reluctant to use fully Bayesian MSC methods with data sets this large because such methods are perceived to be too computationally demanding. This is despite the fact that for dated species trees, the only options are concatenation or fully Bayesian MSC methods. Objections to running these methods in the peer reviewed literature have included the following:

“We did not use the methods *BEAST (Heled and Drummond, 2010) or STEM (Kubatko *et al.*, 2009), because the former has shown poor performance with phylogenomic-scale data (O’Neill *et al.*, 2013), which was confirmed in this case by preliminary exploratory analyses.”

— Pyron *et al.* (2014)

“A true coalescence method such as that implemented in *BEAST (Heled and Drummond, 2010) is not possible given these data, first because the number of loci is prohibitive and because we do not have each species represented for each gene, a requirement¹ for these methods.”

— Mandel *et al.* (2015)

¹StarBEAST2 does not have this requirement at all, and neither does *BEAST if blank sequences (where all sites have a “missing” symbol) are used for the unrepresented species.

“Species-tree analysis can be problematic for UCE data sets in that the large number of loci precludes use of many coalescent-based species-tree methods”

— Streicher *et al.* (2016)

To discover if there was any basis for these concerns in reality, in Chapter 1 my colleagues and I conducted a study to quantify (1) the computational performance of the fully Bayesian method *BEAST and (2) the statistical accuracy of *BEAST compared with concatenation. If *BEAST is slow and concatenation is just as accurate, there is no reason to use fully Bayesian MSC methods.

We did find that the computational performance of *BEAST scales poorly as the number of loci is increased, following a power law. If the number of loci in a given analysis is increased from 16 to 256, we predict that it would require approximately $2400\times$ more CPU time. Because of the way *BEAST is implemented it cannot effectively use more than one CPU core so CPU time will be the same as wall time. This means that an analysis that took 2 weeks with 16 loci would take roughly 92 years with 256.

However we also found that concatenation can be far less accurate than *BEAST when analysing simulated data designed to resemble an empirical data set from the Sino-Himalayan plant clade *Cyathophora* (Eaton and Ree, 2013). For the same number of loci, *BEAST was more accurate than concatenation at estimating species tree topologies (Figure 1.4E). For any number of loci we tested (up to 4096), concatenation was less accurate than *BEAST using as few as 4 loci when estimating branch lengths (Figure 1.4A). The major component of this error was the length of branches at the tips of the tree; concatenation overestimated the lengths of tip branches by approximately 350% (Figure 1.4C).

It makes no sense to use concatenation in order to use more loci, given it will never be as accurate as fully Bayesian MSC methods, no matter how many more loci are used. Lemmon and Lemmon (2013) suggested that researchers should focus on locus selection and proper

model choice in order to increase accuracy. That said, it would be an easier pill to swallow if fully Bayesian MSC methods were faster, so in Chapter 2 my colleagues and I developed a replacement for *BEAST with better performance called StarBEAST2.

We improved the computational performance of StarBEAST2 through a combination of analytical integration, new operators and better defaults. Support for analytical integration sizes was first introduced in BEST (Liu, 2008), and we added it to StarBEAST2. Because StarBEAST2 is a Markov chain Monte Carlo (MCMC) method, it requires operators to traverse the space of phylogenetic trees. If an operator proposes a change to the species tree that is incompatible with a gene tree (or *vice versa*), the change will be rejected. So we designed new operators that make coordinated changes to the species and gene trees which will not be rejected because of incompatibility. MCMC operators each have a default weight, and we adjusted those weights through trial-and-error to improve performance.

The combination of those improvements increased the performance of StarBEAST2 by more than 13-fold when analysing empirical data sets, relative to *BEAST. StarBEAST2 is already being used by researchers to infer species trees (Tougaard *et al.*, 2017; Laver *et al.*, 2017a,b; de Magalhães *et al.*, 2017; Perrot-Minnot *et al.*, 2017) and species delimitation (Afonso Silva *et al.*, 2017), in some cases from previously intractable data sets of many loci and individuals (e.g. Moritz *et al.* 2017).

0.4 EXTENDING THE MULTISPECIES COALESCENT

While the MSC is a more sophisticated model than concatenation, it still makes some assumptions which may be violated in reality or restrict the sources of data used for phylogenetic inference. The latter chapters of my thesis describe extensions to the multispecies coalescent which relax some assumptions.

0.4.1 SPECIES TREE RELAXED CLOCKS

Previous fully Bayesian implementations such as *BEAST and BPP assumed a fixed clock, or relaxed clocks for gene trees that were uncorrelated with the species tree. However species traits such as body size, and (arboreal) tree height are known to be associated with molecular clock rates (Bromham, 2011; Lanfear *et al.*, 2013). So when relaxed clocks are used, it makes little sense for the rate of a gene tree branch to be uncorrelated with the species tree branches it is embedded within.

In Chapter 2 my colleagues and I introduced a new model where relative clock rates are estimated for each species tree branch. The rate of each gene tree branch is derived from the species tree branches it is embedded within, multiplied by a scaling factor to allow for rate variation between loci. We called this model “species tree relaxed clocks” and implemented it in StarBEAST2. We demonstrated that concatenation was less accurate than StarBEAST2 at estimating per-species clock rates simulated under this model, and that using concatenation with unphased molecular data was acutely bad at estimating those rates.

0.4.2 MULTISPECIES NETWORK COALESCENT

A core assumption of the multispecies coalescent is that gene flow ceases immediately and irrevocably after a species divergence. This assumption is violated in the case of introgression where migration/mating occurs between separate lineages of a species tree, or by hybrid species where a new species evolves with roughly equal genetic inheritance from parental species.

More and more examples of introgression and hybrid species in both plants and animals are being reported. Two North American species of *Canis*, *C. rufus* and *C. lycaon* (red wolf and great lakes wolf) are the result of hybridisation between the *C. lupus* (grey wolf) lineage and *C. latrans* (coyote) lineage (vonHoldt *et al.*, 2016). Evidence of hybrid origins has been dis-

covered for six bird species, most recently *Branta ruficollis* (red-breasted goose; Ottenburghs *et al.* 2017). Three species of *Helianthus* are the result of hybridisation between *H. annuus* (common sunflower) and *H. petiolaris* (prairie sunflower), which are not even sister lineages (Rieseberg, 1991).

Extending the MSC, the multispecies network coalescent (MSNC) introduces reticulation nodes which have two parents and a single child, and a γ value indicating the proportion of inheritance from each parent (Yu *et al.*, 2011, 2012, 2014). These reticulation nodes can model introgression and hybrid species. In Chapter 3 my colleagues and I introduced a fully Bayesian implementation of the MSNC called “SpeciesNetwork”.

This implementation is the first to use the birth-hybridization prior, which is also the first process based prior for species networks. We demonstrated its power by confirming that the purple cone spruce *Picea purpurea* is a hybrid of *P. wilsonii* and *P. likiangensis*. Because it is a fully Bayesian implementation, the absolute times of the *Picea* speciation and hybridisation events could be estimated as well as the network topology.

0.4.3 FOSSILIZED BIRTH-DEATH-MULTISPECIES COALESCENT

Fully Bayesian implementations of the MSC have until now assumed that all the data are collected from present-day organisms. However the fossil record is also a rich source of morphological character data and time calibration. It is also, increasingly, a source of ancient DNA (Shapiro and Hofreiter, 2014). The fossilized birth-death (FBD) process can be used to model the evolution of species trees containing fossil data. Bayesian FBD implementations can be used to estimate species trees containing and calibrated by fossil data (Gavryushkina *et al.*, 2014; Matzke and Wright, 2016). They can also be used with concatenated molecular sequence data for “total evidence” analyses (Gavryushkina *et al.*, 2017).

These concatenated total evidence studies will of course suffer the same problems as standalone concatenation, so in Chapter 4 my colleagues and I introduce a new integrative model

of evolution that combines the FBD and MSC models. I implemented this model which we dubbed the “FBD-MSc” in a new version of StarBEAST2 (version 14). We applied the FBD-MSc and other models (i.e. without the FBD and/or without the MSc) to a total evidence data set of the dog and fox subfamily Caninae.

We showed that estimated branch lengths and divergence times within Caninae differ between concatenation and the MSc, and that these differences are exactly what one expects due to coalescent processes. Specifically, concatenation estimates of species divergence times were consistently older than MSc estimates. The failure to account for coalescent processes qualitatively and quantitatively affected lineages-through-time curves of Caninae evolution when using FBD-concatenation instead of FBD-MSc.

1

Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods

ABSTRACT

Under the multispecies coalescent model of molecular evolution, gene trees have independent evolutionary histories within a shared species tree. In comparison, supermatrix concatenation methods assume that gene trees share a single common genealogical history, thereby equating gene coalescence with species divergence. The multispecies coalescent is supported by previous studies which found that its predicted distributions fit empirical data, and that concatenation is not a consistent estimator of the species tree. *BEAST, a fully Bayesian implementation

of the multispecies coalescent, is popular but computationally intensive, so the increasing size of phylogenetic data sets is both a computational challenge and an opportunity for better systematics. Using simulation studies, we characterize the scaling behaviour of *BEAST, and enable quantitative prediction of the impact increasing the number of loci has on both computational performance and statistical accuracy. Follow-up simulations over a wide range of parameters show that the statistical performance of *BEAST relative to concatenation improves both as branch length is reduced and as the number of loci is increased. Finally, using simulations based on estimated parameters from two phylogenomic data sets, we compare the performance of a range of species tree and concatenation methods to show that using *BEAST with tens of loci can be preferable to using concatenation with thousands of loci. Our results provide insight into the practicalities of Bayesian species tree estimation, the number of loci required to obtain a given level of accuracy and the situations in which supermatrix or summary methods will be outperformed by the fully Bayesian multispecies coalescent.

1.1 INTRODUCTION

In recent years a number of new techniques have applied next-generation sequencing to phylogenetics and phylogeography (McCormack *et al.*, 2013). These new methods include target enrichment strategies (Mamanova *et al.*, 2010) like exon capture (Bi *et al.*, 2012), anchored phylogenomics (Lemmon *et al.*, 2012) and ultra-conserved elements (Faircloth *et al.*, 2012), as well as RAD sequencing (Baird *et al.*, 2008; Davey *et al.*, 2011). As a result genome-wide samples of large numbers of loci from multiple individuals and multiple species have become increasingly common. This trend is rapidly shifting the *modus operandi* of systematic biology from phylogenetics to phylogenomics. This move to phylogenomics has also heralded a rapid development and uptake of species tree inference methods that acknowledge and model the discordance among individual gene trees. As with the field of phylogenetics, there is a broad

acceptance that probabilistic model-based methods are preferable, however the amount of data produced by next-generation technologies has also spurred the development of faster methods that do not utilize all the available data and employ statistical shortcuts such as admitting no uncertainty in individual gene trees (Kubatko *et al.*, 2009; Liu *et al.*, 2009b).

1.1.1 BAYESIAN SPECIES TREE ESTIMATION

The theory of incomplete lineage sorting and its implications for phylogenetic inference has been appreciated for some time (Pamilo and Nei, 1988), and early approaches to applying this theory inferred the species tree that minimizes deep coalescences using gene tree parsimony (Maddison, 1997; Page and Charleston, 1997; Slowinski and Page, 1999). The fully probabilistic application of the theory to molecular sequence analysis has only begun more recently with the introduction of Bayesian implementations of the multispecies coalescent (Rannala and Yang, 2003; Edwards *et al.*, 2007; Liu, 2008; Liu *et al.*, 2008; Heled and Drummond, 2010). This model embeds gene trees within a birth-death or pure Yule species tree, and within each lineage (or branch) of the species tree, gene trees are assumed to follow a coalescent process (Heled and Drummond, 2010). Prior to the development of these methods it was necessary to assume that the history of each gene is shared and equal to the history of the species tree being studied.

However, gene trees evolve within a species tree and the approximation of equating them becomes increasingly problematic as one samples more loci, when in reality each have distinct gene tree topologies and divergence times. The multispecies coalescent brings together coalescent and birth-death models of time-trees into a single model. It describes the probability distribution of one or more gene trees that are nested inside a species tree. The species tree describes the relationship between the sampled species, or sometimes, sampled populations that have been separated for long periods of time relative to their population sizes. In the latter case it may be referred to as a *population tree* instead.

The initial implementations of the multispecies coalescent made very simple assumptions including no recombination within each locus and free recombination between loci. While these simple assumptions can be robust to violation, including some forms of gene flow (Heled *et al.*, 2013) (but see Leaché *et al.* (2014)), researchers have begun to acknowledge that additional processes (such as hybridization) may need to be incorporated (Joly *et al.*, 2009; Kubatko, 2009; Chung and Ané, 2011; Yu *et al.*, 2011; Camargo *et al.*, 2012). A number of simulation studies have also looked at various facets of performance of Bayesian species tree estimation including the influence of missing data (Wiens and Morrill, 2011), the influence of low rates and rate variation among loci (Lanier *et al.*, 2014) and comparisons of performance with “supermatrix” concatenation approaches (DeGiorgio and Degnan, 2010; Larget *et al.*, 2010; Leaché and Rannala, 2011; Bayzid and Warnow, 2013).

Although these modelling advances are exciting, in the face of a next-generation data deluge, this study asks and answers the following, heretofore unanswered questions: (i) How do fully Bayesian multispecies coalescent methods scale to data sets of hundreds of loci? (ii) How much more accurate will phylogenetic species tree estimates be with more sequence data? (iii) When should one use a multispecies coalescent approach instead of computationally more efficient Bayesian supermatrix approaches, or summary methods which do not use all available data? To address the first of these questions we investigate the computational performance of the *BEAST implementation of the multispecies coalescent (Heled and Drummond, 2010), so as to assess the feasibility of conducting phylogenomic analyses using existing computational tools. To shed light on the second question we investigate how estimation accuracy improves with increasing loci.

To address the final question, we investigate how the statistical accuracy of the multispecies coalescent compares with concatenation across a broad range of conditions. We also investigate the statistical accuracy of the multispecies coalescent, supermatrix and summary meth-

ods using simulations based on two published sequence data sets; RAD tag sequences from a study of the Sino-Himalayan plant clade *Cyathophora* (Eaton and Ree, 2013), and RNA-seq assemblies from a study of primates (Perry *et al.*, 2012). *Cyathophora*, a section of the genus *Pedicularis* originating in the late Miocene or the Pliocene, is probably no older than 8 Ma (Yang and Wang, 2007) and is therefore a shallow study system. In contrast primates are a deep study system, as the oldest split in this order is estimated to have occurred in the Cretaceous around 80 Ma (Tavaré *et al.*, 2002; Steiper and Young, 2006; Wilkinson *et al.*, 2011).

1.2 METHODS

Using simulation, we investigated the trends in computational performance and statistical accuracy of the multispecies coalescent model as implemented in BEAST 2 (*BEAST), and its statistical accuracy relative to other methods of species tree inference. In designing these simulation studies there were a number of parameters to consider. The key parameters that might determine performance of inference under the multispecies coalescent are:

n : The number of species.

n_i : The number of individuals sampled per species.

n_l : The number of independent loci.

n_s : The number of sites in a single locus.

N_e : The effective population sizes of extant and ancestral species.

τ : The branch lengths in units of time or expected substitutions.

Another factor which may influence *BEAST performance is whether the molecular evolution of each locus has been more or less clock-like. Of all these parameters it is the number of loci n_l , the number of sites in a single locus n_s , and the number of individuals per species

n_i that are largely determined by experimental design. In addition, a complete specification of a multispecies coalescent model requires a speciation model (parameterized model of the species tree), a substitution model (model of the relative rates and base frequencies) and a clock model describing the absolute rate of evolution across the branches of each gene tree. In the following sections we describe the choices of parameters, models and simulation conditions for our computational experiments.

Species and gene trees for all experiments were simulated using *biopy*¹, which simulates gene trees contained within species trees according to the multispecies coalescent process. Sequence alignments were also simulated using *biopy* for experiment 1 and 2, and *Seq-Gen* (Rambaut and Grassly, 1997) was used to simulate nucleotide alignments for experiment 3.

1.2.1 EXPERIMENT 1: PERFORMANCE OF *BEAST WITH INCREASING NUMBERS OF LOCI

The first set of simulations we performed was primarily aimed at understanding the effect that increasing the number of loci has on the computational performance and statistical accuracy of Bayesian species tree estimation. We simulated 100 random (rapidly speciating) species trees of each of three different sizes, $n = 5, 8, 13$, using the birth-death process (Kendall, 1948; Nee *et al.*, 1994; Gernhard, 2008). In all cases the speciation rate was $\lambda = 1$ and the extinction rate was $\mu = 0.2$ (nominally per million years). For 5-species trees we considered $n_i = 2, 4, 8$, for 8-species trees $n_i = 2, 4$ and for 13-species trees $n_i = 2$. For each combination of n and n_i we simulated up to 256 gene trees. Gene alignments were simulated from these gene trees using an HKY substitution model (Hasegawa *et al.*, 1985) and a strict clock. All sequences were simulated with a substitution rate of 1% per lineage per million years, a transition/transversion ratio κ of 4, equal base frequencies and a strict clock. For each *BEAST analysis, the substitution rate was fixed at 1%, and a single κ value and set of base frequencies for all loci was estimated. The locus length was 200 sites each to mimic short-read next-generation sequence

¹<http://www.cs.auckland.ac.nz/~yhel002/biopy/> — accessed 15th December 2017

data. Finally, we drew successively larger subsets of each group of alignments to form a set of *BEAST analyses (Heled and Drummond, 2010). We considered increasing numbers of loci on a logarithmic scale, i.e. $n_l \in \{2, 4, 8, 16, 32, 64, 128, 256\}$.

If the effective sample size (ESS) of either the log posterior or the age of the species tree in an analysis was not ≥ 200 after the initial MCMC chain was completed, we used the *resume* function in BEAST 2 (Bouckaert *et al.*, 2014) to extend the MCMC chain from the final state of the previous run, until sufficient samples were obtained to achieve a minimum ESS of 200. For each combination of n_l , n and n_i , MCMC chains were resumed until at least 90 out of 100 replicates had sufficient ESS values. All statistics and trees were logged at a sampling rate of 1 sample per 25000 states, and the MCMC chains that needed extension were combined into a single long chain. Pseudocode for the experimental protocol can be found in Algorithm S1 in supplementary information.

ESS per hour was not calculated using the total CPU time for the combined chain because resumed runs were not restricted to a single type of CPU and hence were not directly comparable. Instead, the initial MCMC chain for each condition and replicate was restricted to a single type of CPU (Intel E5-2680 @ 2.70 GHz), and million states per hour of CPU time was calculated based on the number of states and CPU time of the initial chain. To calculate ESS per million states, the ESS of the age of the species tree was divided by the million post-burnin states in the combined chain. To calculate ESS per hour, ESS per million states was multiplied by million states per hour. All replicates were used to calculate average ESS rates, including those with ESS values < 200 .

The main measure of error used in this study, “relative species tree error,” incorporates both topological and branch length error by building on the previously described measure “rooted branch score” (RBS; Heled and Bouckaert, 2013). Given two trees T_1 and T_2 , the sets of monophyletic clades c present in each tree are defined as \mathbb{C}_1 and \mathbb{C}_2 . The length of the

branch which extends rootward from the most recent common ancestor (MRCA) of a clade is defined as $b(c)$. Given these definitions, the rooted branch score is defined as the sum of all absolute differences in branch lengths $b(c)$ between trees T_1 and T_2 :

$$RBS(T_1, T_2) = \sum_{c \in \mathbb{C}_1 \cup \mathbb{C}_2} |b^{(1)}(c) - b^{(2)}(c)| \quad (1.1)$$

By convention, the branch length of a clade that is missing from a tree is zero, so the topological error of absent or erroneous clades will be weighted by the true or estimated branch length respectively. We define the relative species tree error e_T to be the posterior expectation of the rooted branch score distance RBS between the estimated species tree \hat{T} and the true species tree T_{true} , normalized by the tree length of the true species tree L_{true} :

$$e_T = \frac{\frac{1}{k} \cdot \sum_{i=1}^k RBS(T_{true}, \hat{T}_i)}{L_{true}} \quad (1.2)$$

This measure summarizes the error over the entire posterior distribution by averaging the RBS for each i posterior sample \hat{T}_i drawn from the entire set of posterior samples of size k . We normalize by the length of the true species tree to make the error comparable between species trees of differing units and/or number of species. Replicates with insufficient ESS values were excluded when calculating average relative species tree error, because the posterior distributions of species trees for those replicates might be inadequately sampled.

A post-hoc analysis was performed to investigate the residual variation in ESS rates and relative species tree error, after accounting for the number of loci, individuals and species in each replicate. Spearman's rank correlation was used to calculate correlation coefficients between the residuals and various tree and alignment parameters. P-values for each correlation were computed using asymptotic t approximation, and then corrected for multiple comparisons based on 48 tests per set of residuals (Benjamini and Hochberg, 1995).

Mean population size was calculated as the mean of all per-branch effective population sizes. Species tree asymmetry is the variance σ_N^2 in the number of nodes between each tip and the tree root (Kirkpatrick and Slatkin, 1993). Mean tree height difference is the mean difference in height between each gene tree and the species tree. Mean deep coalescences is the mean number of deep coalescences for each gene as calculated by DendroPy 4.0.3 (Sukumaran and Holder, 2010). The mean parsimonious mutations is the parsimonious (minimum) number of mutations required per site given the true gene tree, again calculated by DendroPy. Mean variable site count is the mean number of sites per locus with more than one extant allele, and mutations per variable site is the total number of parsimonious mutations required divided by the total number of variable sites.

Experiment 1 was performed using the Pan cluster provided by New Zealand eScience Infrastructure and hosted at the University of Auckland². This high performance compute cluster provides access to Linux compute nodes with 2.7 and 2.8GHz Intel Xeon CPUs, and approximately 8GB of RAM per CPU core.

1.2.2 EXPERIMENT 2: COMPARING A BAYESIAN MULTISPECIES COALESCENT APPROACH WITH A BAYESIAN SUPERMATRIX APPROACH

In the second set of simulations we compare the statistical accuracy of the multispecies coalescent to partitioned concatenation, both as implemented in BEAST 2. We refer to these methods as *BEAST and Bayesian supermatrix respectively. Specifically we tested the hypothesis that the comparative accuracy would depend on mean branch length in coalescent units of $\tau(2N_e)^{-1}$.

For every combination of $n = 4, 5, 6, 8$ and $n_l = 1, 2, 4$ we simulated species trees with a range of branch lengths in coalescent units. In order to vary branch lengths, species trees were

²<https://www.nesi.org.nz/services/high-performance-computing/platforms> — accessed 15th December 2017

simulated with expected root heights of $R = \frac{1}{2}, 1, 2, 4, 8, 16$ (nominally in millions of years) and population sizes chosen from $N_e = \frac{1}{4}, \frac{1}{2}, 1$ (nominally in units of million individuals), changing the coalescent branch length unit numerator and denominator respectively. Additional expected root heights were included where the most accurate method switches from *BEAST to Bayesian supermatrix, to obtain denser sampling in that part of parameter space.

Species trees were generated under the pure birth Yule model (Yule, 1924). The birth rate for each combination of parameters was set to $\lambda = \frac{1}{R} \sum_{k=2}^n \frac{1}{k}$, that is, the birth rate which generates trees with an expected root height of R . These settings roughly correspond to mammalian nuclear genes of species with an effective population size of one-quarter, one half or one million individuals.

A single individual per species was simulated for all loci. We used the Jukes-Cantor substitution model (Jukes and Cantor, 1969) and a strict clock model for each locus, but with rate variation between loci. The mutation rate for the first locus was fixed at $\mu_0 = 0.01$, and the rates for other loci drawn from the range $[\mu_0/F, \mu_0 \times F]$. We used $F = 3$, giving a factor of 9 between the fastest and slowest possible rates. The rate was drawn in log space, so there is equal density of slower and faster rates around μ_0 . The number of sites per alignment (n_s) was fixed at 1000.

We generated 100 replicates for each combination of n, n_l, R and N_e . For each unique combination of n, R and N_e only one set of 100 species trees was generated and used (regardless of n_l) to minimize species tree sampling error when analyzing the effect of increasing n_l . Gene trees and extant sequences were generated separately for each replicate and for each value of n_l .

Both Bayesian supermatrix and *BEAST analyses used a Yule prior on the species tree, with a uniform prior of $[1/100, 100]$ on λ , and a separate partition per locus each with a strict clock model, where the clock rate of the first partition was fixed to the truth (μ_0) and the other rates

were estimated. The *BEAST effective population size hyperparameter (popMean) was given a uniform prior in the range $[\frac{1}{5}, 5]$, and all population sizes were estimated.

The Bayesian supermatrix analysis used a fixed chain length of 4 million states, sampling every 1000 states. The *BEAST analysis used a fixed chain length of 40 million states, sampling every 10,000 states. The ESS values of the posterior, likelihood and prior statistics of each chain were estimated, and replicates where the ESS was <200 for any of those statistics were discarded. For each combination of n , n_l and method there were never more than 4% of replicates discarded for this reason (Figure S10). As with experiment 1, this experiment was performed using the NeSI Pan cluster.

1.2.3 EXPERIMENT 3: MANY-METHOD COMPARISON OF SPECIES TREE INFERENCE USING PARAMETERS ESTIMATED FROM TWO PHYLOGENOMIC DATA SETS

The purpose of the third set of simulations was two-fold: to check that the trends in statistical accuracy observed for the first two sets of simulations held for empirically derived simulations, and to compare statistical accuracy across a range of species tree inference methods. To simulate more realistic trees and sequences, we derived a range of properties and phylogenetic parameters from two empirical phylogenomic data sets for use as simulation parameters.

The biallelic species tree inference method SNAPP (Bryant *et al.*, 2012) was used to estimate speciation birth rates and effective population sizes because it did not require phasing the sequence data. To estimate base frequencies, substitution rates, between-site rate variation and between-locus rate variation we used a Bayesian supermatrix analysis with a Yule prior on the species tree. A detailed description of sequence data processing and SNAPP and BEAST settings is given in supplementary information.

We simulated 100 replicates each of “deep” and “shallow” Yule species trees of $n = 12$ and $n = 8$ respectively, using the inferred empirical birth rates, with per-branch population sizes picked from a gamma distribution of shape 2 and a mean equal to the mean inferred

population sizes. For the deep species trees we simulated 512 gene trees, and for the shallow species trees we simulated 4096 gene trees within each species tree, each with two individuals per species.

For each simulated gene tree we chose a strict clock rate from the gamma distribution defined by the inferred shape parameters and scale parameters. Nucleotide sequences were simulated for every locus using the empirically derived GTR+G base frequencies, substitution rates and gamma rate variation from the applicable study. As the shallow study used 64nt RAD tags, we picked that fixed length for sequence simulations based on that study. For simulations based on the deep study, each simulated alignment length was randomly sampled (with replacement) from the original alignment lengths of the deep study.

Species trees were reconstructed from simulated sequences using five different multi-locus inference methods; *BEAST, Bayesian supermatrix, MP-EST (Liu *et al.*, 2010), RAxML version 8 (Stamatakis, 2014) and BIONJ (Gascuel, 1997). We tested *BEAST performance given $n_l = 1, 2, 4, 8$ for the deep study based simulations and $n_l = 1, 2, 4, 8, 16, 32$ for the shallow study based simulations. For all simulations, we tested the performance of Bayesian supermatrix given $n_l = 1, 2, 4, 8, 16, 32, 64, 128, 256, 512$. For the deep study simulations we tested RAxML, BIONJ and MP-EST with $n_l = 1, 2, 4, 8, 16, 32, 64, 128, 512$. For the shallow study simulations we also analyzed $n_l = 1024, 2048, 4096$. Both *BEAST and MP-EST can infer species trees utilizing more than one individual per species, and we tested both methods using $n_i = 1, 2$.

All GTR+G rates were estimated for *BEAST and Bayesian supermatrix analyses. For RAxML analyses, only GTR+G substitution rates were estimated and empirical base frequencies were used. Clock rate distribution parameters and clock rates for each locus were estimated for *BEAST and Bayesian supermatrix analyses. Loci were not partitioned for RAxML analyses, so per-locus clock rates could not be estimated for that method. The RAxML maxi-

imum likelihood algorithm used was “new rapid hillclimbing”. Pairwise distances matrices calculated by RAxML were used to generate neighbor-joining trees using the BIONJ algorithm implemented in PAUP* version 4.0a142³. *BEAST and BEAST trees are implicitly rooted because they are ultrametric, and RAxML and BIONJ trees were midpoint rooted.

MP-EST uses gene trees as input data, which were inferred using RAxML. The same settings used for RAxML species tree inference were used for gene tree inference, and gene trees were midpoint rooted. For each replicate MP-EST was set to make 10 independent runs, and the species tree with the highest pseudo-likelihood was retained for further analysis.

The BEAST and *BEAST chains were run on the Raijin cluster provided by the National Computational Infrastructure⁴. This cluster provides access to Linux compute nodes with 2.6GHz Intel Xeon Sandy Bridge CPUs, and 4GB of RAM was requested per run. Further details of BEAST and *BEAST chains are provided in supplementary information. RAxML and MP-EST were run on the cluster provided by the Genome Discovery Unit of the Australian Cancer Research Foundation Biomolecular Resource Facility. Jobs on this cluster ran on Linux compute nodes with a variety of Intel Xeon and AMD Opteron CPUs, and 2GB of RAM was requested per RAxML or MP-EST job.

1.3 RESULTS

1.3.1 EXPERIMENT 1: PERFORMANCE OF *BEAST WITH INCREASING NUMBERS OF LOCI

COMPUTATIONAL PERFORMANCE

We evaluated the scaling of computational performance of *BEAST as a function of the number of loci analyzed. We recorded the elapsed computational time for each replicate analysis running in a single thread. This was then used to calculate the effective number of samples per hour (ESS per hour), to measure the computational effort required to produce a sample from

³<http://paup.phylosolutions.com/> — accessed 15th December 2017

⁴<http://nci.org.au/systems-services/peak-system/raijin/> — accessed 15th December 2017

the posterior for a given number of loci. The ESS per hour relationship (Figure 1.1a,S3) suggests that a power law fits the scaling of computational performance. The linear relationship in the log-log plot indicates that a power law fits well for the range from 32 to 256 loci. We extrapolate that for $n = 5$, $n_i = 2$ and $n_l \geq 32$, ESS per hour follows a power law with a slope and intercept of -3.06 ± 0.04 and 16.34 ± 0.18 respectively.

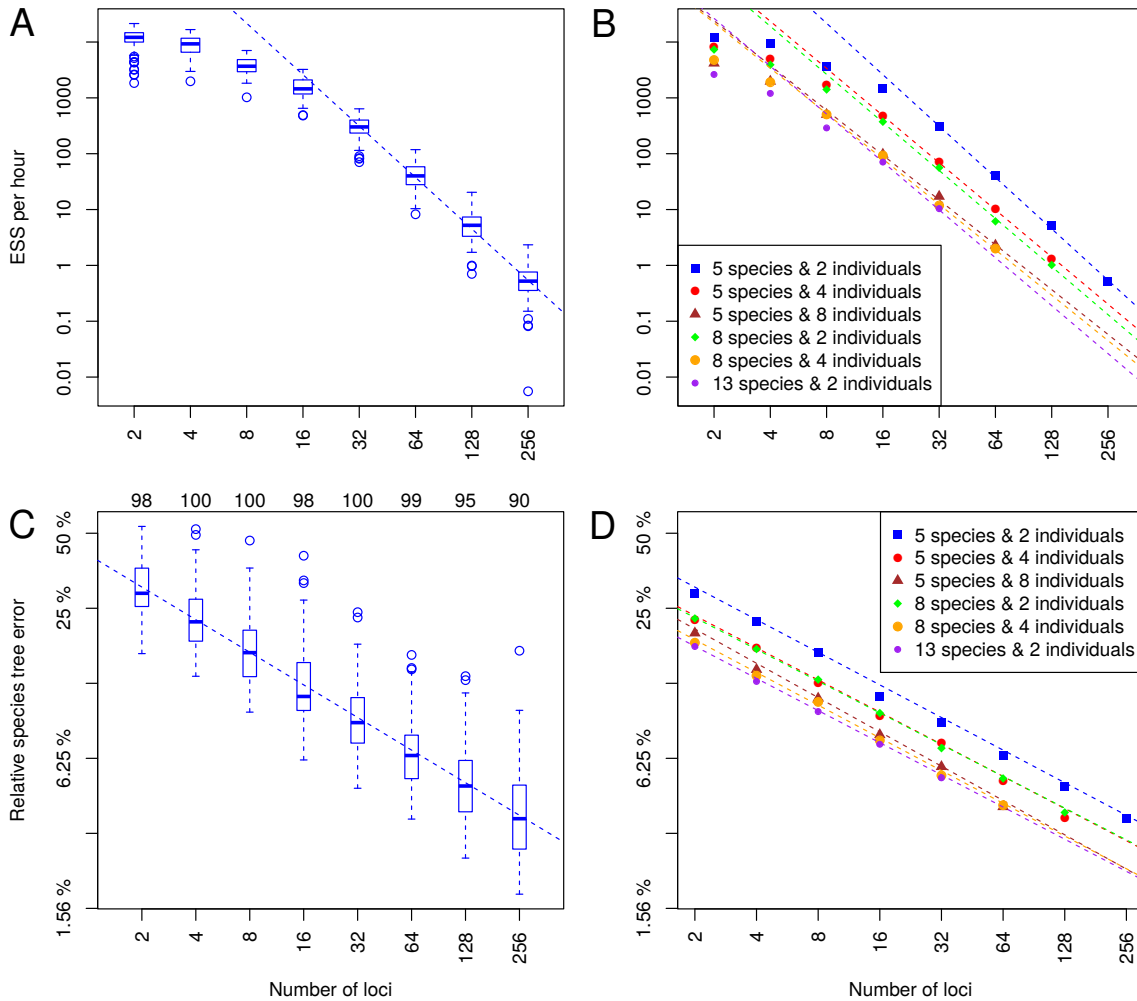


Figure 1.1: Trends in ESS per hour and relative species tree error as a function of the number of loci. (a) ESS per hour for analyses of 5 species each with 2 individuals. Each box-and-whisker shows the variance in mixing across a hundred replicate data sets for each number of loci. (b) The median ESS per hour as a function of number of loci, with trend lines for each combination of number of species and individuals per species. Solid shapes indicate the median value for each category, and regression lines were calculated using all replicates for each category. (c) Relative error for 5 species each with 2 individuals, with each box-and-whisker showing the variance in relative error between replicates. Numbers above the graph area indicate how many replicates were included for each number of loci. (d) The relative error in the estimated species tree as a function of the number of loci, with trend lines for each combination of number of species and individuals per species. Solid shapes indicate the median value for each category, and regression lines were calculated using all replicates for each category with sufficient ESS.

Applying this functional relationship, we could estimate the computational cost to analyze a similar data set with a larger number of loci. For example, given 5 species and 2 individuals

in the simulation, the predicted ESS per hour is 0.54 for 256 genes, which indicates it would take approximately 369 CPU hours to attain an ESS of 200. We can therefore estimate that a similar analysis of 1024 loci would take roughly 1064 CPU days. Nevertheless an analysis this size might be achieved within two months by parallelizing the problem into 20 independent MCMC chains for two months each and discarding a few days of burnin from each of them, to achieve on the order of ten independent samples from each chain.

Variation in ESS per hour between replicates was observed under all tested conditions (Figure S3). The slowest replicate relative to the median rate for any condition was a 5 species, 2 individuals and 256 genes outlier, $94\times$ slower than the median rate for that combination (Figure 1.1a). This replicate would require approximately 1500 CPU days to attain an ESS of 200. However, this was an extreme case as the next slowest replicate for that combination was another outlier only $6.4\times$ slower than the median rate, and would require only 100 CPU days to attain the same ESS value.

The slope of the expected computational performance as a function of number of loci does not vary with the number of species or the number of individuals (Figure 1.1b), although a larger range of n and n_i would need to be examined to understand the scaling relationship of computational performance with those quantities. For analyses larger than 5 species and 2 individuals, the power law range appears to begin at $n_l \geq 16$. Combining all simulation results, a multiple linear regression describing a response variable Y (e.g. ESS per hour) as a function of three explanatory variables: number of loci n_l , number of species n , and number of individuals per species n_i , can be constructed as follows:

$$\log(Y) = \beta_1 \log(n_l) + \beta_2 n + \beta_3 n_i + \alpha \quad (1.3)$$

Taking the ESS per hour as the response variable, the linear regression estimates of the coefficients are $\beta_1 = -2.81 \pm 0.02$, $\beta_2 = -0.42 \pm 0.01$, $\beta_3 = -0.46 \pm 0.01$, and the intercept is

$\alpha = 17.98 \pm 0.13$. At least within the range of parameters examined here, it appears that the β_1 coefficient is not greatly influenced by n and n_i (Figure 1.1b).

We also considered the scaling of the number of effective samples per million states (ESS per million states) in the MCMC analyses. This quantity is complementary to our first result; it is easier to investigate as it does not require running all simulations on identical and dedicated hardware. Computational time for methods like *BEAST is dominated by the phylogenetic likelihood, which is calculated for all site patterns given a proposed tree (Yang *et al.*, 1994). Because *BEAST infers a separate gene tree for each locus, the time per state will be linear with the number of loci assuming the average number of site patterns per locus is independent of the total number of loci. This assumption of independence holds for experiment 1 because loci were subsetted uniformly.

Adapting the terminology of Equation 1.3, the slope of ESS per hour (β_{1h}) will be simply related to the slope of ESS per million states (β_{1s}): $\beta_{1h} = \beta_{1s} + 1$. However because CPU time per site pattern depends on the specific hardware employed, the intercept of ESS per hour (α_h) cannot be predicted from that of ESS per million states (α_s).

As expected, ESS per million states also exhibits a power law in the number of loci (Figure S4). By assigning the ESS per million states to Y in the multiple linear regression in Equation 1.3, the estimated coefficients are $\beta_1 = -1.87 \pm 0.02$, $\beta_2 = -0.28 \pm 0.01$, $\beta_3 = -0.24 \pm 0.01$, and the estimated intercept is $\alpha = 9.07 \pm 0.12$. The difference in slope between ESS per million states and ESS per hour is $(-1.87) - (-2.81) = 0.94$, very close to 1 as predicted. As with ESS per hour, observations used for the linear regression were restricted to $n_i \geq 32$ for the 5 species, 2 individual case and $n_i \geq 16$ for other cases.

Using the example of 5 species and 2 individuals, the slope and intercept are -1.97 ± 0.04 and 7.86 ± 0.18 respectively, so the predicted ESS per million states for 256 individuals is 0.047 (Figure S4a). It would therefore take approximately 4.3 billion states to obtain an ESS

of 200. We can extrapolate that a similar analysis of 1024 loci would require an MCMC chain of roughly $4.3 \times \left(\frac{1024}{256}\right)^{1.97} \approx 66$ billion states.

STATISTICAL ACCURACY

We also calculated the relative error in the species tree estimate for each replicate. For some larger analyses it was challenging to achieve acceptable ESS values for every replicate, even with chain lengths of several billion states and access to high performance computational infrastructure. To retain the larger analyses without biasing statistical accuracy, we excluded replicates in which the ESS of either the log posterior or the species tree age was smaller than 200. All remaining replicates were used for a linear regression analysis of the contribution of the number of loci to relative species tree error. This analysis revealed a power law relationship from 2 to 256 loci (Figure 1.1c,S5). Given 5 species and 2 individuals, the slope and intercept are -0.435 ± 0.007 and -0.889 ± 0.026 respectively, so the relative species tree error predicted by the power law for 256 loci is 0.037. By extrapolation we would therefore estimate that the relative error of a 1024 loci analysis would decrease to $0.037 \times \left(\frac{1024}{256}\right)^{-0.435} \approx 0.020$.

Linear regression analysis of relative species tree error for all combinations of n and n_l showed little variation in the trend line slope between conditions (Figure 1.1d). By assigning the relative species tree error to Y in the multiple linear regression in Equation 1.3, the estimated coefficients are $\beta_1 = -0.433 \pm 0.003$, $\beta_2 = -0.066 \pm 0.002$, $\beta_3 = -0.070 \pm 0.002$, and the estimated intercept is $\alpha = -0.481 \pm 0.022$. More details for all multiple linear regression models are available in supplementary information. Trends in topology-only accuracy inferred using rooted Robinson-Foulds (rRF) scores are also presented in supplementary information (Figure S9, Table S12).

Finally, we also analyzed the number of species tree topologies sampled in each posterior distribution. It appears that for the analyses involving 8 and 13 species there is a rapid reduction in the number of topologies in the 95% credible set with increasing numbers of loci, but

it does not follow a power law (Figure S7).

POST-HOC ANALYSIS OF CONVERGENCE AND SPECIES TREE ERROR

Experiment 1 was designed to investigate the relationship between the number of loci n_l , number of species n and number of individuals n_i on ESS rates and statistical accuracy. While these variables explained most of the variation in ESS rates and accuracy, residual variation was present between the 100 replicates of each combination of n_l , n and n_i (Figure 1.1a,c). The correlations between this residual variation and a collection of phylogenetic statistics that could be extracted from the simulated trees and alignments were studied in a post-hoc analysis.

Table 1.1: Spearman correlation of tree and alignment parameters with ESS per hour.

	$5n, 2n_i$	$5n, 4n_i$	$5n, 8n_i$	$8n, 2n_i$	$8n, 4n_i$	$13n, 2n_i$
Species tree height	0.068	0.222***	0.362***	-0.036	0.180***	0.120
Mean population size	0.075	-0.048	-0.086	-0.020	-0.101	0.121
Species tree asymmetry	-0.238***	-0.088	-0.045	-0.125*	0.013	-0.068
Mean deep coalescences	-0.122**	-0.225***	-0.295***	0.020	-0.079	0.044
Mean parsimonious mutations	0.099	0.148***	0.122*	-0.013	0.124*	0.074
Mean variable site count	0.088	0.228***	0.294***	-0.045	0.146**	0.042
Mean tree height difference	0.246***	0.355***	0.315***	0.421***	0.340***	0.398***
Mutations per variable site	0.030	-0.066	-0.123*	0.046	0.016	0.057

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

The only tree or alignment statistic that was significantly correlated with ESS per hour consistently across all conditions was mean tree height difference (Table 1.1). This statistic is the mean difference in height between each gene tree and the species tree. The positive correlation observed for this parameter suggests that when gene trees are taller relative to the species tree, the ESS rate will be higher and *BEAST will converge more quickly.

In contrast to ESS per hour, several statistics were consistently significantly correlated with relative species tree error (Table 1.2). The height of the species tree and the number of variable sites per locus were negatively correlated with relative error. This result is somewhat intuitive, as taller species trees will have longer branches which are easier to resolve, and the number of

Table 1.2: Spearman correlation of tree and alignment parameters with species tree error.

	$5n, 2n_i$	$5n, 4n_i$	$5n, 8n_i$	$8n, 2n_i$	$8n, 4n_i$	$13n, 2n_i$
Species tree height	-0.734***	-0.582***	-0.330***	-0.702***	-0.537***	-0.580***
Mean population size	0.103*	0.078	0.006	0.118*	0.004	0.076
Species tree asymmetry	0.041	0.011	0.035	-0.170***	-0.181***	-0.050
Mean deep coalescences	0.665***	0.573***	0.273***	0.647***	0.522***	0.591***
Mean parsimonious mutations	-0.387***	-0.199***	-0.025	-0.372***	-0.184***	-0.378***
Mean variable site count	-0.587***	-0.494***	-0.242***	-0.607***	-0.530***	-0.642***
Mean tree height difference	0.194***	0.186***	0.196***	0.173***	0.207***	0.127*
Mutations per variable site	0.416***	0.306***	0.152**	0.333***	0.220***	0.148*

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

variable sites is an obvious proxy for the amount of information in each locus. Relative error was positively correlated with the mean number of deep coalescences and the number of mutations per variable site. Those correlations suggest that data sets with more incomplete lineage sorting will be more difficult to resolve, and that saturated sites may increase uncertainty.

1.3.2 EXPERIMENT 2: STATISTICAL ACCURACY OF *BEAST RELATIVE TO

BAYESIAN SUPERMATRIX

To assess the statistical accuracy of the *BEAST relative to the standard Bayesian supermatrix approach, we conducted a simulation study where we simulated species trees with a broad range of mean branch lengths for varying numbers of species and loci. Gene coalescences occur prior to species divergence times, and the severity of this discrepancy will depend on species tree branch lengths in units of coalescent time. Because the multispecies coalescent accounts for this phenomenon but the Bayesian supermatrix approach does not, we expected the multispecies coalescent to outperform the Bayesian supermatrix approach for trees with shorter branch lengths.

The “species tree error ratio” e_{T_a}/e_{T_b} is a measure of the comparative accuracy and is speci-

fied as follows, where a is *BEAST and b is Bayesian supermatrix:

$$\frac{e_{T_a}}{e_{T_b}} = \frac{\frac{1}{k_a} \cdot \sum_{i=1}^{k_a} RBS(T_{true}, \hat{T}_{ai})}{\frac{1}{k_b} \cdot \sum_{i=1}^{k_b} RBS(T_{true}, \hat{T}_{bi})} \quad (1.4)$$

Values below 1 indicate lower error, or equivalently superior accuracy, when using *BEAST instead of Bayesian supermatrix. For all numbers of species tested, the statistical accuracy of *BEAST was superior to Bayesian supermatrix for trees with shorter mean branch lengths (Figure 1.2). Using LOESS regression, it is clear that as the number of loci increases, *BEAST performance improves relative to Bayesian supermatrix because for a given mean branch length, the species tree error ratio decreases as the number of loci increases (Figure 1.2).

For all numbers of species and loci tested, there is a mean branch length crossover point where for shorter mean branch lengths, *BEAST is expected to outperform Bayesian supermatrix, and *vice versa* for longer mean branch lengths. The crossover point depends on the number of loci; as the number of loci increases, the point shifts right (Figure 1.2), indicating that *BEAST is expected to outperform Bayesian supermatrix for a larger range of mean branch lengths, consistent with the general trend of improved performance of *BEAST when increasing the number of loci.

Within the parameter region explored in this experiment, depending on the number of species, loci and the effective population sizes, the crossover point was found in the range $0.382\tau(2N_e)^{-1}$ to $5.416\tau(2N_e)^{-1}$ (Figure S11). For mean branch lengths shorter than $0.382\tau(2N_e)^{-1}$, *BEAST was preferred regardless of the parameters explored, even when using a single locus (Figure 1.2). The crossover point given a single locus was always below $0.5\tau(2N_e)^{-1}$ (Figure S11) and given longer mean branch lengths the relative performance of Bayesian supermatrix was higher than for multi-locus inference (Figure 1.2). This implies that *BEAST is still useful for single-locus studies of species trees with short branches, but should be applied with caution.

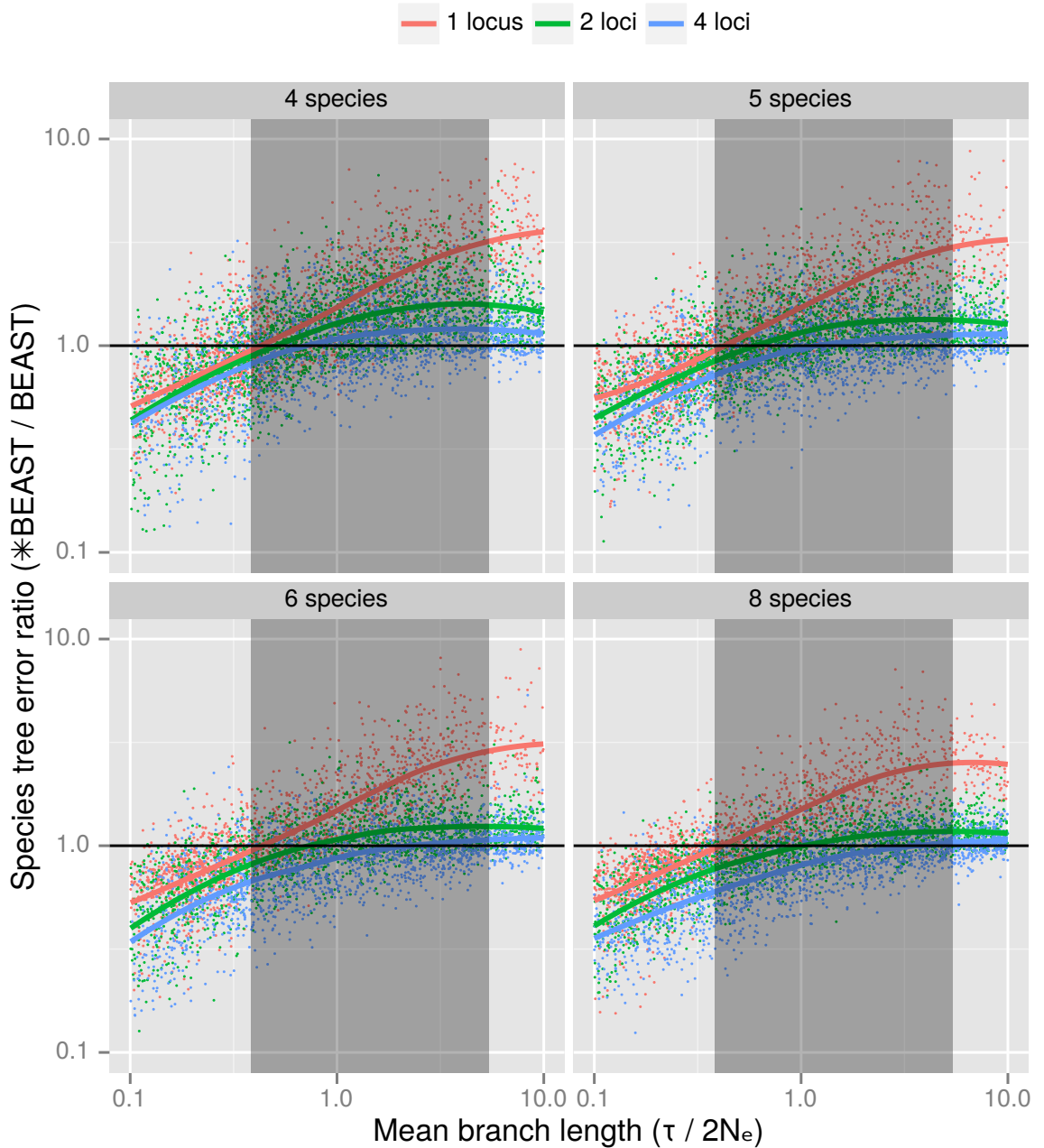


Figure 1.2: Species tree error ratio ($*BEAST/BEAST$) as a function of the average species tree branch length (in coalescent units) for trees of 4, 5, 6 and 8 species. Data points are below 1 (black line) where the $*BEAST$ error is lower than the $BEAST$ error, indicating that $*BEAST$ was more accurate than $BEAST$. Data points above 1 show the opposite. Only results with both mean branch lengths and error ratios between 0.1 and 10.0 are included. The red, green and blue lines show the local regression for one, two and four locus estimates respectively. The shaded region indicates where the crossover point depended on the combination of simulation parameters chosen — $*BEAST$ was always preferred for average branch lengths shorter than this zone.

1.3.3 EXPERIMENT 3: INFERRED PARAMETERS OF PHYLOGENOMIC DATA SETS AND MULTI-METHOD COMPARISON

Sequence data sets from two published studies were realigned and reanalyzed to calculate their empirical properties and phylogenetic parameters. Besides the expected difference in specia-

tion rate (which for the shallow study rate was over six times faster, corresponding to much shorter branch lengths), the shallow plant study sequences were very AT rich, whereas the deep primate study sequences were moderately GC rich (Table 1.3). $C \rightleftharpoons T$ substitutions were a greater proportion of all substitutions for the deep study, but the between-site gamma rate variation was flatter. The mean effective population size N_e of the deep study was estimated to be only 2.4% that of the shallow study.

Table 1.3: Experiment 3 data set properties and mean values of inferred parameters.

Phylogenetic depth	Shallow	Deep
Clade name	Cyathophora	Primates
Taxonomic rank	Section	Order
Sequence data	RAD tag	RNA-seq
In-group nS	8	12
Base frequency: A	0.290	0.266
Base frequency: C	0.212	0.240
Base frequency: G	0.204	0.263
Base frequency: T	0.294	0.231
$A \rightleftharpoons C$ rate	0.367	0.152
$A \rightleftharpoons G$ rate	0.940	0.694
$A \rightleftharpoons T$ rate	0.246	0.100
$C \rightleftharpoons G$ rate	0.305	0.155
$C \rightleftharpoons T$ rate	1.000	1.000
$G \rightleftharpoons T$ rate	0.353	0.127
Gamma rate variation	0.0383	0.233
Speciation birth rate	125.3	20.7
Per-branch N_e	6.35×10^{-3}	1.53×10^{-4}
Locus length	64nt	110–3511nt
Clock variation shape	6.22	5.15
Clock variation scale	0.173	0.195

All inferred parameters are rounded to three significant figures or one decimal place, whichever is more precise.

The original publication of *Cyathophora* sequences and phylogeny suggested that *P. rex* subsp. *rockii* is sister to subsp. *rex* and subsp. *lipskyana* (Eaton and Ree, 2013). The most common species tree topology seen in both SNAPP and Bayesian supermatrix posterior distributions supports this placement (Figure S16,S17). The original study left open the question of *P. thamnophila* monophyly but raised the possibility that the apparent paraphyly of this species, as replicated by our reanalysis, is an artifact of introgression (Eaton and Ree, 2013).

Species trees inferred by SNAPP and Bayesian supermatrix from reanalysis of the deep phylogenetic study (Figure S18,S19) agreed with the accepted primate phylogeny (Perry *et al.*, 2012).

ANALYSIS OF EMPIRICAL-BASED SIMULATIONS

We simulated species trees, gene trees and sequences based on the estimated parameters of both data sets (Table 1.3), and refer to these simulations as shallow and deep phylogenetic simulations respectively. The mean branch length of the simulated shallow species trees was $0.539\tau(2N_e)^{-1}$, compared to $159.8\tau(2N_e)^{-1}$ for the simulated deep species trees. We computed the relative species tree error for all *BEAST analyses of these simulations.

The relative species tree errors for all values of n_l and n_i considered were computed for both simulation types. A power law appeared to fit the relationship between relative error and number of loci for values of $n_l \geq 2$, so log-log linear regression analyses were restricted to $n_l \geq 2$. The log-log slope connecting relative error and the number of loci appears mostly independent of n_i for shallow phylogenetic simulations. For deep simulations, the trend lines for $n_i = 1$ and $n_i = 2$ were very close, implying that multiple individuals did not improve accuracy for those simulations (Figure 1.3).

This result is consistent with the initial set of simulations reported in “Statistical accuracy”. However, the log-log slopes varied substantially between *BEAST inference of shallow and deep phylogenetic simulations. The difference in power law exponents inferred using multiple linear regression (Table S13,S14) between shallow and deep simulations was $(-0.365) - (-0.568) = 0.203$.

Results from the initial simulation study, detailed in “Computational performance,” suggest that a power law relationship of ESS and number of loci only applies to *BEAST analyses of 16 to 32 loci and above. As we only inferred deep phylogenetic trees utilizing up to 8 loci and shallow phylogenetic trees up to 32 loci using *BEAST, we cannot make firm conclusions regarding the scaling laws of ESS performance using this set of simulations.

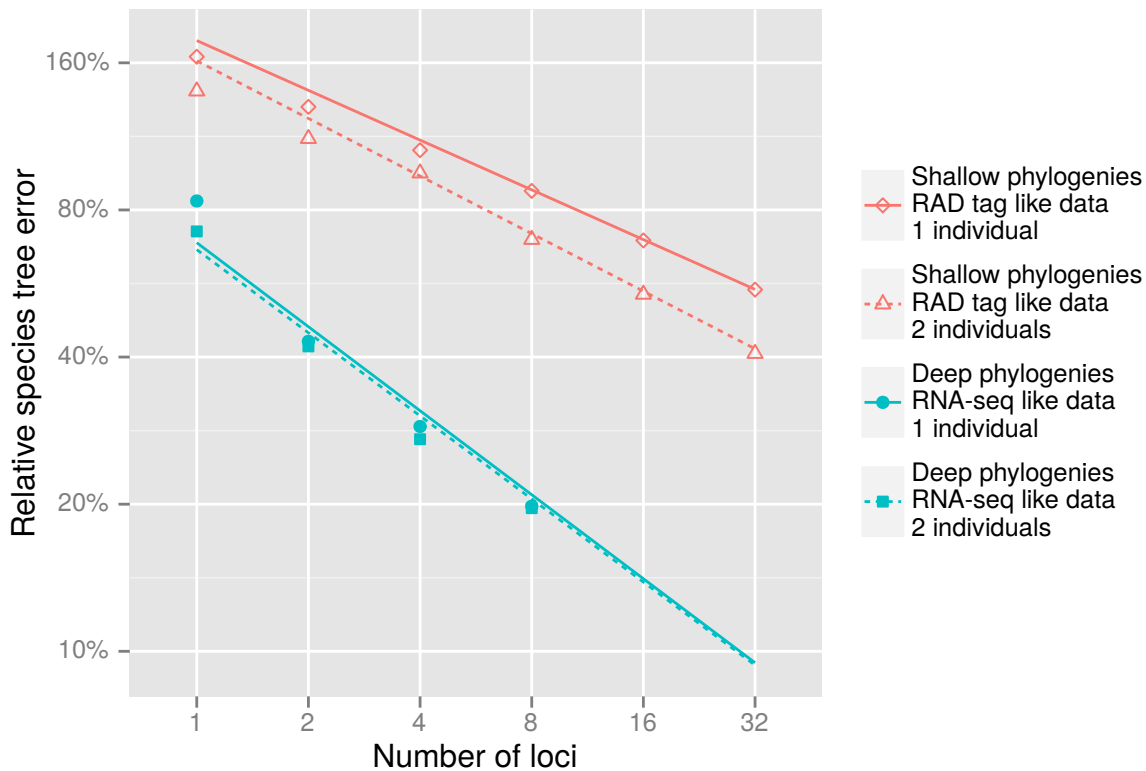


Figure 1.3: The relative species tree error as a function of the number of loci for empirical-based simulations. Both shallow and deep phylogenetic simulation results are presented. Solid and hollow shapes are the median value for each category, and regression lines were calculated using all replicates for each category.

ALTERNATIVE METHODS FOR MULTI-LOCUS PHYLOGENETIC INFERENCE

The second analysis we conducted based on the empirically derived shallow and deep phylogenetic simulations was a comparison of common multi-locus methods of species tree inference. This encompassed the Bayesian multispecies coalescent (*BEAST), Bayesian supermatrix (BEAST), Maximum-likelihood supermatrix (RAxML), neighbor-joining (BIONJ) and summary coalescent (MP-EST) methods. As some methods provide only a single best tree estimate in place of a posterior distribution of trees, we used common ancestor summary trees (CAT; Heled and Bouckaert, 2013) for *BEAST and Bayesian supermatrix analyses in this comparison.

Based on relative species tree error, *BEAST outperformed all other methods for any given number of loci for the shallow simulations. The statistical accuracy of Bayesian supermatrix, RAxML and BIONJ all plateaued beyond 64 loci for the shallow simulations, whereas *BEAST appears to follow a power law as previously suggested (Figure 1.4a). The statistical

accuracy of all methods improves with increasing numbers of loci for the deep simulations, however we limited the simulations to a maximum of 8 loci when running *BEAST. The statistical accuracy of all methods tested was similar up to 8 loci, but for larger numbers of loci Bayesian supermatrix analysis was superior and BIONJ was inferior to RAxML (Figure 1.4b).

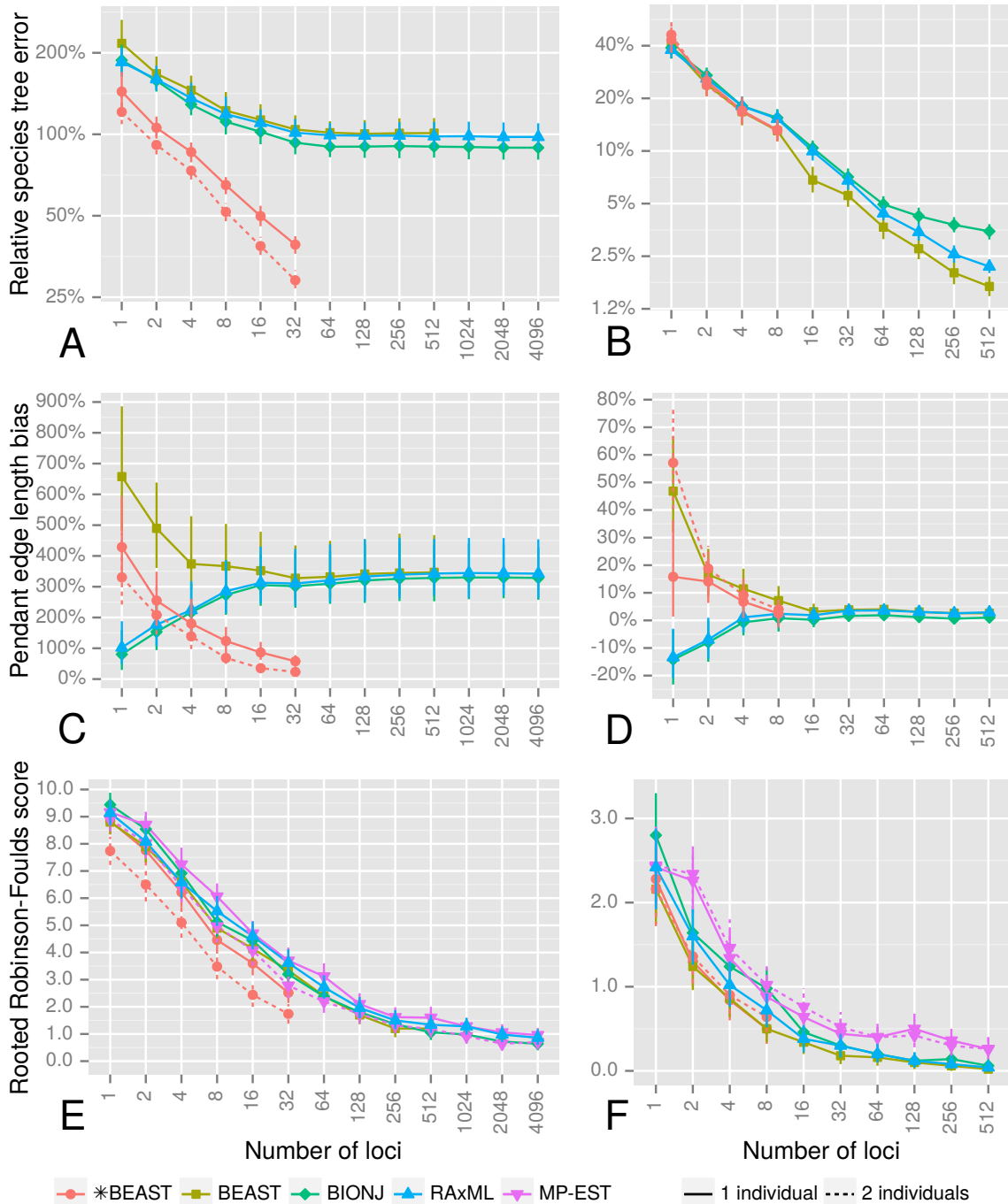


Figure 1.4: Statistical accuracy of multiple species tree inference methods as a function of the number of loci. Shallow phylogenetic simulation results (a, c, e) and deep results (b, d, f) are both presented. Measures of statistical accuracy used here are relative species tree error (a, b) which incorporates branch length and topological error, pendant edge length bias (c, d) which highlights biased branch lengths inferred by non-coalescent methods at the tips of the tree, and rooted Robinson-Foulds scores (e, f) which are a purely topological measure. All solid shapes in subfigures a-d show trimmed means (25% trim to reduce the influence of outliers), or untrimmed means for subfigures e and f. Vertical range lines show 95% confidence intervals for each mean, calculated by bootstrapping.

A major factor causing the poor performance of methods other than *BEAST for the shallow simulations is a bias when estimating pendant edge (also known as leaf or tip) length. While the mean bias of estimated pendant edge length trends towards zero for *BEAST, other methods converge on a bias of approximately 350%, meaning estimated pendant edges are on average $4.5\times$ the true length (Figure 1.4c). In contrast, there is only a small positive bias using methods other than *BEAST for the deep simulations (Figure 1.4d).

Relative species tree error incorporates both topological error and branch length error. To separate these two components we calculated the mean rRF score as a measure of purely topological error — estimated topologies more distant from the truth will have higher rRF scores. For shallow simulations, *BEAST was the best-performing method, and the topological accuracy of both *BEAST and MP-EST was improved given two individuals per species (Figure 1.4e). For deep simulations, all methods other than *BEAST and MP-EST converged at near-zero topological error given 512 loci (Figure 1.4f). *BEAST was limited to a maximum of 8 loci, but its performance for a given number of loci was very close to Bayesian supermatrix. The topological accuracy of MP-EST was inferior to all other methods analyzed.

1.4 DISCUSSION AND CONCLUSIONS

We have demonstrated by simulation that the multispecies coalescent (as implemented in *BEAST) can be applied to some problems involving hundreds of loci. In order to analyze the performance of *BEAST with hundreds of loci under various conditions, with 100 replicates per condition and given finite computational resources, we made choices partly based on computational expediency. These included relatively limited numbers of species and individuals, and assuming a strict molecular clock. More complexity in the sense of more parameters to estimate, for example denser taxon sampling or relaxed clocks, would be expected to require more computational time than the analyses reported here.

Researchers studying the evolutionary histories of organisms are not burdened by the need to test hundreds of replicates across many conditions, and can therefore conduct larger analyses using *BEAST. For example, a recent study of Neotropical cotingas (Cotingidae: Aves) applied *BEAST to resolve a species tree of 67 extant bird lineages, and used a lognormal relaxed clock for each locus with molecular rate calibrations to infer absolute divergence times. ESS rates for all logged statistics were greater than 200 and convergence was also confirmed graphically, demonstrating that *BEAST can be applied to real phylogenetic data sets with many taxa, and may also be used with a relaxed clock (Berv and Prum, 2014).

1.4.1 POWER LAWS DESCRIBE *BEAST SCALING BEHAVIOUR

For the various numbers of species, individuals and loci analyzed in this study, power laws could be used to describe the observed trends in computational performance of *BEAST, and in the statistical accuracy of the fully Bayesian multispecies coalescent. In terms of computational performance, this provides a benchmark for the efficiency of Bayesian MCMC approaches to inference under the multispecies coalescent. Our results are a product of the particular algorithm design decisions that the authors of *BEAST have made, and we hope that power law exponents can be improved upon by subsequent efforts to produce more efficient algorithms for inference under the multispecies coalescent model.

In contrast, the power law that describes the decrease in estimation uncertainty associated with inference of the species tree with increasing number of loci is a fundamental property of the model itself, and will hold regardless of the details of the algorithmic approach to inference under this model. It therefore represents a fundamental feature of the problem of species tree inference. With these results it is possible to extrapolate what one might expect to achieve by expanding data from a small pilot study to a more comprehensive sample of the genomic material of a set of study species or individuals.

The decrease in relative species tree error given different numbers of species and individ-

uals was investigated in experiment 1. Other phylogenetic parameters were fixed, including the locus length, substitution model and population size distributions. Possibly because of this, the variation in power law exponents was minimal. Experiment 3 by contrast compared shallow and deep phylogenies with larger and smaller population sizes respectively, and associated alignments of short fixed-length loci and longer variable-length loci respectively. Clock rate variation and substitution model rates also differed between conditions. Power law exponents did vary between experiment 1 and both the shallow and deep inferences in experiment 3; exponents were -0.433, -0.365 and -0.568 respectively. This is important because larger exponents imply a greater decrease in relative species tree error, so additional loci will lead to a larger improvement in accuracy of inferred species trees than with a smaller exponent.

Given a hypothetical pilot study of 16 loci, it may be of interest what the decrease in error would be for a full study of 256 loci. Because the number of loci in this scenario is increased 16 times, the reduction in relative species tree error of the full study compared to the pilot study would be $1.0 - 16^{-0.433} \approx 70\%$ if the study is similar to experiment 1, $1.0 - 16^{-0.365} \approx 64\%$ if it is similar to the shallow phylogenetic simulations, or $1.0 - 16^{-0.568} \approx 79\%$ if it similar to the deep phylogenetic simulations. What these calculations should remind us about the power law relationship is that expanding data from 1 to 16 loci provides as great an increase in statistical accuracy as expanding from 16 to 256 loci. That is, for each subsequent locus added there is a diminishing return with regards to statistical accuracy.

The power laws describing computational performance can also be used to predict the increase in computational time and chain length required to achieve sufficient sampling of the posterior distribution. In experiment 1, the power law coefficient for the log number of loci was -2.81 for ESS per hour and -1.87 for ESS per million states. Given the previous example going from 16 to 256 loci, the amount of time required for sufficient sampling of data sets similar to experiment 1 would increase by $16^{2.81} \approx 2408$ times. The chain length (number of

states) required would increase by $16^{1.87} \approx 180$ times.

Some residual variation in ESS rates was observed after accounting for the number of individuals, species and loci in each analysis. This was unsurprising as the operators used by *BEAST are stochastic (Höhna and Drummond, 2012), so even when applied to the same data ESS rates are expected to vary between runs. Consistent with this expectation, the only non-stochastic contribution identified in our post-hoc analysis was a moderate correlation between residual ESS per hour and the average gene and species tree height difference.

It is possible that the parameters which were kept constant in our analysis (e.g. the substitution rate, or the number of sites per loci, or the choice of a strict molecular clock) may change the relationship between the number of loci and computational performance or statistical accuracy. Given a sequence data set with substantially different properties from experiment 1, increasing the number of loci might have a smaller or larger effect on computational performance.

1.4.2 *BEAST COMPARED WITH OTHER METHODS

A previous simulation study which analyzed the scaling behaviour of *BEAST and other methods used just two species trees to report on topological accuracy given a range (5, 10, 25 and 50) of number of loci, and produced ambiguous results (Bayzid and Warnow, 2013). Because we simulated a new species tree for each replicate, we are able to make more general observations regarding relative performance. As expected, the relative performance of *BEAST is higher when branch lengths are shorter. The relative performance of *BEAST is also higher as the number of loci is increased (Figure 1.2).

The primary measure we chose to explore statistical accuracy, relative species tree error, incorporates both branch length and topological error. This measure is particularly relevant for molecular dating and downstream analyses of macroevolution and ecology. For example, the PD_C measure of phylogenetic diversity and the BiSSE model of binary character influence

on birth and death rates both assume accurate tree topologies and branch lengths (Maddison *et al.*, 2007; Cadotte *et al.*, 2008). When inferring species trees with shorter branch lengths, *BEAST using tens of loci outperformed supermatrix methods by this measure, even when other methods were able to utilize thousands of loci (Figure 1.4a).

If instead branch lengths are irrelevant for a study, *BEAST still outperformed other methods for a given number of loci when inferring the topology of shallow species trees (Figure 1.4e). However, when using thousands of loci, other methods were able to outperform *BEAST because *BEAST was restricted to tens of loci.

For certain species trees concatenation is statistically inconsistent (Roch and Steel, 2015) and might not outperform *BEAST even when using thousands of loci. For deeper phylogenetic trees, *BEAST performed similarly to the Bayesian supermatrix method, which in turn was superior to RAxML given larger numbers of loci (Figure 1.4b,f). Unpartitioned concatenation is known to potentially change the branch lengths and topology of estimated trees relative to partitioned concatenation (Kainer and Lanfear, 2015), so this difference may be due to method configuration rather than a quality of the statistical method employed (maximum likelihood). Regardless, as *BEAST requires substantially more computational time, concatenation methods may be preferable in this case.

Multispecies coalescent methods assume free recombination between loci, and no recombination within loci. Short sequences dispersed throughout a genome, including RAD tags, can be justifiably used with coalescent methods as violations of both assumptions are likely to be limited. However shortcut coalescence methods like MP-EST suffer from high gene tree estimation error when applied to these short sequences (Mirarab *et al.*, 2016; Springer and Gatesy, 2016). In our study MP-EST was inferior to *BEAST and similar to concatenation when inferring shallow phylogenies using short, RAD tag-like sequences (Figure 1.4e). When inferring deep phylogenies MP-EST was inferior to both *BEAST and concatenation (Figure 1.4f), de-

spite the longer loci used for those simulations.

Newer fast multispecies coalescent methods such as ASTRAL (Mirarab *et al.*, 2014a) and SVDquartets (Chifman and Kubatko, 2014) may perform better at inferring species tree topology — the latest iteration of ASTRAL is both faster and less sensitive to gene tree error than MP-EST (Mirarab and Warnow, 2015). However because these methods compute unrooted species trees without branch lengths, they cannot be compared with other methods using relative species tree error or rRF scores.

1.4.3 PRACTICAL IMPLICATIONS FOR APPLIED PHYLOGENETICS

Systematists can use the results of this study as a guide to choosing an appropriate phylogenetic method. If both *a priori* estimates or boundaries of root height (clade age) and extant effective population sizes are available for a particular study system, and the Yule process is a good fit for that system, an approximate estimate of branch length in coalescent units can be made before selecting a particular method.

Previous work has shown that the expected mean branch length of a Yule tree is equal to $1/2\lambda$ (Steel and Mooers, 2010). Under the Yule model this value is related to the expected root height:

$$\frac{1}{2\lambda} = \frac{R}{2(H_n - 1)} \quad (1.5)$$

where R is the expected root height and H_n is the n^{th} harmonic number (where n is the number of species). The expected branch length \bar{b} in coalescent units of $\tau(2N_e)^{-1}$ is therefore:

$$\bar{b} = \frac{1}{2\lambda} \cdot \frac{1}{2N_e} = \frac{1}{4} \cdot \frac{R}{H_n - 1} \cdot \frac{1}{N_e} \quad (1.6)$$

The mean root height of the shallow simulations was 0.01315, and the mean of the reciprocal extant population sizes $1/N_e$ was 302.05. The approximate branch length in coalescent

units based on these averages is:

$$\bar{b} = \frac{1}{4} \cdot \frac{R}{H_n - 1} \cdot \frac{1}{N_e} = \frac{1}{4} \cdot \frac{0.01315}{H_n - 1} \cdot 302.05 = 0.578 \quad (1.7)$$

This approximate value is quite close to the sample mean of simulated branch lengths; $0.539\tau(2N_e)^{-1}$. Based on the results of experiment 2, this value of \bar{b} is towards the lower bound of the crossover zone, and *BEAST will be preferred under most conditions (Figure 1.2). As with experiment 1, parameters which were kept constant may move this crossover point to be more or less favorable to *BEAST.

The results of experiment 3 will inform researchers with access to phylogenomic data in the order of hundreds or thousands of loci on how to select an appropriate inference method. If branch lengths are at all important, either for reporting divergence times or for downstream analyses which require a species tree, using a subset of loci with *BEAST will be superior to using all loci with other methods tested for shallow phylogenies (Figure 1.4a). If instead only the topology of the species tree is of interest, concatenation methods may be superior to fully Bayesian multispecies coalescent methods like *BEAST until improvements can be made to their computational performance (Figure 1.4e,f).

1.4.4 OPEN QUESTIONS IN PHYLOGENOMIC INFERENCE

Our results point to a number of areas for further research into the performance of species tree inference.

When using a single locus for species tree inference, experiment 2 shows Bayesian supermatrix analysis outperforming *BEAST for trees with longer branch lengths. This may be due to the population size priors used in *BEAST. However our many-method comparison shows similar performance for both methods given species trees with long branch lengths. Because deep phylogenetic trees from experiment 3 were longer than the longest trees from experi-

ment 2, this may point to a zone of intermediate branch lengths where *BEAST performs poorly given a single locus.

For all simulations we assumed a constant rate of speciation, however many lineages of life have undergone rapid radiations. It may be that when inferring species trees of clades containing ancient rapid radiations the performance of phylogenetic methods is closer to the shallow simulations than the deep simulations, and hence *BEAST becomes the preferred method.

Sequence alignments were generated and subsetting uniformly for all simulations regardless of the number of loci used for each analysis. In practice, researchers may reasonably choose longer, more informative loci when subsetting phylogenomic data sets for use with methods like *BEAST which are computationally intensive. This may improve the relative performance of *BEAST given a subset of the most informative loci relative to supermatrix or summary methods using thousands of loci.

However, whole proteins and transcripts can span genomic regions hundreds of thousands of nucleotides long, so recombination within loci will be common. The use of whole proteins or transcripts with coalescent methods has been dubbed “concatalescence” to reflect this violation (Gatesy and Springer, 2013, 2014). If these long sequences are instead split into their constituent exons, the assumption of free recombination between loci may be violated due to short intronic distances. Further studies are needed to resolve which violation is less harmful to statistical accuracy.

1.4.5 CONCLUSION AND FUTURE DIRECTIONS

The multispecies coalescent is applicable to a wider range of conditions than has been suggested by more limited simulation studies. Our results confirm that the multispecies coalescent is especially suited to the estimation of shallower evolutionary relationships. We have also demonstrated that scaling of *BEAST to problems involving hundreds of loci is feasible, however very long chains and/or crude parallelization approaches need to be employed.

We anticipate that the increasing availability of phylogenomic sequence data will motivate further improvements to the computational efficiency of fully Bayesian inference under the multispecies coalescent model, which should allow for analysis of hundreds or even thousands of loci across tens or hundreds of species. These improvements will need to scale efficiently on many-core systems such as cluster supercomputers, as such systems offer vastly greater computing power than any desktop workstation.

1.5 SUPPLEMENTARY INFORMATION AND DATA

Supplementary information and data are available at the Dryad Digital Repository.⁵

1.6 ACKNOWLEDGEMENTS

This work was supported by a Rutherford Discovery Fellowship awarded to AJD by the Royal Society of New Zealand. HAO was supported by an Australian Laureate Fellowship awarded to Craig Moritz by the Australian Research Council (FL110100104). The authors wish to acknowledge the contribution of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities to the results of this research, which are funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure program. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. The authors also thank Craig Moritz who provided valuable suggestions to improve this work.

⁵<https://doi.org/10.5061/dryad.02tf9> — accessed 15th December 2017

2

StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates

ABSTRACT

Fully Bayesian multispecies coalescent (MSC) methods like *BEAST estimate species trees from multiple sequence alignments. Today thousands of genes can be sequenced for a given study, but using that many genes with *BEAST is intractably slow. An alternative is to use heuristic methods which compromise accuracy or completeness in return for speed. A common heuristic is concatenation, which assumes that the evolutionary history of each gene tree is identical to the species tree. This is an inconsistent estimator of species tree topology, a

worse estimator of divergence times, and induces spurious substitution rate variation when incomplete lineage sorting is present. Another class of heuristics directly motivated by the MSC avoids many of the pitfalls of concatenation but cannot be used to estimate divergence times. To enable fuller use of available data and more accurate inference of species tree topologies, divergence times, and substitution rates, we have developed a new version of *BEAST called StarBEAST2. To improve convergence rates we add analytical integration of population sizes, novel MCMC operators and other optimisations. Computational performance improved by $13.5\times$ and $13.8\times$ respectively when analysing two empirical data sets, and an average of $33.1\times$ across 30 simulated data sets. To enable accurate estimates of per-species substitution rates we introduce species tree relaxed clocks, and show that StarBEAST2 is a more powerful and robust estimator of rate variation than concatenation. StarBEAST2 is available through the BEAUTi package manager in BEAST 2.4 and above.

2.1 INTRODUCTION

The throughput of sequencing technologies has improved remarkably over the past two decades culminating in next generation sequencing (NGS), and it is now feasible to sequence whole or partial genomes or transcriptomes for phylogenetic studies (Lemmon and Lemmon, 2013). NGS produces hundreds or thousands of phylogenetically useful loci (see for example Blom *et al.*, 2016) with potentially millions of sites spread across a data set of multiple sequence alignments.

While NGS offers hundreds or thousands of loci at relatively low cost, making accurate inferences from the enormous amount of data produced is particularly challenging. In the case of *BEAST, a fully Bayesian method of species tree inference which implements a realistic and robust evolutionary model in the multispecies coalescent (MSC; Degnan and Rosenberg, 2009; Heled and Drummond, 2010), it becomes exponentially slower as the number of

loci in an analysis is increased. This scaling behaviour causes *BEAST to become intractably slow after a certain number of loci (the exact number will depend on other parameters of the data set, see Chapter 1). Given the current challenges of using large phylogenomic data sets with *BEAST there have been three broad alternatives available to researchers; concatenate sequences from multiple loci, use heuristic methods statistically consistent with the MSC, or choose a tractable subset of loci to use with a fully Bayesian method like *BEAST, BEST (Liu, 2008), or BPP (Yang, 2015).

Using maximum likelihood phylogenetic methods to infer a species tree based on concatenated sequences will return the single tree that best fits the combined sequence alignment according to the phylogenetic likelihood function (Felsenstein, 1981). Popular maximum-likelihood concatenation methods include RAxML, PAML and PhyML (Stamatakis, 2014; Yang, 2007; Guindon *et al.*, 2010). Bayesian methods, such as ExaBayes and BEAST (Aberer *et al.*, 2014; Drummond and Rambaut, 2007), will instead return a distribution of trees which are probable given the combined sequence alignment, a set of priors, and the same likelihood function. Recent results show that likelihood-based concatenation can be counterproductive, producing statistically inconsistent results which assign high confidence to incorrect nodes due to model misspecification (Liu *et al.*, 2015). In the so-called “anomaly zone” of short branch lengths, the most probable gene tree topology will be different from the species tree, and estimated tree topologies will likely differ from the true species tree topology (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007).

Likelihood-based concatenation has been shown to produce systematic errors when estimating branch lengths, including overestimation of divergence times. Because some time is required for genes to coalesce looking backwards from a speciation event, the expected molecular distance between two species is greater than if coalescent events occurred simultaneously with the speciation event. This leads concatenation to overestimate the divergence

times across a species tree in proportion to effective population size (Arbogast *et al.* 2002, see also Chapter 1).

Incomplete lineage sorting (ILS) also causes systematic errors in estimated branch lengths when using concatenation. When a gene tree topology is discordant with the species tree topology, then the gene tree will contain one or more branches that define splits not occurring in the species tree. If a substitution occurs on one of these discordant gene tree branches, the resulting site pattern would define a homoplasy on the species tree, implying multiple substitutions. This effect has been termed “substitutions produced by ILS” (SPILS) and causes concatenation to overestimate the lengths of specific branches and underestimate the lengths of others, which produces apparent substitution rate variation where none exists (Mendes and Hahn, 2016). For all the above reasons, trees inferred using concatenation are therefore not a reliable approximation of the species tree in terms of branch lengths or topology.

As an alternative to concatenation for use with phylogenomic data, heuristic methods which do not perform phylogenetic likelihood calculations but are statistically consistent with the MSC have been developed. These include summary methods which utilise distributions of estimated gene tree topologies as input, such as the rooted triplet method MP-EST (Liu *et al.*, 2010) and the quartet method ASTRAL (Mirarab *et al.*, 2014a). Another quartet method is SVDquartets, which utilises single-nucleotide polymorphism (SNP) matrices (Chifman and Kubatko, 2014). Recent results show that MP-EST should be used with caution as it is sensitive to gene tree errors (Mirarab and Warnow, 2015; Xi *et al.*, 2015). At low levels of ILS, MP-EST is less accurate than likelihood-based or neighbour-joining concatenation at inferring topologies, and even at high levels of ILS it may be no more accurate than concatenation (see Chapter 1). No available heuristic method is both statistically consistent and can infer branch lengths in expected substitutions or calendar units. Therefore heuristic methods cannot be reliably used to estimate divergence times. If concatenation is used to estimate

branch lengths or divergence times for a species tree topology estimated by another heuristic method, then those estimates will be unreliable for the same reasons as pure concatenation.

An issue specific to summary methods occurs when the assumption of no recombination within loci is substantially violated because overly long loci are used (Gatesy and Springer, 2013). To resolve larger and deeper species trees using summary methods, longer and more informative loci may be required to infer more accurate gene trees. However the larger and deeper a tree, the more recombination events will have occurred. The use of longer loci and the higher incidence of recombination will both increase the risk of recombination occurring within loci, which has been dubbed the “recombination ratchet” (Springer and Gatesy, 2016).

As an alternative to increasing locus length, fully Bayesian MSC methods like *BEAST infer more accurate gene trees by sharing information between loci through the species tree (Szöllösi *et al.*, 2015). In this way very accurate species trees can be estimated using only weakly informative loci, which may not be possible using MP-EST (Xu and Yang, 2016). To avoid the recombination ratchet summary methods can use naïvely binned subsets of gene trees estimated by *BEAST (Zimmermann *et al.*, 2014), or statistically binned subsets of genes trees estimated by concatenation (Mirarab *et al.*, 2014b). Statistical binning has been criticised as statistically inconsistent (Liu and Edwards, 2015), and for either binning method the resulting species trees still cannot be used for molecular dating.

With the aim of improving the computational performance of fully Bayesian MSC inference of species trees, we have developed an upgrade to *BEAST — StarBEAST2 — which is available as a package for BEAST 2 (Bouckaert *et al.*, 2014). By improving computational performance StarBEAST2 enables the use of more loci, which will improve the precision of estimated parameters and provide an alternative to concatenation. We have also developed and include in StarBEAST2 new MSC relaxed clock models to enable accurate inference of per-species substitution rates.

2.2 NEW APPROACHES

2.2.1 ANALYTICAL INTEGRATION OF POPULATION SIZES

Markov Chain Monte Carlo (MCMC) methods like *BEAST jointly integrate over many parameters by proposing small changes at each step to eventually produce a probability distribution over all random variables. From a researcher’s perspective, some random variables may be “nuisance” parameters not of scientific interest. For example species tree topology and divergence times may be of interest, but not effective population sizes. For tractable parameters, an analytic solution will integrate over the entire range of values at each MCMC step, and may be faster than MCMC integration. However explicit estimates will not be produced so this approach is suitable only for nuisance parameters. Among-site rate variation is already integrated out at each step; the likelihood of each site is calculated for all possible discrete gamma rate categories at each step, so individual site rates are not estimated (Yang, 1994).

Analytical integration of constant per-branch population sizes was first implemented as part of BEST (Liu *et al.*, 2008), and is described in detail by Jones (2017). The analytic solution, which we have added to StarBEAST2, uses an inverse gamma conjugate prior for population sizes. By default StarBEAST2 fixes the shape of the distribution $\alpha = 3$ and only estimates the mean of the distribution μ , which is proportional to the scale parameter β :

$$\mu = \frac{\beta}{\alpha - 1} = \frac{\beta}{2} \quad (2.1)$$

In this special case where $\alpha = 3$, the standard deviation is identical to the mean:

$$\sigma = \sqrt{\frac{\beta^2}{(\alpha - 1)^2 \times (\alpha - 2)}} = \sqrt{\frac{\beta^2}{2^2}} = \frac{\beta}{2} = \mu \quad (2.2)$$

The coefficient of variation $c_v = \sigma/\mu$ of the prior distribution for effective population sizes

is therefore 1.

2.2.2 COORDINATED TREE TOPOLOGY CHANGING OPERATORS

One approach to improving the performance of MSC analyses which simultaneously estimate gene and species trees (such as *BEAST) is to develop MCMC operators which propose coordinated changes to both the species tree and the gene trees in the same step. Yang and Rannala (2014) introduced a Metropolis-Hastings (MH; Metropolis *et al.*, 1953; Hastings, 1970) operator which makes nearest-neighbour interchange (NNI) changes to the species tree topology, and simultaneously makes changes to gene tree topologies which preserve compatibility of the gene trees within the proposed species tree. Later, both Jones (2017) and Rannala and Yang (2017) introduced more general coordinated operators which make subtree prune and regraft (SPR) changes to the species tree. We have reimplemented these coordinated NNI and SPR moves in StarBEAST2 as a single new operator called “CoordinatedExchange”. Rannala and Yang (2017) also describe a proposal distribution which favours topological changes on shorter branches as well as less radical changes in topology. StarBEAST2 implements a simpler proposal distribution but still favours less radical changes by applying adjustable proposal probability weights to (less radical) NNI moves and (more radical) SPR moves.

2.2.3 COORDINATED NODE HEIGHT CHANGING OPERATORS

A novel class of coordinated Metropolis operators was introduced by Jones (2017), which pick at random a non-root non-leaf species tree node S with an existing height of $t(S)$. A new height $t'(S)$ is chosen from a uniform distribution with lower and upper bounds D and U . The height of the species tree node and the heights of subtrees of gene tree nodes (termed “connected components”) are all shifted by the amount $\eta = t'(S) - t(S)$.

The D and U bounds limit the minimum and maximum values of η to those which do not require modifying the topology of the gene tree or of the species tree, and the algorithm

to determine those bounds is given by Jones (2017). The species tree root node is excluded because there is no natural upper bound in that case. As long as the connected components are chosen with reference only to the topology of the species tree, the topology of the gene trees, and the mapping of sampled individuals to species, operators of this class are symmetric.

We have developed a new operator called “CoordinatedUniform” that belongs to this general class but has not been implemented before. Individuals from extant species which descend from a species tree node, or are directly descended from a gene tree node, are referred to here as “descendant individuals”. The gene tree nodes s selected by this operator to be shifted in height are all those for which:

1. at least one descendant individual of s is also a descendant individual of the *left* child of S
2. at least one descendant individual of s is also a descendant individual of the *right* child of S
3. all descendent individuals of s are also descendent individuals of S

An example of how gene tree nodes are selected and node heights shifted is given in Supplementary Material.

We have also developed a new adaptive MH (Andrieu and Thoms, 2008) operator called “CoordinatedExponential” which changes the height of the species tree root node and the height of connected components by an amount η . The gene trees nodes to be shifted are chosen using identical criteria as for CoordinatedUniform. Because this operator changes the height of the root node, a different method must be used to pick η compared to CoordinatedUniform.

First the lower bound D is identified in the same way as for CoordinatedUniform and as described in Jones (2017). The difference between D and the current root height is referred to as x , and a new random value x' is chosen from an exponential distribution. The value of $x' - x$ is then used for η . The median of the exponential distribution is adaptively modified over the course of an MCMC chain to equal the posterior expectation of x .

Because the proposal distribution for a new species tree root height is independent of the current height, the Hastings ratio which is usually $q(x',x)/q(x,x')$ (Hastings, 1970) can be simplified to $\pi(x)/\pi(x')$. The natural logarithm of the Hastings ratio may then be derived from the respective probability densities of x and x' drawn from an exponential distribution with the rate λ :

$$\frac{\pi(x)}{\pi(x')} = \frac{\lambda e^{-\lambda x}}{\lambda e^{-\lambda x'}} = \frac{e^{-\lambda x}}{e^{-\lambda x'}} \quad (2.3)$$

$$\therefore \ln \left(\frac{\pi(x)}{\pi(x')} \right) = \ln (e^{-\lambda x}) - \ln (e^{-\lambda x'}) \quad (2.4)$$

$$= \lambda x' \cdot \ln (e) - \lambda x \cdot \ln (e) \quad (2.5)$$

$$= \lambda (x' - x) = \lambda \eta \quad (2.6)$$

2.2.4 SPECIES TREE RELAXED CLOCKS

The overall rate of evolution occurring at a given locus within a species will be influenced by the nature of the particular gene and also by the natural history of the particular species. For a given gene, the average substitution rate may depend on the effects of selection such as the accelerated molecular evolution of sex-biased genes in *Arabidopsis thaliana* (Gossmann *et al.*, 2014), and on within-genome variation in mutation rate (Baer *et al.*, 2007). For a given species, the average substitution rate is correlated with a multitude of traits including metabolic rate, body size, and fecundity, although causal relationships are difficult to pin down (Bromham, 2011). Unsurprisingly in light of the above, empirical analysis has shown that two major factors contributing to rate variation among gene branches are the per-gene rate and the per-species rate (Rasmussen and Kellis, 2007).

Because variation is expected in the nature of different genes and species, and therefore variation is also expected in the average substitution rate of different genes and species, mul-

tispecies coalescent models should take both per-gene and per-species rate variation into account. *BEAST can accommodate both types of rate variation using gene tree relaxed clock models (for examples see Berv and Prum, 2014; Lambert *et al.*, 2015). This involves estimating per-branch substitution rates separately for each branch of each gene tree. While gene tree relaxed clocks may accommodate variation in substitution rates between species, they do not produce estimates of species branch rates. To enable accurate inference of species branch rates, we have developed a new species tree relaxed clock model.

The challenge of applying a relaxed clock to the species tree is that phylogenetic likelihood calculations require branch rates for each branch of each gene tree. Our clock model computes those rates using the total expected number of substitutions $\Sigma\mathbb{E}(S)$ accumulated along the entire length of a gene tree branch. In each species tree branch, substitutions are expected to be accumulated at the mean clock rate of the gene tree c , multiplied by the length of time L spent traversing the species tree branch, multiplied by the rates R of the corresponding species tree branch. Typical nuclear substitution rates for mammals are around 10^{-3} substitutions per site per million years (Phillips *et al.*, 2009).

The gene tree branch rates r can then be derived by dividing the total expected number of substitutions by the total length of that branch l . The gene tree branch rates for the illustrated example (Figure 2.1; Table 2.1) are therefore:

$$r_a = \frac{\Sigma\mathbb{E}(S_a)}{l_a} = \frac{0.00135}{1.5} = 0.0009 \quad (2.7)$$

$$r_b = \frac{\Sigma\mathbb{E}(S_b)}{l_b} = \frac{0.00165}{1.5} = 0.0011 \quad (2.8)$$

The new species tree relaxed clock model is available in StarBEAST2. Branch rate models that can be used with a species tree relaxed clock currently include the well-established un-

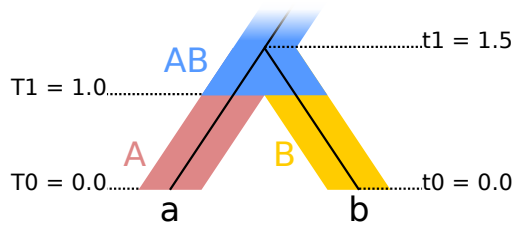


Figure 2.1: Two-species phylogeny used to illustrate species tree relaxed clocks. There are two extant species “A” and “B”, and one ancestral species “AB”. Within the species tree there is a single gene tree with extant individuals “a” and “b”. The single speciation event occurs at time T1, and the single coalescence event occurs at time t1. Gene tree rates are computed according to Table 2.1.

Table 2.1: Expected number of substitutions $\Sigma\mathbb{E}(S)$ for gene branches a, b under a species tree relaxed clock

Gene branch	Gene rate ¹ c	Length ² L			Species rate ³ R			$\mathbb{E}(S) = c \cdot L \cdot R$			$\Sigma\mathbb{E}(S)$
		A	B	AB	A	B	AB	A	B	AB	
a	0.001	1.0	0.0	0.5	0.7	1.0	1.3	0.00070	0.00000	0.00065	0.00135
b		0.0	1.0	0.5	0.7	1.0	1.3	0.00000	0.00100	0.00065	0.00165

¹ The overall substitution rate at a given locus.

² The length of a given gene tree branch within species tree branch A, B or AB.

³ The substitution rate of species tree branch A, B or AB.

correlated log-normal (UCLN) and uncorrelated exponential (UCED) models (Drummond *et al.*, 2006), as well as the newer random local clock model (Drummond and Suchard, 2010). The current species tree relaxed clock implementation estimates — separately for each species tree branch — a single relative rate. However it is possible to imagine a further relaxed model that estimates — again separately for each species tree branch — hyperparameters for a prior distribution on substitution rates. This would enable gene tree branch rates to be guided by the species tree, but still allow some difference in response between genes.

2.3 RESULTS AND DISCUSSION

2.3.1 STARBEAST2 CORRECTLY IMPLEMENTS THE MULTISPECIES COALESCENT

New methods must be shown to be correct implementations of the target model. One way to accomplish this for MCMC methods is to estimate parameters from a prior distribution using the MCMC kernel, and to also draw independent samples from the same distribution by simulation. The resulting parameter distributions should be identical if the implementation

is correct. We used this method to test the correctness of the novel features in StarBEAST2; analytical population size integration, coordinated operators, and species tree relaxed clocks. Simulated and StarBEAST2 distributions were identical for species and gene tree topologies (Figure S1,S2), species and gene tree node heights (Figure S3,S4), and for gene tree branch rates (Figure S5,S6). This combination of results supports the correctness of the StarBEAST2 implementation.

2.3.2 SPECIES TREE RELAXED CLOCKS PREVENT SPILS

When using concatenation to infer a species tree, SPILS causes apparent substitution rate variation. However in an ultrametric (time tree) framework like BEAST, branch lengths are constrained so that terminal species begin at time zero. We hypothesised that if a relaxed clock is used with concatenation in an ultrametric framework, SPILS will be absorbed as faster substitution rates for lineages that would be lengthened by SPILS in a non-ultrametric framework.

In an ultrametric framework with a strict clock and no external (e.g. fossil, biogeographical or known clock rate) calibrations, the substitution rate of each branch is set to 1. This ensures that 1 unit of time is equivalent to 1 expected substitution. Using a relaxed clock with no external calibrations the substitution rate of each branch can vary, but the expectation of the mean rate of all branches is 1, preserving the relationship of 1 unit of time = 1 expected substitution. Therefore when SPILS causes the rates of some branches to be faster than 1, the rates of some other branches will be slower than 1 to keep the expected mean constant.

We used BEAST concatenation and StarBEAST2 with a species tree relaxed clock to infer the branch lengths and substitution rates of simulated species trees with the topology (((A,B),C),D),E), using sequence alignments simulated using a strict clock. Gene tree discordance will increase the estimated length of A, B and C branches for these species trees (Mendes and Hahn, 2016), and as hypothesised substitution rates for A and B branches inferred using concatenation were biased towards being faster than the true rate of 1 (Figure 2.2). Estimated

substitution rates for the C branch were more variable, and could be faster or slower than 1. Substitution rates estimated for the D and E branches were biased towards being slower than 1, presumably to balance the mean rate. Concatenation also overestimated the lengths of tip branches, another known bias when using concatenation to infer a species tree (Chapter 1). No biases were observed for the branch rates or lengths estimated using StarBEAST2 (Figure 2.2).

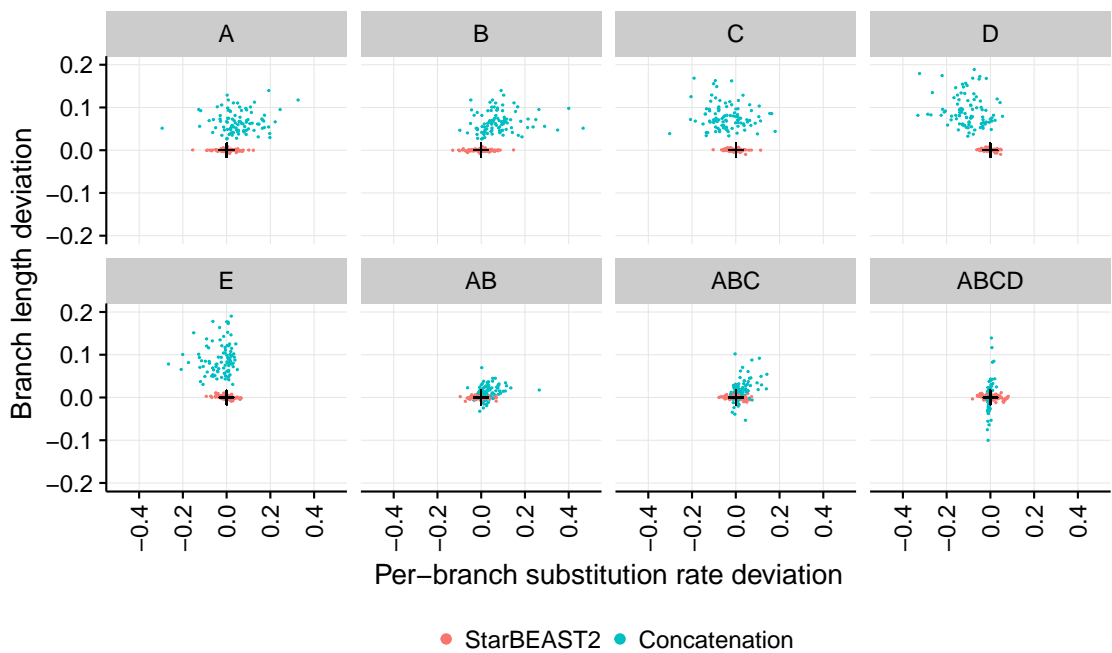


Figure 2.2: Accuracy of branch substitution rates and lengths inferred by BEAST concatenation and StarBEAST2. Deviation is the difference of each estimated rate and length from the true value. Estimated rates and lengths are the posterior expectation of the overall substitution rate and length for each species tree branch. Black crosses in each panel indicate the point of perfect accuracy. Each panel shows the distributions for the labelled extant or ancestral branch. $N = 96$.

A number of estimated branch rates had 95% credible intervals that excluded the true rate of 1 when using concatenation. If a study is testing whether substitution rates vary across a species tree, those branch rates could be erroneously interpreted as faster or slower than average. In our simulations, the clock rate of the D branch would be inferred as slower than average in 37 out of 96 replicates (Figure S7), despite the sequence data being simulated using a strict clock. When applying the same 95% credible intervals to branch lengths, the true simulated length was excluded with just two exceptions for all tip branches across all replicates using concatenation (Figure S8). In contrast, no erroneous results would be inferred for

branch rates given the same data using StarBEAST2, and out of the 768 total simulated non-root branch lengths, only five erroneous results would be inferred (Figure S7,S8).

Mendes and Hahn (2016) demonstrated that SPILS causes systematic bias when estimating branch lengths, and we show that this translates into systematic bias when estimating per-branch substitution rates. Because the bias is caused by ILS which is a function of population sizes and branch lengths, there is no reason to expect that large trees with varying population sizes and branch lengths would be any less biased.

2.3.3 DATASETS USED TO CHARACTERISE THE NEW METHODOLOGICAL APPROACHES

To characterise the performance of coordinated operators, methods of population size integration and relaxed clocks, we tested StarBEAST2 using empirical and simulated sequence data. The empirical data set used for this analysis is from the North American chorus frog genus *Pseudacris*, and was originally collected and analysed by Barrow *et al.* (2014). This data set has sequences from 26 nuclear loci across 44 sampled individuals. The individuals belong to 19 extant *Pseudacris* lineages and two outgroup species. Barrow *et al.* (2014) reported phased haplotypes but to avoid wasting computational resources we used a single haplotype per individual.

A key metric of phylogenies that can be used to judge whether it is necessary to employ MSC models is the average branch length in coalescent units $b = \tau(2N_e)^{-1}$. In this study N_e will always refer to the effective population size of diploid individuals. Given short branch lengths, likelihood-based or neighbour-joining concatenation is unable to infer accurate species trees regardless of the number of loci used, but for long branch lengths, concatenation is approximately as accurate as *BEAST (see Chapter 1). Indeed concatenation can be considered a special case of the MSC as the models converge when gene trees are identical to the species tree (Liu *et al.*, 2015). Using StarBEAST2, the average branch length within this genus was estimated to be 2.81 coalescent units. This is an intermediate average length com-

pared to the shallow simulations analysed in Chapter 1 which had a shorter average length of 1.08 coalescent units.

Each replicate of each *Pseudacris* empirical analysis used the same sequence data, and the true species tree topology, dates and rates were not known with certainty. For performance results more generally applicable than a single empirical system, and to measure the coverage and accuracy of StarBEAST2 inference, we created a simulated data set of 30 replicates. A unique species tree was simulated for each replicate, and gene trees and locus sequences were simulated according to the MSC.

We simulated 26 nuclear loci from 21 extant species with two individual haplotypes per species, very similar to the empirical data set size. The simulation parameters, including the birth rate, death rate and population sizes, were also chosen to be similar to estimated *Pseudacris* parameters. The simulated data set had an average branch length of 2.99 coalescent units, so the relative accuracy of MSC models compared to concatenation should be comparable with empirical systems like *Pseudacris*.

2.3.4 COORDINATED HEIGHT CHANGING OPERATORS AND ANALYTICAL INTEGRATION IMPROVE PERFORMANCE

To determine which configuration of new features would achieve the best performance, we ran StarBEAST2 using different combinations of operators, methods of population size integration and clock models. To measure convergence both effective sample size (ESS) per hour and ESS per million states were computed for each independent chain. ESS per hour can be used to calculate the total time required for a converged chain (nominally where ESS equals or exceeds 200), and reflects how effectively operators explore the space of trees and parameters, as well as the computational time required by each operator proposal and likelihood calculation. In contrast, ESS per million states reflects only the exploration of tree and parameter space independently of calculation times. A variety of statistics were recorded for each analysis

(Table S2-S7), and for each replicate the slowest ESS rate out of all statistics recorded for that individual chain was used for all subsequent analyses.

Multiple linear regressions with log transformed ESS rates as the response variables were used to measure the effect of coordinated topology changing operators, coordinated node height changing operators, and the method of population size integration. Each additional feature was treated as a binary indicator variable so that we could quantify the relative performance as a percentage by exponentiating the coefficient for each addition (Table 2.2).

Table 2.2: Relative performance of operators, population size integration and clock models.

Clock model	ESS rate per	Topology ³	Height ⁴	Analytical ⁵
<i>Pseudacris</i> reanalysis				
Strict	hour	73%***	120%***	130%***
Strict	million states	101%	129%***	143%***
GT-UCLN ¹	hour	72%***	289%***	100%
GT-UCLN	million states	100%	310%***	108%
ST-UCLN ²	hour	78%**	499%***	154%***
ST-UCLN	million states	95%	484%***	163%***
Simulated data				
Strict	hour	70%***	137%***	208%***
Strict	million states	100%	148%***	225%***
GT-UCLN	hour	68%***	231%***	228%***
GT-UCLN	million states	98%	248%***	248%***
ST-UCLN	hour	72%*	927%***	135%*
ST-UCLN	million states	86%	907%***	144%**

¹ Gene Tree Uncorrelated Log-Normal relaxed clock

² Species Tree Uncorrelated Log-Normal relaxed clock

³ Coordinated topology changing operators relative to naïve operators

⁴ Addition of coordinated height changing operators

⁵ Analytical integration of population sizes relative to MCMC integration

Values higher than 100% indicate faster convergence, lower than 100% indicate slower.

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. N = 30.

Coordinated topology operators consistently and significantly reduced ESS per hour, but had no significant effect on ESS per million states (Table 2.2), suggesting that coordinated topology operators are no more effective than naïve operators at proposing new states. A decrease in the number of states per hour (Figure S9) shows that they are more computationally expensive than naïve operators, and explains the negative effect on ESS per hour.

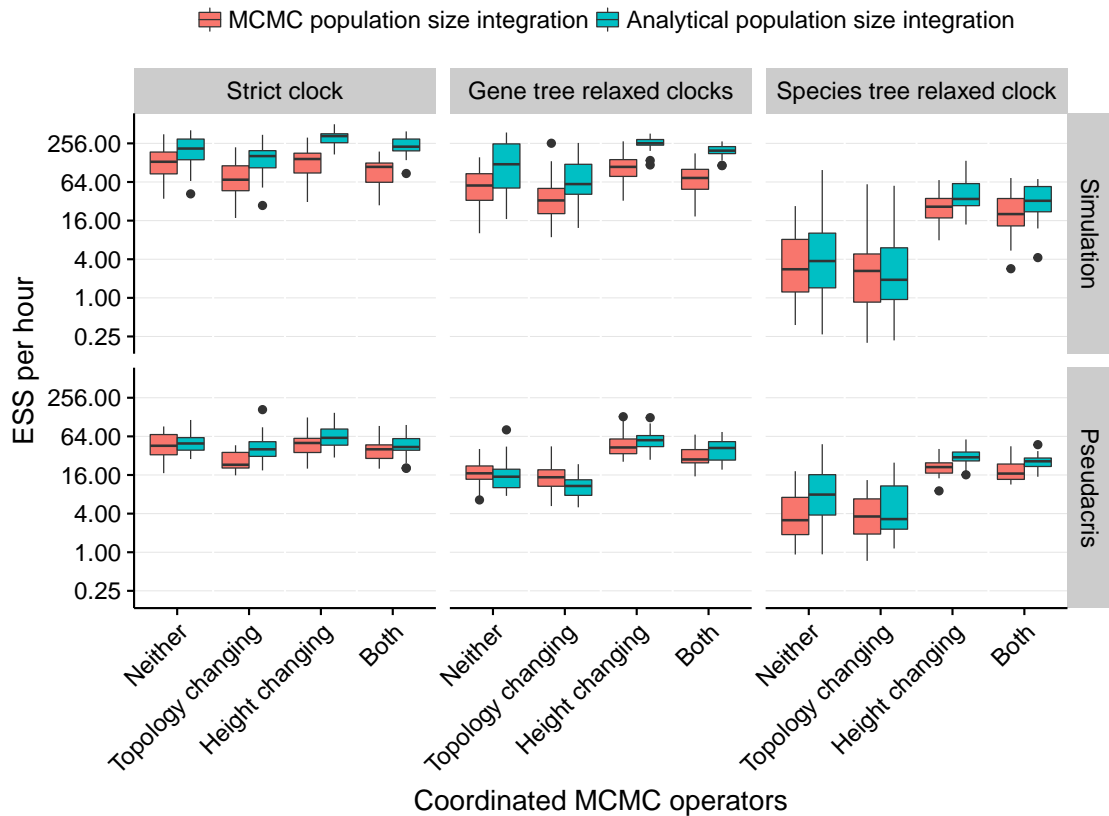


Figure 2.3: Impact of operators, population size integration and clock models on convergence. The estimated sample size (ESS) per hour for a given replicate used the smallest ESS out of all recorded statistics. Topology refers to the replacement of naïve nearest-neighbour interchange and subtree prune and regraft operators with coordinated operators. Height refers to the addition of operators which make coordinated changes to node heights. Uncorrelated log-normal relaxed clocks were applied to each gene tree (GT-UCLN) or to the species tree (ST-UCLN). N = 30.

Coordinated height changing operators consistently and significantly increased ESS per hour and per million states, however the degree of improvement depended on the clock model (Figure 2.3). For strict clock analyses the increase in ESS per hour was modest at 1.2 times and 1.37 times for empirical and simulated data respectively, whereas for species tree relaxed clocks the increase was 4.99 times and 9.27 times respectively (Table 2.2). The difference in species tree relaxed clock performance suggests that coordinated height changing operators are necessary for practical implementations of that model.

Analytical population size integration significantly improved ESS per hour and per million states performance in all cases, with the exception of gene tree relaxed clocks applied to the *Pseudacris* data set (Table 2.2).

Even with new operators and analytical population size integration, the ESS per hour rates

for species tree relaxed clocks were slower than for other clock models (Figure 2.3). One reason is that changing a species tree branch rate requires updating the phylogenetic likelihood for all gene trees, so the computational cost is much higher than for strict or gene tree relaxed clocks (Figure S9).

2.3.5 STARBEAST2 IS AN ORDER OF MAGNITUDE FASTER THAN *BEAST

StarBEAST2 also optimises the core multispecies coalescent algorithms by caching intermediate values and by using fast data structures. Operator weights have also been refined by manual iteration for better performance. Building on our results, by default StarBEAST2 enables coordinated height changing operators and analytical population size integration, but keeps naïve topology operators. To measure the combined improvement when StarBEAST2 is applied to *Pseudacris* data we compared the performance of StarBEAST2 with default settings to *BEAST. For the simulation data set, we compare StarBEAST2 with *BEAST and also with concatenation.

NGS data sets may have hundreds or thousands of loci. To gauge the performance of StarBEAST2 applied to these data sets we tested an empirical NGS data set; ultraconserved element (UCE; Faircloth *et al.*, 2012) sequences from Philippine shrews of the genus *Crocidura* (Giarla and Esselstyn, 2015). This data set consists of 1112 loci sampled from a total of 19 individuals, which belong to 9 extant lineages. Again multiple statistics were recorded to compute ESS rates for each replicate.

Our simulation study confirmed that StarBEAST2 is many times faster than *BEAST (Figure 2.4). For simulated data the average log convergence rate of StarBEAST2 with gene tree relaxed clocks was $5.54 \ln(ESS/hour)$. This compares to 2.04 using *BEAST, an increase in performance of $\exp(5.54 - 2.04) = 33.1$ times (Table S2). In fact, StarBEAST2 was $\exp(4.18 - 2.04) = 8.5$ times faster at analysing 52 loci than *BEAST was when analysing

StarBEAST2 was an order of magnitude faster when analysing either empirical data set. For gene tree relaxed clock reanalyses of *Pseudacris* the difference was $\exp(3.98 - 1.38) = 13.5$ times. For 50-locus *Crocidura* reanalyses it was $\exp(2.79 - 0.17) = 13.8$ times (Table S4,S6).

The ESS per hour convergence of species tree relaxed clocks was lower than for gene tree relaxed clocks. When applying StarBEAST2 to simulated data, gene tree relaxed clocks were $\exp(5.54 - 3.71) = 6.2$ times faster than using species tree relaxed clocks (Table S2). The difference was much smaller for empirical data; for *Pseudacris* reanalyses gene tree relaxed clocks were $\exp(3.98 - 3.44) = 1.7$ times faster, and for *Crocidura* they were $\exp(2.79 - 2.12) = 2.0$ times faster (Table S4,S6). In all three cases species tree relaxed clocks using StarBEAST2 were still faster than gene tree relaxed clocks using *BEAST (Figure 2.4).

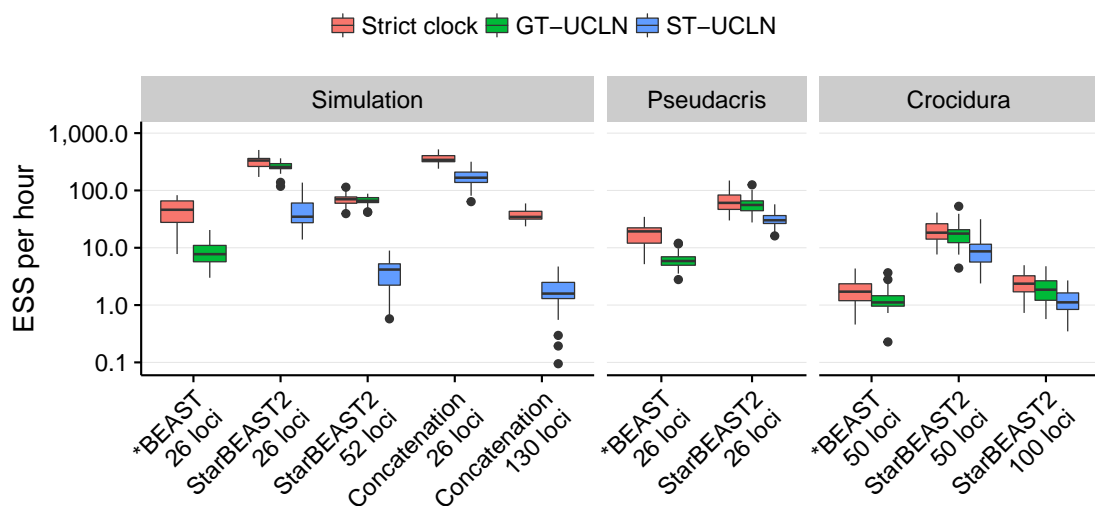


Figure 2.4: Convergence of different methods applied to simulated and empirical data sets. The estimated sample size (ESS) per hour for a given replicate used the slowest ESS rate out of all recorded statistics. Methods are BEAST concatenation, *BEAST, and StarBEAST2 with uncorrelated log-normal relaxed clocks applied to each gene tree (GT-UCLN) or to the species tree (ST-UCLN). Two *Pseudacris* *BEAST outliers with ESS rates below 0.1 are not shown. N = 30.

The increased performance of StarBEAST2 will enable researchers to analyse sequence data more quickly and disseminate their findings sooner; a large MCMC analysis which would currently take three months may now be performed in one week. In the case of phylogenomic data which has been subsetted for use with *BEAST, StarBEAST2 can be used to analyse more data for more precise estimates of species trees and other parameters in the same amount of

time as a more limited *BEAST analysis.

2.3.6 SPECIES TREE BRANCH LENGTH COVERAGE AND ACCURACY

Bayesian methods like StarBEAST2 produce both point estimates and credible intervals of inferred parameters. Ideally the point estimates will have low error, and the credible intervals will cover the corresponding true values. For well calibrated Bayesian methods, a 95% credible interval will include the true value 95% of the time.

Using a species tree relaxed clock with StarBEAST2 improved the coverage of branch length credible intervals (Figure 2.5A,B), but even when using a strict clock most simulated tip and internal branch lengths were within the corresponding credible intervals. This suggests that a strict clock model may be sufficient for studies using StarBEAST2 where substitution rate variation is not of direct interest. When using a strict clock with concatenation most internal branch lengths were outside the credible interval, but when using a relaxed clock were usually within the credible interval (Figure 2.5B). However even when using a relaxed clock, tip branch lengths were usually outside the credible intervals inferred by concatenation (Figure 2.5A).

Point estimates made by *BEAST and StarBEAST2 of both tip and internal branch lengths were more accurate than those made by concatenation (Figure 2.5C,D). The inaccuracy of tip branch lengths inferred using concatenation was driven by a strong bias towards overestimating tip branch lengths. For some replicates the sum of estimated tip branch lengths was more than double the sum of simulated tip branches lengths (Figure 2.5E). Relatively little overestimation of internal branch lengths was observed when using concatenation (Figure 2.5F).

Biased tip branch lengths are important because many published phylogenies show evidence of a slowdown in diversification rate (Moen and Morlon, 2014). If the ages of extant species are overestimated, this will artificially reduce the number of recent speciation events, mimicking a slowdown. We suggest that accurate inference of changing diversification rates

requires species trees inferred by fully Bayesian MSC methods like StarBEAST2.

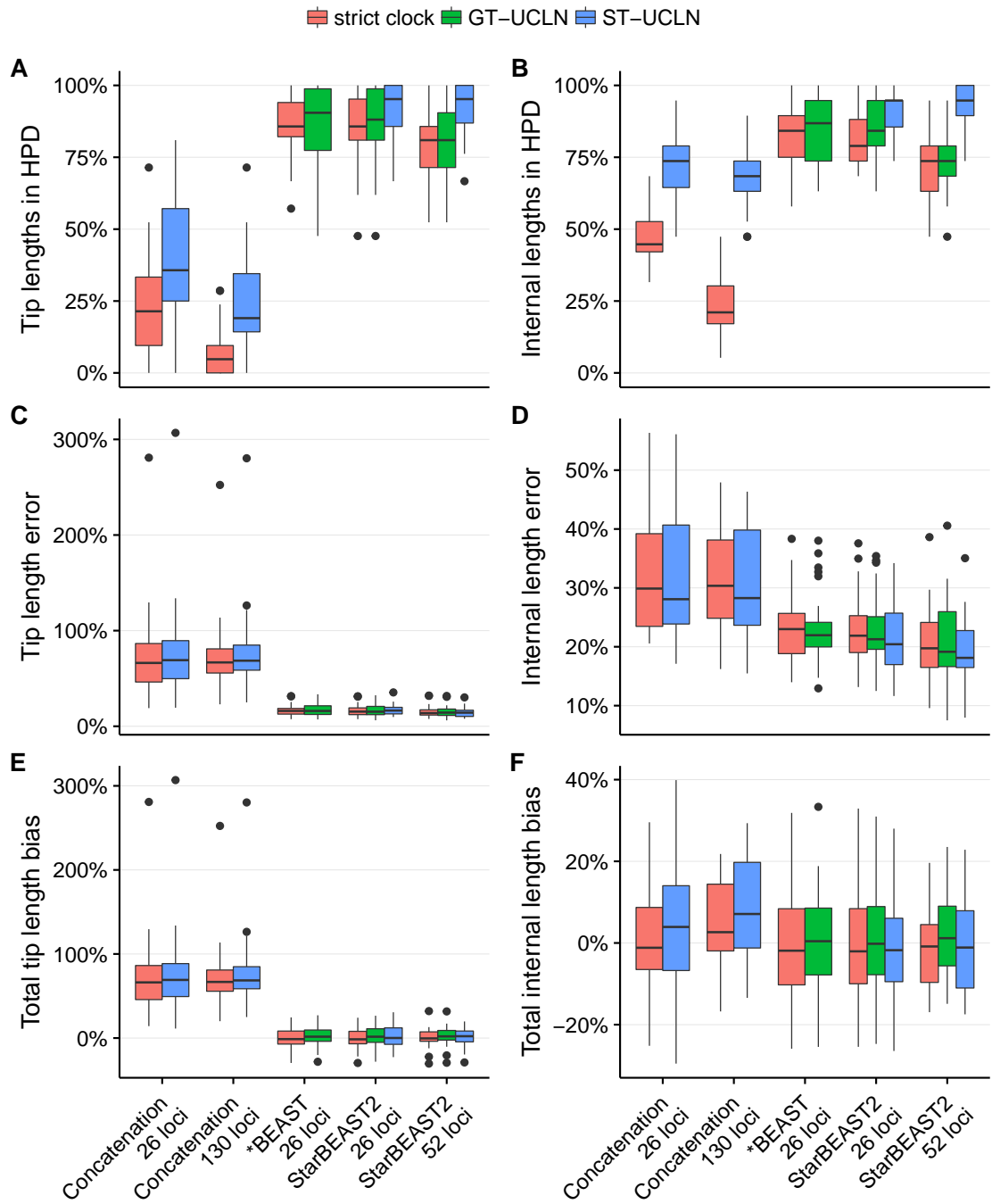


Figure 2.5: Coverage and accuracy of species branch lengths using different methods. Methods are StarBEAST2, *BEAST and BEAST concatenation with uncorrelated log-normal relaxed clocks applied to each gene tree (GT-UCLN) or to the species tree (ST-UCLN). (A,B) The percentages of true branch lengths present within the corresponding 95% highest posterior density (HPD) credible intervals. (C,D) The difference between the sum of estimated branch lengths and the sum of true branch lengths as a percentage of the sum of true branch lengths. (E,F) The sum of absolute differences between estimated and simulated branch lengths as a percentage of true tree length. $N = 30$.

Using unphased sequences with ambiguity codes for heterozygous sites improved the accuracy of concatenation by reducing the bias in tip lengths to less than 40% (Figure S10). Am-

ambiguity codes are treated by most phylogenetic methods (including BEAST) as base call uncertainty, indicating the nucleotide at a given site could be one of several possibilities. When used with unphased sequences, they actually indicate the presence of two nucleotides simultaneously, which is therefore a model violation. Using concatenation to analyse unlinked loci is also a model violation, but in the region of parameter space investigated by this simulation study the two errors may *partially* cancel out.

2.3.7 SPECIES TREE TOPOLOGY COVERAGE AND ACCURACY

We measured the coverage of species tree topologies and used the rooted Robinson-Foulds distance metric to measure the error associated with the maximum clade credibility (MCC) topology point estimates. As with branch lengths, using a relaxed clock with concatenation or a species tree relaxed clock with StarBEAST2 improved coverage. Regardless of clock model the coverage of concatenation was low; in less than 50% of replicates was the simulated topology in the credible set (Figure 2.6A).

In terms of error rates, using 130 loci was similar to StarBEAST2 using 26 loci (Figure 2.6B). Using species tree relaxed clocks with StarBEAST2 was slightly more accurate than using strict clocks, but relaxed clocks did not improve the accuracy of concatenation (Figure 2.6B). Unlike branch lengths, topological accuracy was not improved by using unphased sequences (Figure S11).

2.3.8 STARBEAST2 IS SUPERIOR AT INFERRING SUBSTITUTION RATES

While the convergence of species tree relaxed clock analyses took longer than for gene tree relaxed clocks in StarBEAST2, species tree relaxed clocks enable inference of species branch rates within an MSC framework. To gauge the accuracy of estimated branch rates, we used simple linear regressions with the true rate of each simulated branch as the explanatory variable, and the posterior expectation of the rate of that branch (conditional on the corresponding

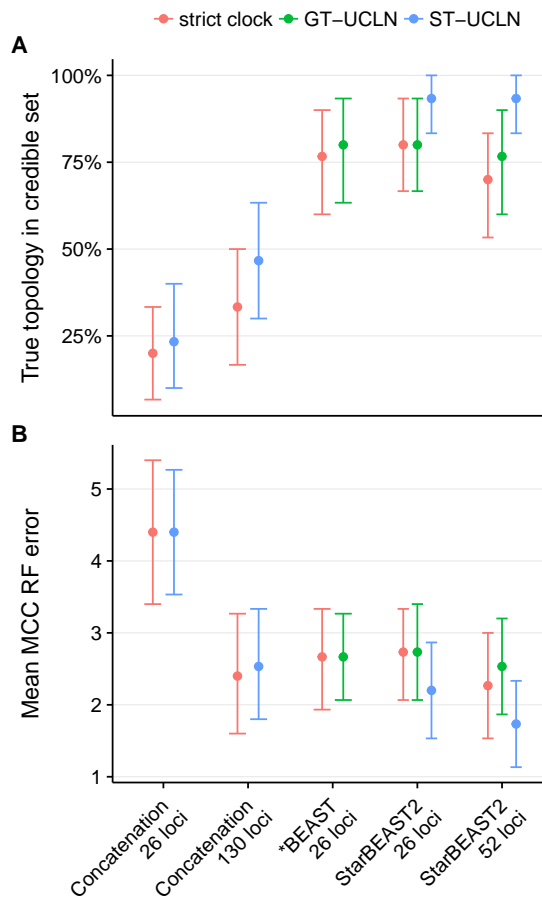


Figure 2.6: Coverage and accuracy of species tree topologies using different methods. Methods are StarBEAST2, *BEAST and BEAST concatenation with uncorrelated log-normal relaxed clocks applied to each gene tree (GT-UCLN) or to the species tree (ST-UCLN). (A) The percentage of true species tree topologies within the 95% credible set of topologies. (B) The average rooted Robinson-Foulds (RF) distance between the maximum clade credibility (MCC) species tree topology and the simulated true topology. Error bars are 95% confidence intervals calculated by bootstrapping. N = 30.

clade being monophyletic in the posterior samples) as the response variable. If all estimates are equally proportional to the truth, then the R^2 coefficient of determination will equal 1. There are intrinsic limits to our ability to estimate substitution rates, namely that branch length is confounded with substitution rate (Thorne and Kishino, 2002).

For analyses of simulated data using 26 loci the R^2 using StarBEAST2 was 0.39 and by doubling the number of loci to 52 was increased to 0.43. In contrast the R^2 when using concatenation with 26 loci was 0.26 and even after increasing the number of loci to 130 it was only 0.33, in either case worse than StarBEAST2 using 26 loci (Figure 2.7). StarBEAST2 is clearly superior to concatenation at inferring branch rates.

Concatenation is an even worse estimator of branch rates when using unphased sequences

with ambiguity codes for heterozygous sites. When applying concatenation to either 26 or 130 loci, R^2 was very weak at 0.12 regardless of the number of loci (Figure S12).

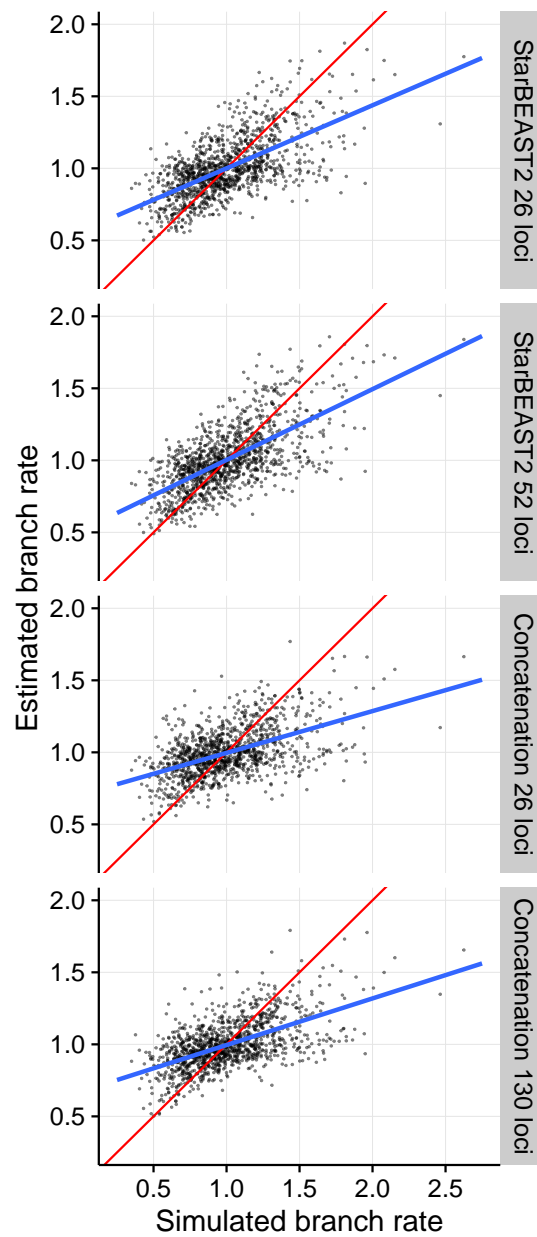


Figure 2.7: Estimates of species tree branch rates using BEAST concatenation versus StarBEAST2. Estimated rates are the posterior expectations of each branch rate from each replicate. Root branch rates, which were fixed at 1, were excluded. In blue are simple linear regression lines of best fit, and in red are the $y = x$ lines showing a perfect relationship between estimates and truth. $N = 30$.

2.4 CONCLUSIONS

When estimating divergence dates and substitution rates, the choice is often between using a subset of available loci with a fully Bayesian MSC method, or all available loci with concatena-

tion. Researchers have often opted for the second choice, but we have shown that concatenation may not accurately estimate species ages or per-species substitution rates, even for trees of intermediate branch lengths. The increased performance of StarBEAST2 should further encourage the adoption of fully Bayesian MSC methods for estimating divergence times, and the new species tree relaxed clock will enable accurate inference of species branch rates despite ILS. StarBEAST2 is free and open source software; source code, development history and multiple tutorials are available on GitHub¹.

2.5 MATERIALS AND METHODS

For all StarBEAST2, *BEAST and concatenation analyses, the version of BEAST used was 2.4.4. For all simulations, the version of biopy (Heled, 2013) used was 0.1.9. Scripts used to perform all analyses will be available on GitHub².

2.5.1 MATHEMATICAL CORRECTNESS OF STARBEAST2

Simulated trees were generated using biopy, and trees sampled from a prior distribution were generated using StarBEAST2 with all new features enabled. This included analytical integration of population sizes, coordinated tree topology and node height changing operators, and a species tree relaxed clock. 100,000 species trees were simulated, one gene tree was simulated per species tree with a rate of 0.5, and a second gene tree was simulated per species tree with a rate of 2.0.

100,000 species trees, 100,000 half rate gene trees and 100,000 double rate gene trees were sampled from the prior at a rate of one every 1000 after a 10% burn-in period.

Identical parameters were used for the simulation and for the StarBEAST2 run including the prior distributions. We fixed the number of species at 5 and the number of sampled haplo-

¹<https://github.com/genomescale/starbeast2> — accessed 15th December 2017

²<https://github.com/genomescale/starbeast2-manuscript/tree/master/scripts> — accessed 15th December 2017

types per species at 1. The birth and death rates were fixed at 200 and 100 per substitution per site respectively. Haploid population sizes were drawn from an inverse gamma distribution with shape $\alpha = 3$ and scale $\beta = 0.004$.

This procedure was repeated for both UCLN and for UCED species branch rates. Branch rates were sampled from a lognormal or exponential distribution, in either case with a mean of 1, discretised into 100 bins.

2.5.2 REANALYSIS OF *PSEUDACRIS* SEQUENCE DATA

Phased and aligned *Pseudacris* sequence data were retrieved from Dryad³. Replicating the original analysis we applied the HKY nucleotide substitution model (Hasegawa *et al.*, 1985) to 22 out of 26 nuclear loci and the GTR model (Tavaré, 1986) to the remaining 4. For all models we used four discrete gamma categories to accommodate among-site rate variation (Yang, 1994). Transition/transversion rates and ratios and rate variation shape parameters were estimated, and empirical base frequencies used, all separately for each locus. The relative substitution rate of each locus was estimated using a lognormal prior with a mean μ in real space of 1 and a standard deviation σ of 0.6. We used a single haplotype sequence per individual per locus, halving the total number of sampled sequences to avoid wasting computational resources.

For inference of *Pseudacris* trees, we ran 30 independent StarBEAST2 chains for all 24 conditions for a total of 720 chains. The conditions were each possible combination of strict, species tree relaxed or gene tree relaxed clocks, analytical or MCMC population size integration, coordinated or naïve topology changing operators, and the inclusion or exclusion of coordinated height changing operators. Each chain used the same sequence data but was an independent estimate of convergence because a different random seed was used to initialise each chain.

³<http://dx.doi.org/10.5061/dryad.23rc0> — accessed 15th December 2017

A birth-death prior was used for the species tree and both the net diversification and extinction fraction hyperparameters were estimated. A gamma prior was used for MCMC estimated population sizes with a shape fixed at 2 and an estimated mean population size hyperparameter, matching the original *BEAST model (Heled and Drummond, 2010). The number of branch rate categories was equal to the number of estimated branch rates (as is the default in BEAST 2), and the standard deviation of the UCLN clock model was fixed at 0.3.

To ensure convergence of all chains, we ran each chain for an initial length of $2^{24} = 16,777,216$ states, sampling every $2^{11} = 2,048$ states. Initial chain lengths and sampling rates for all other analyses are in Table S8. ESS values were computed for all recorded statistics after discarding 12.5% of state samples as burn-in. Recorded statistics included (1) the posterior probability, (2) the coalescent probabilities of gene trees, (3) the overall prior probability, (4) the birth death prior probability of the species tree, (5) the phylogenetic likelihood, (6) the net diversification rate, (7) the extinction fraction, (8) the mean population size, and (9) the height of the species tree.

If any recorded statistic had an ESS below 200, the chain was resumed until the length of the chain had doubled. ESS values were then re-evaluated, again after discarding 12.5% of state samples. The length of a chain was continually doubled and ESS values re-evaluated until the ESS values of all recorded statistics were above 200. The rate at which trees and statistics were sampled was halved with every chain doubling so that the total number of samples remained constant. Two *BEAST GT-UCLN chains still had insufficient ESS values after running for $2^{34} = 17,179,869,184$ states, but all other chains had converged. Estimated ESS values for all chains were used for analyses of computational performance.

ESS per hour was calculated by dividing the final ESS value for a given statistic by 87.5% of the total CPU time used by that chain to account for burn-in. Likewise ESS per million states was calculated by dividing the final ESS value by 87.5% of the total number of the states

in the chain, then multiplied by one million. For all analyses of computational performance including graphs and linear models, the ESS rate for any given chain was that of the slowest converging statistic for that particular chain.

Average branch length in coalescent units was calculated by concatenating the output (after discarding the first 12.5% of states as burn-in from each chain) of all 30 chains which used the combination of MCMC population size integration, naïve topology operators, coordinated node height operators and species tree branch rates. For every sample in the combined posterior distribution, the coalescent length of each branch $\tau(2N_e)^{-1}$ was calculated from its length in substitution units τ and its effective population size N_e . The mean coalescent length of all branches across all samples was taken as the average.

2.5.3 TESTING THE EFFECTS OF SPILS ON ESTIMATED SUBSTITUTION RATES

To test how SPILS affected estimates of per-species branch substitution rates, 96 fully asymmetric species trees were simulated with the topology (((((A,B),C),D),E). All species trees were simulated according to a pure birth Yule process (Yule, 1924) with a speciation rate of 10 per substitution.

Haploid population sizes for each branch were chosen independently from an inverse gamma distribution with a shape of 3 and a scale of 0.2. 100 gene trees with one individual per extant species were then simulated for each species tree according to the MSC process using biopy. Finally 1000nt sequence alignments were then simulated for each gene tree according to the Jukes-Cantor substitution model (Jukes and Cantor, 1969), no among-site rate variation, a strict molecular clock, and a substitution rate of 1 for each locus. Sequence alignments were simulated using Seq-Gen (Rambaut and Grassly, 1997).

BEAST concatenation and StarBEAST2 were then used to estimate the branch rates and divergence times with the species tree topology fixed to the truth. The same substitution model used for simulating sequences (i.e. Jukes-Cantor, no rate variation among sites or loci) was

also used for inference. UCLN relaxed clocks were applied to the tree inferred by concatenation and to the StarBEAST2 species tree.

The same strategy as applied to *Pseudacris* was used to ensure convergence of StarBEAST2, but for concatenation mean population sizes and coalescent probabilities are not part of the model and so were not recorded.

For every converged chain, the posterior expectation and 95% credibility intervals of per-species branch rates were calculated using the TreeAnnotator program supplied with BEAST.

2.5.4 SIMULATIONS TO MEASURE COMPUTATIONAL AND STATISTICAL PERFORMANCE

All simulation parameters were chosen to be broadly similar to those observed in or estimated from the *Pseudacris* data set.

First, 30 species trees were simulated according to a birth-death process (Gernhard, 2008) using biopy with 21 extant species, a speciation rate of 100 and a death rate of 30. This corresponds to a net diversification rate of 70 and an extinction fraction of 0.3. Haploid population sizes for each branch were chosen independently from a gamma distribution with a shape of 2 and a scale of 0.002. For a species with annual generation times, as is the case for at least some *Pseudacris* species (Caldwell, 1987), and a substitution rate of 10^{-9} per year this corresponds to an effective population size N_e of around 2 million individuals per generation. Species branch rates were chosen from a log-normal distribution with a mean in real space of 1 and a standard deviation of 0.3, then scaled so that the mean of the branch rates for a given species tree was exactly 1. This ensured that per-branch rates always reflected relative differences in substitution rates.

For each species tree, 130 gene trees with two sampled haplotype sequences per species were simulated according to the MSC process using biopy. The mean clock rate for each locus was chosen from a log-normal distribution with a mean in real space of 1 and a standard deviation of 0.6.

For each gene tree, 600nt long sequence alignments were simulated using Seq-Gen (Rambaut and Grassly, 1997). An HKY model was used for all sequence alignments with equal base frequencies, a κ value of 3, and a four rate category discretised gamma model of among-site rate variation with a shape α value of 0.2. Hence all inference based on simulated data applied the HKY+ Γ substitution model to all loci.

The same combinations of clock models, population size integration and new operators were explored using the simulated data as for *Pseudacris* to provide more generally applicable results regarding those new techniques. The same number of loci, convergence strategy and calculations of ESS rates were used for both. Both haplotype sequences were used for each species for *BEAST and StarBEAST2. One concatenation chain using phased haplotypes and a relaxed clock still had insufficient ESS values after running for $2^{33} = 8,589,934,592$ states, but all other chains had converged. Estimated ESS values for all chains were used for analyses of computational performance and statistical coverage and accuracy.

2.5.5 COMPARISON OF STARBEAST2 WITH *BEAST AND CONCATENATION

To compare the performance of StarBEAST2 with *BEAST, we ran 30 strict clock and 30 gene tree relaxed clock replicates of the *Pseudacris* reanalysis using the *BEAST package built into BEAST 2. We also reran each simulation replicate using *BEAST with a strict clock and gene tree relaxed clocks. The same priors, substitution models, and convergence strategies as used for StarBEAST2 were used with *BEAST.

For both data sets we reused the StarBEAST2 results for the combination of analytical population size integration, coordinated height-changing operators and naïve topology operators, which are all enabled by default in StarBEAST2. To demonstrate the scaling of StarBEAST2, we also reran each simulation replicate with an additional 26 loci (for a total of 52 loci) for all three clock models.

To compare concatenation with StarBEAST2, we reran each simulation replicate for each

combination of either unphased ambiguity coded sequences or a single haplotype sequence per species, either a strict clock or species tree relaxed clock, and either the original 26 loci or with an additional 104 loci (for a total of 130 loci). We estimated the per-locus rates in the same way as for StarBEAST2, and applied the same convergence strategy as for SPILS concatenation. For species tree clock rates we used the same UCLN parameters as StarBEAST2 but applied to the concatenated tree, a model equivalent to that described by Rasmussen and Kellis (2007).

We also generated 30 replicates from a UCE data set of *Crocidura* shrews to show that StarBEAST2 can scale to 100 loci. For 1020 out of 1112 loci, the best fitting substitution model was either HKY or a nested model (Giarla and Esselstyn, 2015). To simplify configuring substitution models, we chose 100 unique loci at random, and separately for each replicate, from the set of loci which best fit HKY or a nested model. For each replicate we ran *BEAST with a strict clock or gene tree relaxed clock and a subset of 50 loci, StarBEAST2 with all three clock models and the same subset, and StarBEAST2 with all three clock models and all 100 loci. The same priors, substitution model and convergence strategies were used as for the simulated data set. All *Crocidura* MCMC chains converged.

2.5.6 MEASUREMENTS OF SPECIES TREE COVERAGE AND ACCURACY

Branch length error is defined as $\sum_b |\hat{l}_b - l_b| / \sum_b l_b$ where l_b is the true simulated branch length, and \hat{l}_b is the point estimate of the branch length, for a given species tree branch b in a set of branch lengths B . Total branch length bias is defined as $\sum_b \hat{l}_b - \sum_b l_b / \sum_b l_b$. In this study, B is either the set of tip branches, or the set of internal branches excluding the root branch. Point estimates of branch lengths were calculated using the common ancestor method conditioned on the true simulated topology (Heled and Bouckaert, 2013). Highest posterior density regions were used for all credible intervals.

In this study, the rooted Robinson-Foulds distance (Robinson and Foulds, 1981) is the

number of clades present in only one of the true tree T_1 or the maximum clade credibility tree T_2 . The 95% credible set contains tree topologies selected from topologies present in the posterior sample, in order from high to low posterior probability, until the cumulative probability reached or exceeded 95%.

2.6 SUPPLEMENTARY MATERIAL

Supplementary material is available at Molecular Biology and Evolution online. ⁴

2.7 ACKNOWLEDGMENTS

This work was supported by a Rutherford Discovery Fellowship awarded to A.J.D. by the Royal Society of New Zealand. H.A.O. was supported by an Australian Laureate Fellowship awarded to Craig Moritz by the Australian Research Council (FL110100104). This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We wish to thank Jason Bragg and Renee Catullo for testing StarBEAST2 before its official release, Timothy Vaughan for suggesting the addition of a root height changing operator, Joseph Heled for insight into the multispecies coalescent, and Graham Jones for input regarding operator performance. We also thank Tanja Stadler for hosting H.A.O. and A.J.D. during part of the development of StarBEAST2, and thank Fábio Mendes, Matthew Hahn, Diego Mallo, two anonymous reviewers and the editor for suggesting valuable improvements to this manuscript.

⁴<https://doi.org/10.1093/molbev/msx126> — accessed 15th December 2017

3

Bayesian Inference of Species Networks from Multilocus Sequence Data

ABSTRACT

Reticulate species evolution, such as hybridization or introgression, is relatively common in nature. In the presence of reticulation, species relationships can be captured by a rooted phylogenetic network, and orthologous gene evolution can be modeled as bifurcating gene trees embedded in the species network. We present a Bayesian approach to jointly infer species networks and gene trees from multilocus sequence data. A novel birth-hybridization process is used as the prior for the species network, and we assume a multispecies network coalescent (MSNC) prior for the embedded gene trees. We verify the ability of our method to correctly sample from the posterior distribution, and thus to infer a species network, through simula-

tions. To quantify the power of our method, we reanalyze two large datasets of genes from spruces and yeasts. For the three closely related spruces, we verify the previously suggested homoploid hybridization event in this clade; for the yeast data, we find extensive hybridization events. Our method is available within the BEAST 2 add-on `SpeciesNetwork`, and thus provides an extensible framework for Bayesian inference of reticulate evolution.

3.1 INTRODUCTION

Hybridization during speciation is relatively common in animals and plants (Mallet, 2005, 2007). However, when reconstructing the evolutionary history of species, typically non-reticulating species trees are inferred (Guindon *et al.*, 2010; Stamatakis, 2014; Drummond and Bouckaert, 2015; Ronquist *et al.*, 2012a), and the potential for hybridization events is ignored.

To account for the distribution of evolutionary histories of genes inherited from multiple ancestral species, the multispecies coalescent model (Rannala and Yang, 2003; Liu *et al.*, 2009a) was extended to allow reticulations among species, named multispecies network coalescent (MSNC) model (Yu *et al.*, 2014). Orthologous genes are modeled as gene trees embedded in the species network. The MSNC model accounts for gene tree discordance due to incomplete lineage sorting and reticulate species evolution events, such as hybridization or introgression. There have been computational methods developed based on the MSNC to infer species networks using maximum likelihood (Yu *et al.*, 2014; Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016) and Bayesian inference (Wen *et al.*, 2016). These methods use gene trees inferred from other resources as input. Due to the model complexity, applying the MSNC model in a full Bayesian framework, i.e., to infer the posterior distribution of species network and gene trees directly from the multilocus sequence data, is challenging. Recently Wen and Nakhleh (2017) have developed a Bayesian method that can co-estimate species networks and gene trees from multilocus sequence data, but a process-based prior for the species network is still lack-

ing. Their method also integrates over all possible gene tree embeddings at each MCMC step, which means that the estimated histories of individual gene trees within the species network are not available for subsequent analysis, and the method does not co-estimate base frequencies or substitution (transition and transversion) rates.

In this paper, we present a Bayesian method to infer ultrametric species networks jointly with gene trees and their embeddings from multilocus sequence data. Our method assumes a birth-hybridization model for the species network, the MSNC model for the embedded gene trees with analytical integration of population sizes, and employs novel MCMC operators to sample the species network and gene trees along with associated parameters. It is able to use the full range of substitution models implemented in BEAST 2 (Bouckaert *et al.*, 2014), including models with gamma rate variation across sites (Yang, 1994).

3.2 NEW APPROACHES

In this section, we specify our approach to sample from the posterior distribution of species networks and gene trees, given a multilocus sequence alignment. First we derive the unnormalized posterior distribution. Then we introduce operators to move through the space of species networks, the space of gene trees, and finally to update the gene tree embeddings within species networks.

3.2.1 THE POSTERIOR DISTRIBUTION OF SPECIES NETWORKS AND GENE TREES

THE PROBABILITY DENSITY OF A SPECIES NETWORK

The birth-hybridization process provides a prior probability for a given species network Ψ (Figure 3.1). The process starts from t_0 (time of origin) in the past with a single species. A species gives birth to a new species with a constant rate λ (speciation rate), and two species merge into one with a constant rate ν (hybridization rate). That is, at the moment of k species,

the speciation rate is $k\lambda$, the hybridization rate is $\binom{k}{2}\nu$, and the waiting time to the next event is an exponential distribution. The process ends at time 0 (the present).

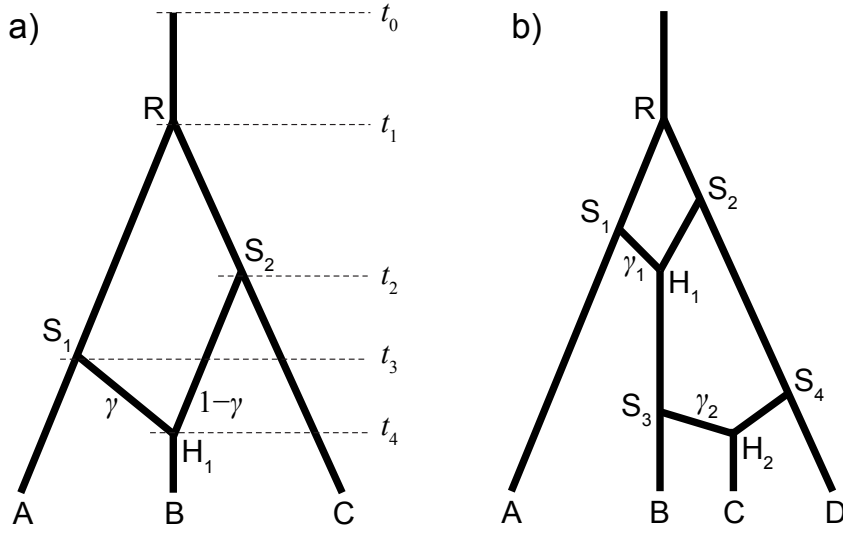


Figure 3.1: a) A species network with 3 tips, 3 bifurcations, and 1 reticulation. The inheritance probability at branch S_1H_1 is γ , and that at S_2H_1 is $1 - \gamma$. b) Another network with 4 tips and 2 reticulations, with γ_1 and γ_2 associated with S_1H_1 and S_3H_2 , respectively.

The probability density of a species network Ψ with n extant species descending from $n - 1 + m$ speciation events and m hybridization events, and these events happening at time $t_1 > t_2 > \dots > t_{n+2m-1}$, conditioned on t_0 , λ and ν , is,

$$f(\Psi | \lambda, \nu, t_0) = \lambda^{n+m-1} \nu^m \prod_{i=0}^{n+2m-1} e^{-(\lambda k_i + \nu \binom{k_i}{2})(t_i - t_{i+1})}, \quad (3.1)$$

where k_i is the number of lineages within time interval (t_i, t_{i+1}) and $t_{n+2m} = 0$ is the present

time. For the network shown in Figure 3.1a, the probability density of the species network is

$$\begin{aligned}
 f(\Psi \mid \lambda, \nu, t_0) &= \lambda e^{-\lambda(t_0-t_1)} \\
 &\quad \lambda e^{-(2\lambda+\nu)(t_1-t_2)} \\
 &\quad \lambda e^{-(3\lambda+3\nu)(t_2-t_3)} \\
 &\quad \nu e^{-(4\lambda+6\nu)(t_3-t_4)} \\
 &\quad e^{-(3\lambda+3\nu)t_4}.
 \end{aligned}$$

In our Bayesian analysis, the parameters λ , ν , and t_0 can be assigned hyperpriors.

Hybridizations or gene flow are modeled by reticulations in the species network. $\gamma = \{\gamma_1, \dots, \gamma_H\}$ are the inheritance probabilities, one per reticulation node in Ψ (Figure 3.1). The inheritance probability measures the average proportion of genetic material inherited from the corresponding parent (or donor) (Long, 1991; Yu *et al.*, 2014; Wen and Nakhleh, 2017). While the prior for γ can be any distribution on $[0, 1]$, in this study we use $f(\gamma_h) \sim U(0, 1)$ throughout.

THE PROBABILITY OF THE SEQUENCE DATA GIVEN THE GENE TREES

Assuming complete linkage within each locus, the probability of the data

$D = \{D_1, D_2, \dots, D_L\}$ given gene trees $G = \{G_1, G_2, \dots, G_L\}$ is the product of phylogenetic likelihoods (Felsenstein, 1981) at individual loci:

$$\Pr(D \mid G, \boldsymbol{\mu}, \boldsymbol{\varphi}) = \prod_{i=1}^L \Pr(D_i \mid G_i, \mu_i, \varphi_i), \quad (3.2)$$

where G_i is the gene tree with coalescent times, μ_i is the substitution rate per site per time unit, and φ_i represents the parameters in the substitution model (e.g., the transition-transversion rate ratio κ in the HKY85 model (Hasegawa *et al.*, 1985)), at locus

i ($i = 1, \dots, \mathbb{L}$).

There are two sources of evolutionary rate variation: across gene tree lineages at the same locus and across different gene loci. In the strict molecular clock model (Zuckermandl and Pauling, 1965), μ is the global clock rate, i.e., no rate variation across gene lineages at each locus. To extend to a relaxed molecular clock model (e.g., Thorne and Kishino, 2002; Drummond *et al.*, 2006; Lepage *et al.*, 2007; Rannala and Yang, 2007), the molecular clock rate is variable across gene lineages following certain distributions with μ as the mean. To account for rate variation across genes, gene-rate multipliers $\{m_1, m_2, \dots, m_{\mathbb{L}}\}$ are constrained to average to 1.0 ($\sum_{i=1}^{\mathbb{L}} m_i x_i = 1$, where x_i is the proportion of sites in locus i to the total number of sites). Then the substitution rate at locus i is $\mu_i = \mu m_i$. Thus, when multiplying the gene tree lineages in G_i by μ_i , all the branch lengths are then measured in units of expected substitutions per site.

The gene-rate multipliers are assigned a flat Dirichlet prior. The average substitution rate (clock rate) μ can be either fixed to 1.0 such that branch lengths are measured by expected substitutions per site, or assigned an informative prior to infer branch lengths measured in absolute time.

THE PROBABILITY DENSITY OF THE GENE TREES GIVEN A SPECIES NETWORK

The gene trees $G = \{G_1, G_2, \dots, G_{\mathbb{L}}\}$ are embedded in the species network Ψ under the multispecies network coalescent (MSNC) model (Yu *et al.*, 2014) (Figure 3.2). The effective population sizes $N = \{N_1, N_2, \dots, N_{\mathbb{B}}\}$ are assumed to be identically and independently distributed (i.i.d.) for each of the \mathbb{B} branches in Ψ , while each locus has the same effective population size N_i at branch i ($i = 1, \dots, \mathbb{B}$). For each locus j , the number of coalescences of gene tree G_j within branch b of Ψ is denoted by k_{jb} , and the number of lineages at the tipward end of b is denoted by n_{jb} , thus the number of lineages at the rootward end of b is $n_{jb} - k_{jb}$. The $k_{jb} + 1$ coalescent time intervals between the tipward and rootward of branch b are denoted by

c_{jbi} ($0 \leq i \leq k_{jb}$). p_j is the gene ploidy of locus j (e.g., 2 for autosomal nuclear genes and 0.5 for mitochondrial genes in diploid species). For each lineage of G_j traversing the reticulation node H_h backward in time, with probability γ_h it goes to the parent branch associated with that inheritance probability, and to the alternate parent branch with probability $1 - \gamma_h$. The corresponding number of traversing lineages are denoted by u_{jh} and v_{jh} respectively.

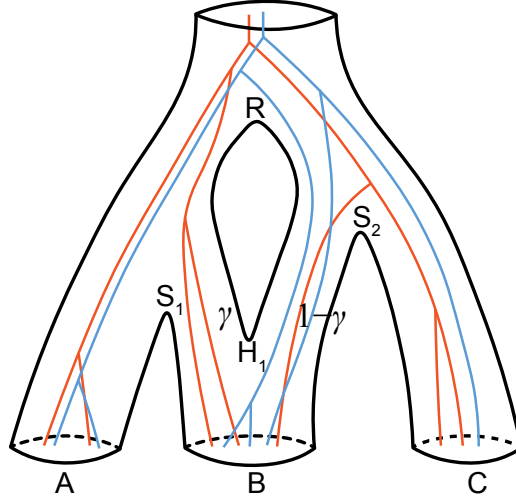


Figure 3.2: Two gene trees embedded in the species network of Figure 3.1a. There are 2 samples from species A , 3 samples from B , and either 1 or 2 samples from C . For each gene tree lineage traversing the reticulation node H_1 backward in time, it goes to the left population with probability γ , and to the right with probability $1 - \gamma$.

The coalescent probability of the gene trees G in species network Ψ with time being measured in calendar units is thus:

$$\begin{aligned}
 f(G \mid \Psi, \boldsymbol{\gamma}, N) &= \prod_{j=1}^L \left[\prod_{b=1}^B (p_j N_b)^{-k_{jb}} \exp \left(-(p_j N_b)^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2} \right) \prod_{h=1}^H \gamma_h^{u_{jh}} (1 - \gamma_h)^{v_{jh}} \right] \\
 &= \Lambda \prod_{b=1}^B r_b N_b^{-q_b} \exp(-\sigma_b N_b^{-1}), \tag{3.3}
 \end{aligned}$$

where $q_b = \sum_j k_{jb}$, $r_b = \prod_j p_j^{-k_{jb}}$, $\sigma_b = \sum_j p_j^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2}$, and

$\Lambda = \prod_j \prod_h \gamma_h^{u_{jh}} (1 - \gamma_h)^{v_{jh}}$. When there is no reticulation in the species network (i.e., it is a species tree), then $\Lambda = 1$ and Equation 3.3 is equivalent to Equation 2 in Jones (2017).

Note here, when time is measured by expected substitutions per site, we use $\theta_b = N_b\mu$ as the population size parameter of branch b , and $\tau_i = t_i\mu$ as the height of node i . In the next section, we discuss how to integrate out the population sizes, which will improve computational speed.

INTEGRATING OUT THE POPULATION SIZES ANALYTICALLY

Equation 3.3 has the form of unnormalized inverse gamma densities. The population sizes N can be integrated out through the use of i.i.d. inverse-gamma $IG(\alpha, \beta)$ conjugate prior distributions (Jones, 2017; Hey and Nielsen, 2007), that is,

$$\begin{aligned}
f(G | \Psi, \gamma) &= \int f(G | \Psi, \gamma, N) f(N | \alpha, \beta) dN \\
&= \Gamma \prod_{b=1}^B \int_0^{\infty} r_b N_b^{-q_b} \exp(-\sigma_b N_b^{-1}) \frac{\beta^\alpha}{\Gamma(\alpha)} N_b^{-\alpha-1} \exp(-\beta N_b^{-1}) dN_b \\
&= \Gamma \prod_{b=1}^B \frac{r_b \beta^\alpha}{(\beta + \sigma_b)^{\alpha+q_b}} \frac{\Gamma(\alpha + q_b)}{\Gamma(\alpha)}. \tag{3.4}
\end{aligned}$$

The symbolic notations follow Equation 3.3.

THE JOINT POSTERIOR DISTRIBUTION

The joint posterior distribution of the parameters is

$$\begin{aligned}
f(\Psi, G, \Theta | D) &\propto \Pr(D | G, \boldsymbol{\mu}, \boldsymbol{\varphi}) f(G | \Psi, \boldsymbol{\gamma}) f(\Psi | \lambda, \nu, t_0) \\
&f(\boldsymbol{\mu}) f(\boldsymbol{\varphi}) f(\boldsymbol{\gamma}) f(\lambda, \nu) f(t_0). \tag{3.5}
\end{aligned}$$

Here Θ represents $(\boldsymbol{\mu}, \boldsymbol{\varphi}, \boldsymbol{\gamma}, \lambda, \nu, t_0)$.

3.2.2 MCMC OPERATORS FOR THE SPECIES NETWORK

NODE SLIDER

The node-slider operator only changes the node heights of the species network, not the topology. It selects an internal node or the origin randomly, then proposes a new height centered at the current height according to a normal distribution: $t' | t \sim N(t, \sigma^2)$, where σ is a tuning parameter controlling the step size. The lower bound is the oldest child-node height, the upper bound is the youngest parent-node height (except for the origin, Figure 3.3). If the proposed value is outside this range, the excess is reflected back into the interval. Note that for the origin, if the proposed height is outside the range of its prior, this move is aborted. A variation of this operator can use a uniform proposal instead of the normal proposal: $t' | t \sim U(t - w/2, t + w/2)$, where w is the window size. The proposal ratio is 1.0 in both cases.

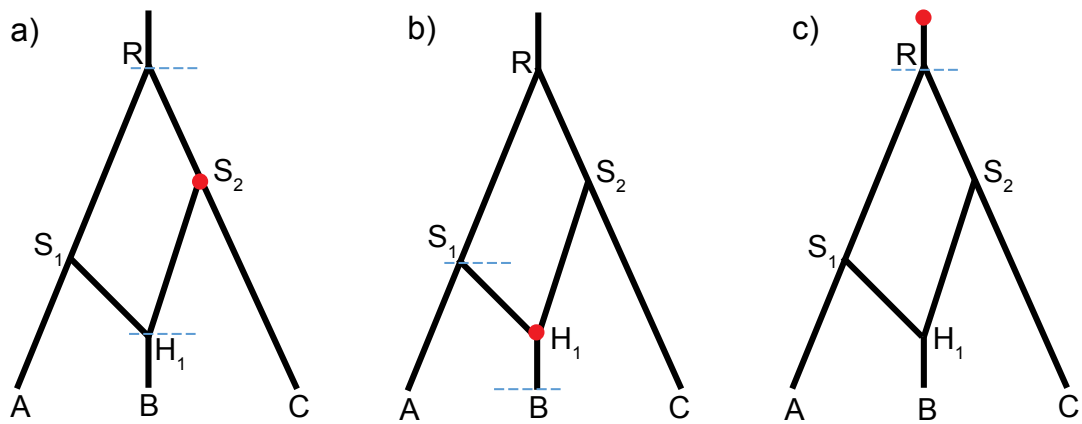


Figure 3.3: Three cases when the node-slider operator is applied: a) a bifurcation node S_2 is selected; b) the reticulation node H_1 is selected; c) the origin is selected. The dashed lines are the lower and upper bounds for changing its height (only the lower bound is applicable in c)). For the node-uniform operator, a) and b) apply but c) does not.

NODE UNIFORM

The node-uniform operator also changes the internal-node heights of the species network while keeping the topology. It selects an internal node randomly, then proposes a new height uniformly between the lower and upper bounds (Figure 3.3ab). The lower bound is the old-

est child-node height, the upper bound is the youngest parent-node height. The proposal ratio is 1.0. Unlike node slider, this operator does not change the time of origin. A separate operator for the origin, such as multiplier or scaler, can be coupled to update all the node heights.

RELOCATE BRANCH

The relocate-branch operator can change the topology, but keeps the number of reticulations in the species network constant. It first selects an internal node at random. If the selected node is a bifurcation node, the rootward end of either its child branches is selected (Figure 3.4a); if the selected node is a reticulation node, the tipward end of either its parent branches is selected (Figure 3.4b). Then the selected branch is detached at the side of the selected node, and a destination branch to be attached is chosen randomly from all possible candidate branches (including the original position). A new height of the selected node is proposed uniformly between the heights of the two ends of the destination branch (v' and u' in Figure 3.4). When the relocated branch has a bifurcation node at one end and a reticulation node at the other end, the candidate branches include all the remaining branches, and the reticulation direction can be changed depending on the proposed new height (Figure 3.4b). When the relocated branch has the same type of nodes at both ends and the resulted network is invalid, the move is aborted. For example, moving the rootward end lower than the tipward end if the two ends are both bifurcation nodes, or moving the tipward end higher than the rootward end if the two ends are both reticulation nodes, will result in an invalid network. We denote with v and u the lower and upper bounds of the backward move. The proposal ratio is $(u' - v')/(u - v)$.

ADD- AND DELETE-RETICULATION

The add-reticulation and delete-reticulation operators are reversible-jump MCMC (rjMCMC) proposals that can add and delete a reticulation event respectively.

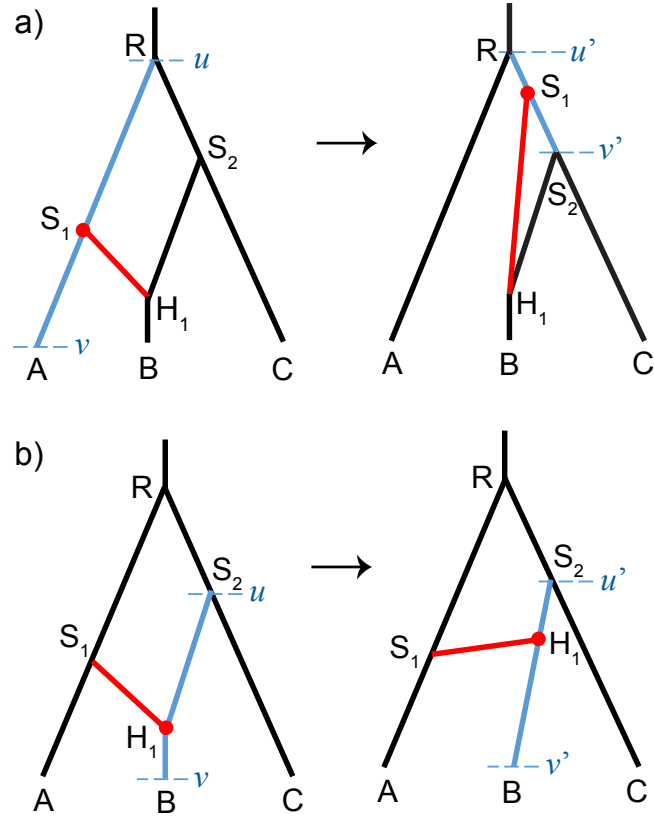


Figure 3.4: Two cases when the relocate-branch operator is applied. a) A bifurcation node S_1 is selected, and branch S_1H_1 is relocated to attach to RS_2 . b) A reticulation node H_1 is selected, and branch S_1H_1 is still attaching to S_2B with flipped reticulation direction. The lower and upper bounds of proposing the new attaching point are v' and u' , and the corresponding bounds of the backward move are v and u .

In the add-reticulation operator, a new branch is added by connecting two randomly selected branches with length l_1 and l_2 (Figure 3.5). The same branch can be selected twice so that $l_1 = l_2$ (Figure 3.5b). Then three values ω_1, ω_2 and ω_3 are drawn from $U(0, 1)$. One attaching point cuts the branch length l_1 to $l_{11} = l_1\omega_1$ (and thus $l_{12} = l_1(1 - \omega_1)$); the other attaching point cuts the branch length l_2 to $l_{21} = l_2\omega_2$ (and thus $l_{22} = l_2(1 - \omega_2)$). Analogously, if we select the same branch twice, the attachment times of the new branch are $l_1\omega_1$ and $l_1\omega_2$. An inheritance probability $\gamma = \omega_3$ is associated to the new branch. We will operate on the inheritance probability γ of this added branch, while the inheritance probability of the second reticulation branch (i.e., $1 - \gamma$) changes accordingly. We denote k as the number of branches in the current network, and m as the number of reticulation branches (parent branches of the reticulation nodes) in the proposed network. The Hastings ratio is then $(1/m)/[(1/k)(1/k) \times 1 \times 1 \times 1] = k^2/m$. The Jacobian is $|\frac{\partial(l_{11}, l_{21}, \gamma)}{\partial(\omega_1, \omega_2, \omega_3)}| = l_1l_2$. Thus

the proposal ratio of add-reticulation is $l_1 l_2 k^2 / m$.

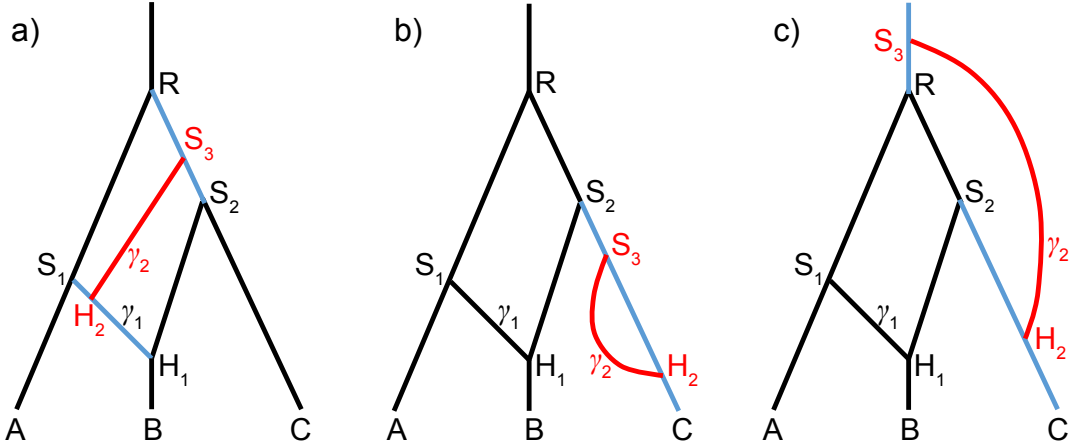


Figure 3.5: Three cases when the add-reticulation operator is applied. The number of branches in the current network (i.e., the network without the red branch) is $k = 8$. The probability of selecting the illustrated branches (in blue) is $1/k^2$. The number of reticulation branches in the proposed network is $m = 4$. In the reverse move, delete-reticulation, the probability of selecting the added branch (in red) is $1/m$. a) Branches S_1H_1 and RS_2 are selected and a new branch S_3H_2 is added together with γ_2 . The length of S_1H_1 is $l_1 = l_{S_1H_1}$, and that of RS_2 is $l_2 = l_{RS_2}$. In the delete-reticulation move, if H_1H_2 is selected, the operator is aborted. b) The same branch S_2C is selected twice. $l_1 = l_2 = l_{S_2C}$, $l_{11} = l_{S_2S_3}$, $l_{21} = l_{S_2H_2}$. c) The root branch and S_2C are selected. S_3 becomes the new root.

In the delete-reticulation operator, a random reticulation branch together with the inheritance probability γ is deleted (Figure 3.5). Joining the singleton branches at each end of the deleted branch, resulting in two branches with length l_1 and l_2 completes the operator ($l_1 = l_2$ when forming a single branch, Figure 3.5b). If there is no reticulation, or the selected branch is connecting two reticulation nodes, the move is aborted. For example in Figure 3.5a, deleting reticulation branch H_1H_2 will result in an invalid network. We denote k as the number of branches in the proposed network, and m as the number of reticulation branches in the current network. The proposal ratio of delete-reticulation is $m/(k^2 l_1 l_2)$.

INHERITANCE-PROBABILITY UNIFORM

The inheritance-probability uniform operator selects a reticulation node randomly, and proposes a new value of the inheritance probability $\gamma' \sim U(0, 1)$. The proposal ratio is 1.0.

INHERITANCE-PROBABILITY RANDOM-WALK

The inheritance-probability random-walk operator selects a reticulation node randomly, and applies a uniform sliding window to the logit of the inheritance probability γ , that is $y' \mid y \sim U(y - w/2, y + w/2)$, where $y = \text{logit}(\gamma) = \log(\gamma) - \log(1 - \gamma)$. Since the proposal ratio for the transformed variable y is 1.0, and $\frac{d\gamma}{dy} = \frac{d}{dy} [e^y / (1 + e^y)] = e^y / (1 + e^y)^2$, the proposal ratio for the original variable γ is $\frac{d\gamma'}{dy'} / \frac{d\gamma}{dy} = e^{(y'-y)} (1 + e^y)^2 / (1 + e^{y'})^2$.

3.2.3 MCMC OPERATORS FOR GENE TREES

The standard tree operators in BEAST 2 (Bouckaert *et al.*, 2014) are applied to update the gene trees, including the scale, uniform, subtree-slide, narrow- and wide-exchange, and Wilson-Balding (Wilson and Balding, 1998). The scale and uniform operators only update the node heights without changing the tree topology, while the other operators can change the topology (Drummond and Bouckaert, 2015). The species network is kept unchanged when operating on the gene trees, and vice versa.

3.2.4 MCMC OPERATOR FOR THE GENE TREE EMBEDDING

The gene trees must be compatibly embedded in the species network (Figure 3.2). When a new gene tree is proposed using one of the tree operators, the rebuild-embedding operator proposes a new embedding for that gene tree. When a new species network is proposed, the rebuild-embedding operator proposes a new embedding for each gene tree in the species network. If there is no valid embedding for any gene tree, the gene tree or species network proposal is rejected.

The rebuild-embedding operator proposes a new embedding proportional to the product of traversal probabilities across all traversed reticulation nodes. Specifically, we define the (un-normalized) likelihood of a compatible embedding x as $w_x = \prod_{h=1}^H \gamma_h^{u_{xh}} (1 - \gamma_h)^{v_{xh}}$, where

\mathbb{H} is the number of reticulation nodes in the species network, u_{xh} is the number of lineages traversing node H_h to the branch associated with γ_h , and v_{xh} is the number of lineages traversing node H_h to the alternative branch associated with $1 - \gamma_h$. If there is no reticulation in the species network (i.e., it is a species tree), $w_x = 1$. For example in Figure 3.2, there are two possible embeddings for one gene tree (orange) while the likelihoods are $\gamma^2(1 - \gamma)$ (current) and $(1 - \gamma)^3$ respectively, and four possible embeddings for the other gene tree (blue) while the likelihoods are γ^2 , $\gamma(1 - \gamma)$, $(1 - \gamma)\gamma$, and $(1 - \gamma)^2$ (current), respectively.

The proposal ratio of moving from embedding x to x' is

$$\frac{w_x}{\sum_{i=1}^{\mathbb{E}} w_i} \bigg/ \frac{w_{x'}}{\sum_{j=1}^{\mathbb{E}'} w_j},$$

where \mathbb{E} and \mathbb{E}' are the number of possible embeddings in the current and new states respectively. If $\mathbb{E}' = 0$ (no valid embedding), the move is aborted. This proposal distribution is chosen to have a superior acceptance ratio than if a new embedding is proposed randomly from all possible embeddings.

3.2.5 SUMMARIZING POSTERIOR DISTRIBUTION OF SPECIES NETWORKS

Reducing many hundreds of posterior or bootstrap samples to a summary result is essential in order to describe the underlying distribution. For phylogenetic trees, many summary methods have been developed such as “majority rule consensus” and “maximum clade credibility” trees (Heled and Bouckaert, 2013). By comparison, methods to summarize samples of phylogenetic networks are underdeveloped. As part of the `SpeciesNetwork` package, we have implemented a basic method for summarizing networks, where unique network topologies are reported in descending order of their posterior probabilities. For each unique topology, each subnetwork is annotated with its posterior probability and node age credible interval.

To facilitate the calculation of posterior probabilities and credible intervals, we have de-

veloped an algorithm to enumerate each unique subnetwork, and label all occurrences of a unique subnetwork in a sample of phylogenetic networks. After running this algorithm, the label of a network’s root node uniquely identifies its topology, and the generation of a sorted summary of posterior topologies becomes trivial. Details of the algorithm are given in Appendix A. The default setting of our summary tool eliminates all parallel branches (e.g., S_3H_2 in Figure 3.5b) from all samples in the posterior before summarizing, which simplifies the posterior distribution of networks and reduces the number of unique topologies.

Alternatively, users may generate a summary network using the “major displayed tree” method as implemented in the PhyloNetworks package (Solís-Lemus *et al.*, 2017).

3.3 SIMULATIONS

The components from the last section, i.e., the unnormalized posterior density and the operators, allow us to implement a Markov chain Monte Carlo (MCMC) procedure to sample species networks and gene trees from the posterior distribution, given a multilocus sequence alignment. The implementation is available within BEAST 2 (Bouckaert *et al.*, 2014) as an add-on `SpeciesNetwork`. A convenient format for the species networks, and a link to our source code, is presented in Appendix A.

We investigate the performance of the implementation using simulations in this section. Time is measured by expected substitutions per site throughout the simulations, so that $\theta = N\mu$ is used for all population sizes and $\tau_i = t_i\mu$ for the time of node i . The substitution rate μ is fixed to 1.0 across all gene lineages (strict molecular clock) and all loci (no rate variation).

3.3.1 SIMULATION AND MCMC SAMPLING WITHOUT SEQUENCE DATA

To verify the implementation of our Bayesian MCMC method, we compared stochastic simulation to MCMC sampling of species networks and gene trees. We first generated networks under the birth-hybridization process. The simulator starts from the time of origin (t_0) with

one species. A species splits into two (speciation) with rate λ , and two species merge into one (hybridization) with rate ν . At the moment of k branches, the total rate of change is $r_{tot} = k\lambda + \binom{k}{2}\nu$. We generate a waiting time $\sim \exp(r_{tot})$ and a random variable $u \sim U(0, 1)$. If $u < k\lambda/r_{tot}$, we randomly select a branch to split; otherwise, we randomly select two branches to join, and generate an inheritance probability $\gamma \sim U(0, 1)$. The simulator stops at time 0 (cf. Figure 1). In this simulation, $\tau_0 = 0.1$, $\lambda = 20$, $\nu = 10$, and we kept 200,000 networks with exactly three tips. All the population sizes were fixed to $\theta = 0.01$. Given each simulated species network, we then simulated a gene tree with two samples from each species (2, 2, 2) under the backward-in-time MSNC, resulting in 200,000 gene trees.

In the MCMC, we used all the operators for the species network (with 3 tips), gene tree (with 2 samples per species), and embedding (see above). The parameters τ_0 , λ , ν and θ were fixed to the truth. The likelihood of data was set to be constant (no data). The chain was run 500 million steps and sampled every 2000 steps. The last 200,000 sampled species networks and gene trees were kept (i.e., the burn-in was 20%).

Theoretically, we expect the distributions of species network and gene trees to be identical from both simulation and MCMC sampling. Indeed, the networks obtained from the simulator and MCMC match when comparing the network length, root height, number of hybridizations, and time of the youngest hybridization (Figure 3.6). The tree sets from MSNC and MCMC also give rise to the same distributions of tree length, the gamma-statistic (Pybus and Harvey, 2000), and Colless index (Blum *et al.*, 2006) as expected (Figure 3.7).

3.3.2 INFERENCE OF SPECIES NETWORKS FROM SEQUENCES

We simulated sequence alignments of multiple loci to reveal the ability of our method to recover the true species network from multilocus sequence data. The true network is shown in Figure 3.1a, with $\tau_1 = 0.05$, $\tau_2 = 0.03$, $\tau_3 = 0.02$, $\tau_4 = 0.01$, $\gamma = 0.3$, and population sizes $\theta = 0.01$. A random gene tree was generated for each locus under the MSNC. Then DNA

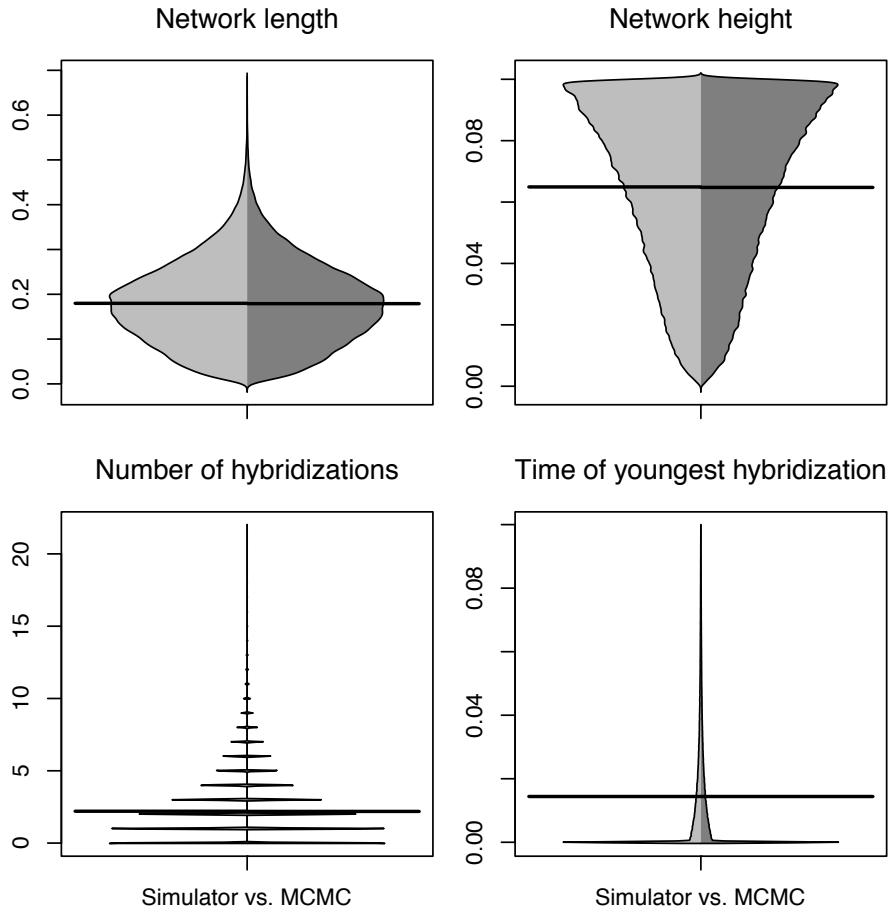


Figure 3.6: Beanplot of network summary statistics comparing 3-tips networks simulated under the birth-hybridization process (left, light gray) with those sampled using the MCMC operators (right, dark gray). The horizontal bar is the mean.

sequences of length 200 bp were simulated under JC69 model (Jukes and Cantor, 1969) along each tree. The sample configurations were $(2, 4, 2)$ (meaning species A has 2, B has 4, and C has 2 sampled sequences) and $(5, 10, 5)$, and the number of loci was 2, 5, 10, 20, 40, respectively. Under each setting, the simulation was repeated 100 times. In the inference, the priors were $\tau_0 \sim \exp(10)$ with mean 0.1, $d = \lambda - \nu \sim \exp(0.1)$ with mean 10, $r = \nu/\lambda \sim U(0, 1)$, and $\gamma \sim U(0, 1)$. The population sizes were integrated out analytically using inverse-gamma $N \sim IG(5, 0.05)$ (Eq. 3.4). The substitution model was set to JC69 (the true model). We fixed $\mu = 1.0$ for all genes as in the simulation (strict molecular clock and no rate variation). The MCMC chain was run for 50 million steps and sampled every 2000 steps. The first 35% of samples were discarded as burn-in. The results are shown in Figure 3.8.

With only 2 loci, the species trees are inferred with the highest posterior probability, the

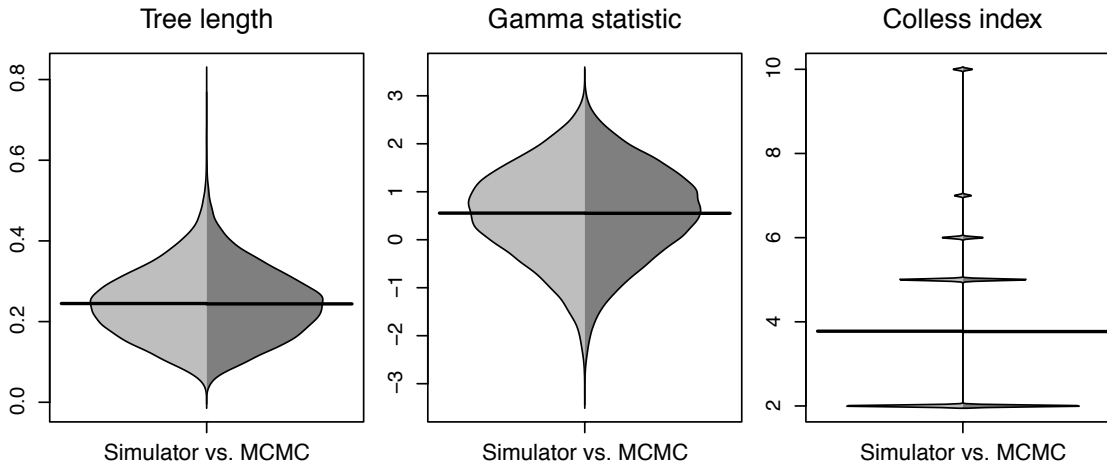


Figure 3.7: Beanplot of three tree summary statistics comparing gene trees simulated under MSNC (left, light gray) with those sampled using the MCMC operators (right, dark gray). The sample configuration was (2, 2, 2).

distribution of species network topologies is sensitive to the prior (Figure 3.8a). The HPD intervals of the network height are also very wide (Figure 3.8b). As sample size increases, the posterior estimates become increasingly accurate. Conditional on the true species network topology inferred (i.e., Figure 3.1a), the estimates of inheritance probability γ and time of hybridization become increasingly accurate as the number of loci increases (Figure 3.8cd). We also observe that adding loci increases the accuracy of inference more than adding individuals. For example, by comparing (5, 10, 5) individuals \times 5 loci with (2, 4, 2) individuals \times 10 loci, the latter has higher probability of recovering the true species network (Figure 3.8a).

To reveal the power of our method to detect both ancient and recent hybridization events, we simulated gene trees and sequences subsequently under the true species network shown in Figure 3.1b, with $\tau_R = 0.05$, $\tau_{H_1} = 0.03$, $\gamma_1 = 0.6$, $\tau_{H_2} = 0.01$, $\gamma_2 = 0.7$, $\tau_{S_1} = 0.035$, $\tau_{S_2} = 0.04$, $\tau_{S_3} = 0.012$, $\tau_{S_4} = 0.015$, and population sizes $\theta = 0.01$. The sample configurations were (2, 2, 2, 2) and (5, 5, 5, 5) respectively. The other settings were kept the same as in the previous simulation. The results are shown in Figure 3.9.

We find that an ancient hybridization close to the root is much harder to detect than a recent hybridization. With up to 8 samples and 20 loci, the posterior probabilities of the true network topology are all close to zero. The estimates start to increase with 20 samples and 20

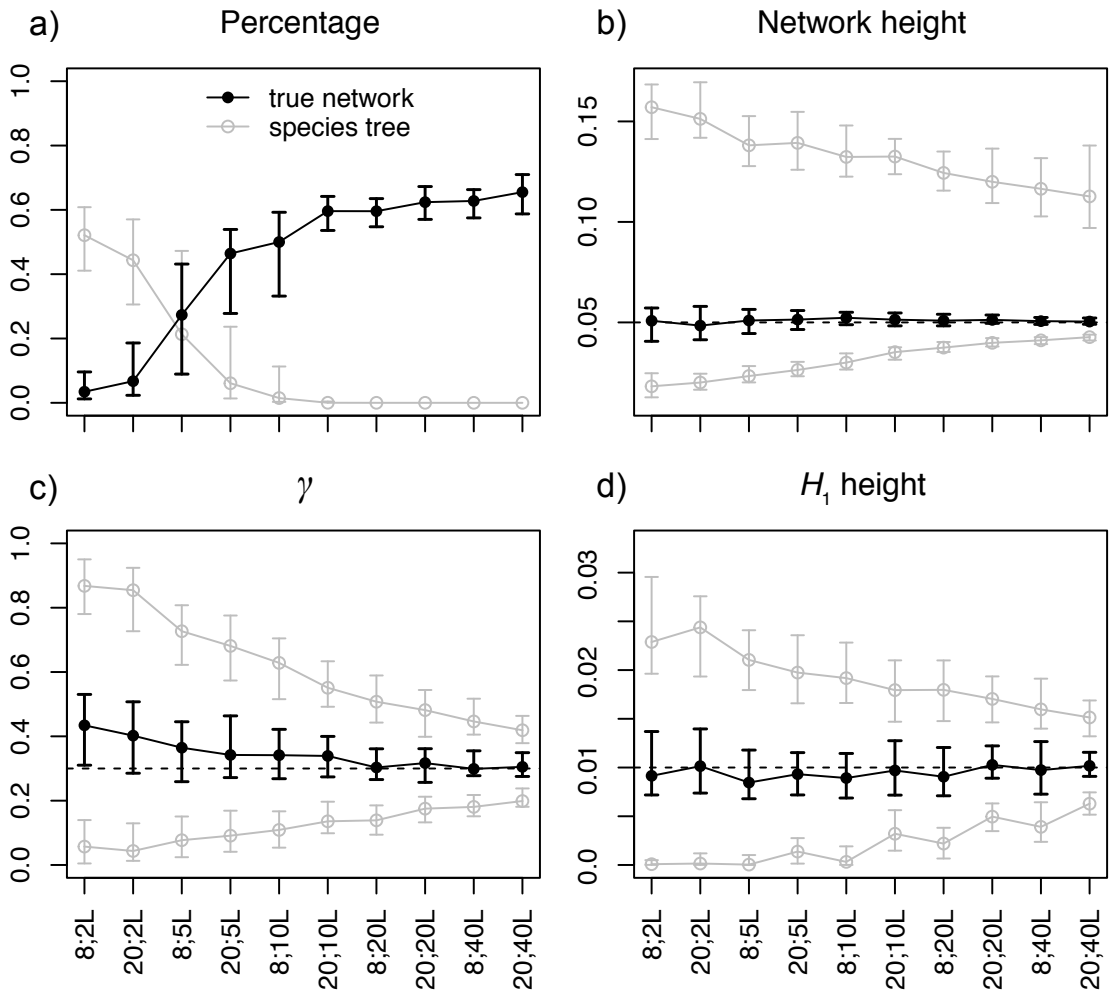


Figure 3.8: Posterior estimates of a) probability of the true network (black) and species trees (gray), b) network height, c) γ in the true network topology, and d) H_1 height in the true network topology, when the data were simulated under the network in Figure 3.1a with sample configuration (2, 4, 2) or (5, 10, 5), and 2, 5, 10, 20, or 40 loci, respectively. For each setting in a), the dot/circle with error bars are the median and the 1st and 3rd quartiles of the percentages of 100 replicates. For each setting in b), c) and d), the black dot with error bars are the median and the 1st and 3rd quartiles of the posterior medians of 100 replicates, the gray circles with error bars are the same summaries for the 95% HPD intervals. The dashed lines indicate the true values.

loci or more (Figure 3.9a). The difficulty is mainly due to the fact that there are few gene-tree lineages close to the root of the network, making it hard to distinguish the true hybridization event from incomplete lineage sorting in the ancestral populations. However, the recent hybridization event is inferred with high probability using 10 to 40 loci (Figure 3.9a). More specifically, we looked at the posterior probability of networks having the $BCDH_2$ subnetwork structure (cf. Figure 3.1b). Conditional on having this subnetwork inferred, the estimates of inheritance probability γ_2 become increasingly accurate as the number of loci increases (Figure 3.9c), although the time of hybridization H_2 is generally underestimated (Fig-

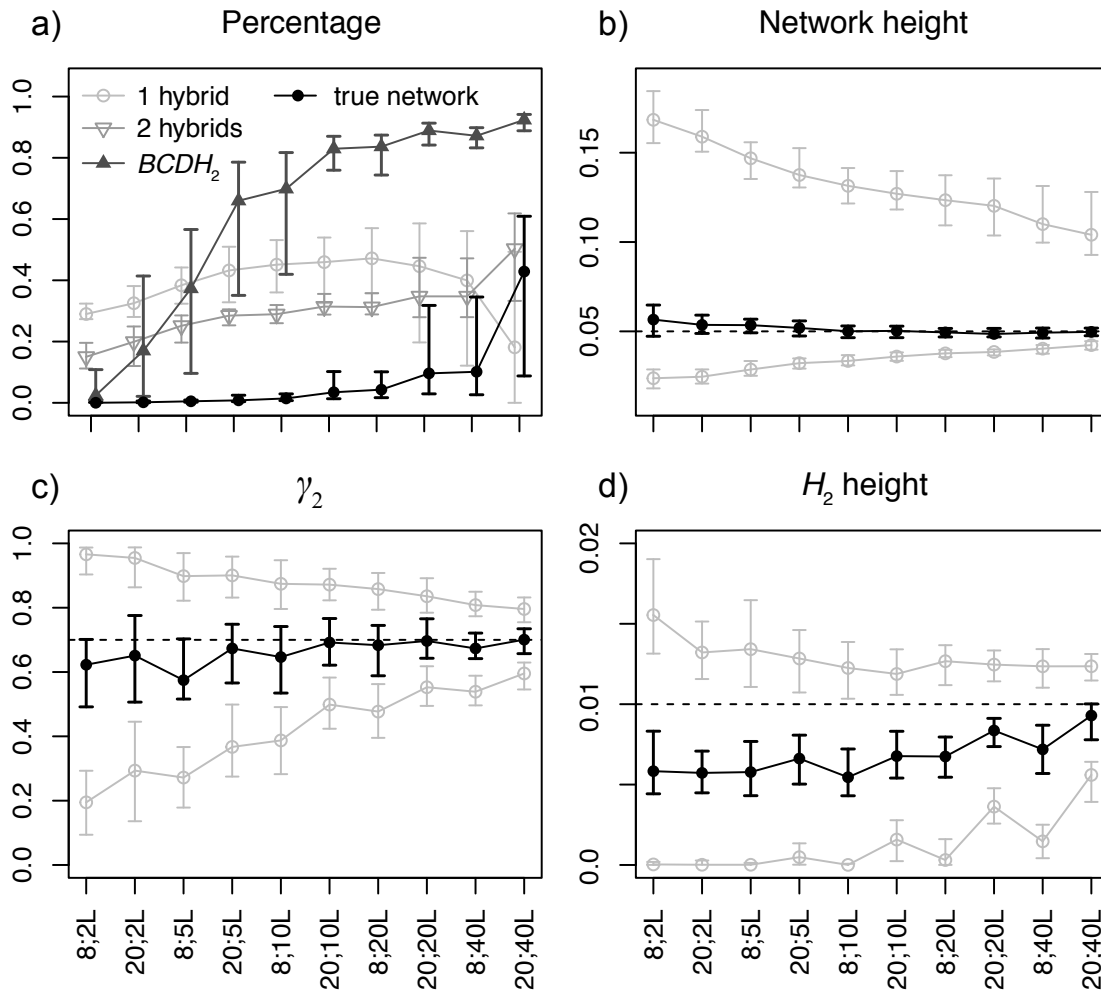


Figure 3.9: Posterior estimates of a) probability of the true network (black), networks with 1 or 2 hybridizations (light gray), and networks with the $BCDH_2$ subnetwork (dark gray), b) network height, c) γ_2 in the $BCDH_2$ subnetwork, and d) H_2 height in the $BCDH_2$ subnetwork, when the data were simulated under the network in Figure 3.1b with sample configuration (2, 2, 2, 2) or (5, 5, 5, 5), and 2, 5, 10, 20, or 40 loci, respectively. For each setting in a), the dot/circle with error bars are the median and the 1st and 3rd quartiles of the percentages of 100 replicates. For each setting in b), c) and d), the black dot with error bars are the median and the 1st and 3rd quartiles of the posterior medians of 100 replicates, the gray circles with error bars are the same summaries for the 95% HPD intervals. The dashed lines indicate the true values.

ure 3.9d). It is not feasible to perform larger scale simulations, e.g., using 100 loci or more, to investigate the power of recovering the ancient hybridization (thus the true species network). Further studies need to be carried out after the efficiency of the operators is improved (see Discussion).

3.4 ANALYSIS OF BIOLOGICAL DATA

3.4.1 THREE CLOSELY RELATED SPRUCE SPECIES

We analyzed a dataset of three spruce species (*Picea purpurea*, *P. likiangensis* and *P. wilsonii*) in the Qinghai-Tibet Plateau (Sun *et al.*, 2014). *P. purpurea* was inferred to be a homoploid hybrid of *P. likiangensis* and *P. wilsonii* (Sun *et al.*, 2014). The original data has 11 nuclear loci and 166 diploid individuals (50 from *P. wilsonii*, 56 from *P. purpurea*, 60 from *P. likiangensis*, and two phased haplotype sequences per individual per locus).

To achieve proper mixing and convergence in a reasonable time, the data was truncated into two non-overlapping datasets by randomly selecting individuals, resulting in 20 individuals from *P. purpurea*, 15 from *P. likiangensis*, and 15 from *P. wilsonii* (100 sequences per locus) for each. The priors for the species network were $\tau_0 \sim \exp(500)$ with mean 0.002, $d = \lambda - \nu \sim \exp(0.01)$ with mean 100, $r = \nu/\lambda \sim U(0, 1)$, and $\gamma \sim U(0, 1)$. The population sizes were integrated out analytically (Eq. 3.4) using $IG(3, 0.003)$ with mean 0.0015 and mode 0.00075. The substitution model was HKY85 (Hasegawa *et al.*, 1985), with independent κ (transition-transversion rate ratio) and state frequencies at each locus. The clock rate was fixed to 1.0 (strict molecular clock across branches) and gene-rate multipliers were used to account for rate variation across loci. The MCMC chain was run for 1 billion steps and sampled every 20,000 steps. The first 35% of samples were discarded as burn-in. For each dataset we obtained two independent runs, and the two runs were checked for effective sample sizes (ESS) and the consistency of trace plots of inferred parameters. The MCMC samples from the two runs were combined.

The species network shown in Figure 3.10 has a posterior probability > 0.95 for both datasets. This confirms that *P. purpurea* is a hybrid species of *P. likiangensis* and *P. wilsonii*. The estimates of γ are 0.33 (0.18, 0.52) and 0.37 (0.17, 0.57) respectively (median and 95%

HPD interval). To investigate the impact of prior on population sizes, we fixed the species network topology to the one in Figure 3.10, and used three priors for the population size parameter: $IG(3, 0.0003)$ with mean 0.00015 (small), $IG(3, 0.003)$ with mean 0.0015 (medium), and $IG(3, 0.03)$ with mean 0.015 (large), respectively. The population sizes were either inferred using MCMC or integrated out analytically. The other priors and MCMC settings were unchanged.

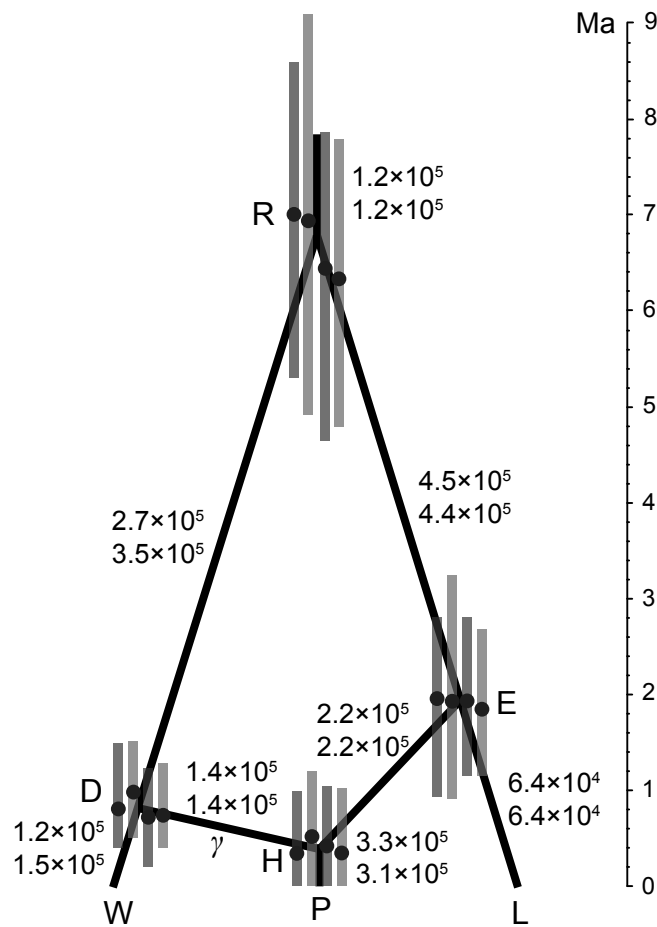


Figure 3.10: The species network with highest posterior probability ($> 95\%$) inferred from the spruce data. The medians and 95% HPD intervals of node heights (black dots with error bars) are in unit of million years. From left to right, they are for dataset 1 with population sizes inferred and integrated out, and dataset 2 with population sizes inferred and integrated out, under the inverse-gamma $IG(3, 0.003)$ prior. The numbers beside the branches are the medians of effective population sizes inferred from dataset 1 (above) and 2 (below). See also Table S1 and S2.

The posterior estimates of γ , node heights, and population sizes are summarized in Supplemental Table S1 and S2. The estimates are similar for both datasets regardless of whether the population sizes were inferred or integrated out under the same prior, but some estimates vary slightly across different priors. Below we summarize the results from the $IG(3, 0.003)$

prior (medium mean) for population sizes (Figure 3.10, and middle column of Table S1 & S2). Around 31–37% of the nuclear genome of *P. purpurea* was derived from *P. wilsonii* (and thus 63–69% from *P. likiangensis*). This estimate is concordant with the original estimate of 31% made using approximate Bayesian computation (ABC) (Sun *et al.*, 2014). Assuming an average substitution rate $\mu = 2 \times 10^{-4}$ per site per million years (Sun *et al.*, 2014), and dividing the node heights (τ 's in Table S1 & S2) by μ , we get the times measured by million years (Figure 3.10). The time of hybridization is inferred to be around 1 Ma. The estimate was 1.3 (0.73, 2.2) Ma in the original analysis assuming the same height for nodes *D*, *E*, and *H*. Moreover, we get an older and narrower estimate for the root age (Figure 3.10), compared to 2.7 (1.4, 6.5) Ma in the original analysis. Similarly, dividing estimates of θ 's (Table S1 & S2) by $\mu = 1 \times 10^{-8}$ per site per generation, we get the effective population sizes (Figure 3.10). The inferred population sizes of *P. purpurea*, *P. wilsonii*, and *P. likiangensis* are smaller than those estimated using ABC (cf. Table 4 in Sun *et al.*, 2014).

3.4.2 SEVEN YEAST SPECIES (*Saccharomyces*)

We re-analyzed another dataset of seven yeast species, including *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. bayanus* (Sbay), *S. castellii* (Scas), and *S. kluyveri* (Sklu). There are in total 106 orthologous gene loci and one sequence per species per locus (Rokas *et al.*, 2003). This data analyzed using concatenation under maximum likelihood yielded a single tree (Figure 3.11a) with 100% bootstrap values at every branch (Rokas *et al.*, 2003). The analysis using BEST (Liu, 2008) showed two main species trees in the posterior (Figure 3.11ab)(Edwards *et al.*, 2007). Both studies discovered extensive incongruent phylogenies from individual genes, with phylogenetic conflict often involving Scas and Sklu. Recently, the full dataset was also analyzed using a Bayesian method co-estimating species networks and gene trees. Extensive hybridization events were found, usually involving Scas and Sklu as the donor or recipient (Wen and Nakhleh, 2017).

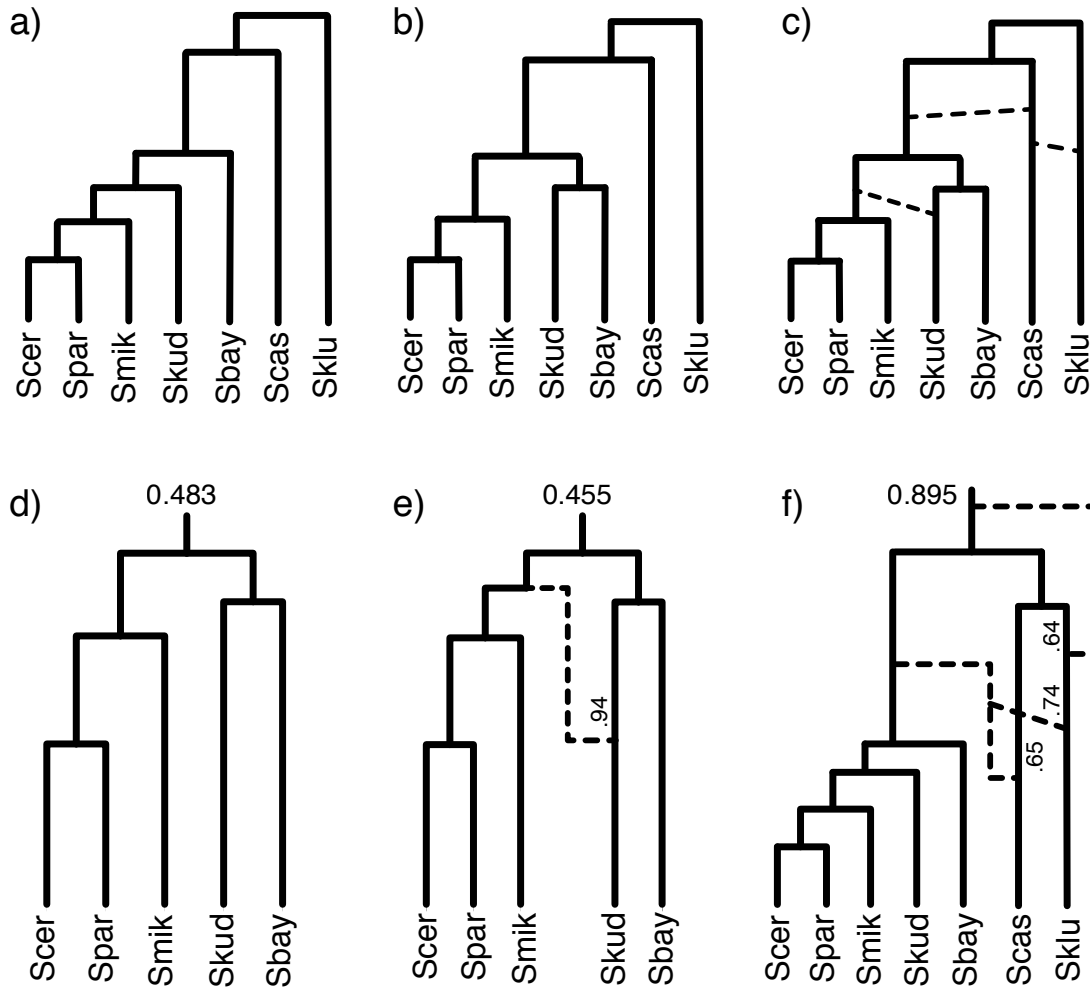


Figure 3.11: The species networks inferred from the yeast data. a) The species tree estimated using concatenation under maximum likelihood (Rokas *et al.*, 2003). a) and b) are the two main species trees in the posterior analyzed using BEST (Edwards *et al.*, 2007). c) The representative species network inferred using our method on all seven species and 106 loci. The dashed lines indicate possible hybridization events. d) and e) are two species networks in the 95% posterior credible set using five species and 106 loci (excluding Scas and Sklu). The posterior probabilities are labelled at the root. f) The species network inferred using seven species and 28 loci with the strongest phylogenetic signal. The larger inheritance probabilities are labelled beside the corresponding branches.

For the inference using our method, the priors for the species network were $\tau_0 \sim \exp(10)$ with mean 0.1, $d = \lambda - \nu \sim \exp(0.2)$ with mean 5, $r = \nu/\lambda \sim U(0, 1)$, and $\gamma \sim U(0, 1)$. The population sizes were integrated out analytically with the prior $IG(3, 2\theta)$, while the mean population size θ was assigned a $\Gamma(2, 100)$ prior, which has a mean of 0.02. We still used the HKY85 substitution model (Hasegawa *et al.*, 1985), gene-rate multipliers for rate variation across loci, and the same MCMC chain settings as in the spruce data analysis.

Similarly, we observed extensive hybridizations among Scas, Sklu, and the remaining five species (Figure 3.11c) in the posterior estimates from independent runs. The incongruence

among gene tree phylogenies are well captured and explained by the hybridization events. These patterns are similar to the results in Wen and Nakhleh (2017). The backbone tree (by removing the reticulation branches with smaller inheritance probabilities from the networks) is consistent with the species tree in Figure 3.11b. However, the complexity of hybridizations caused difficulty and poor mixing of MCMC using the full data. The species network topology may stay unchanged for long durations of the MCMC chain and independent runs give different numbers and directions of hybridizations, although the hybridization pattern and the backbone tree are the same across runs.

Using only five species by excluding Scas and Sklu produced consistent results across runs, and the posterior samples from the three runs are combined. About half of the samples in the 95% posterior credible set are species trees (Figure 3.11d) and another half are species networks with one reticulation leading to Skud (Figure 3.11e). The result of Wen and Nakhleh (2017, Figure 22e) showed only one species network in the 95% posterior credible set with opposite hybridization direction (from Skud to Sbay) and smaller γ than ours (0.75 vs. 0.94). But both analyses have the same backbone tree as in Figure 3.11d. The difference is probably due to the different priors and evolutionary models we used (see Discussion). The inheritance probability of 0.94 is fairly high, such that only a small amount of genes in Skud are horizontally transferred from the ancestral species of Smik. Thus we do not interpret Skud as a hybrid species between Sbay and the ancestral species of Smik. Further investigations are needed to fully understand the underlining biological mechanism. The root heights are both 0.094 (0.092, 0.096) (median and 95% HPD interval) in Figure 3.11de. The branch lengths are measured by the mean substitutions per site. The posterior estimate of mean population sizes θ is 0.00086 (0.00015, 0.0018). The rate multipliers range from 0.55 to 1.5 for the 106 loci, indicating small amount of evolutionary rate variation among loci.

We further investigated the 28 loci with the strongest phylogenetic signal. The gene trees

inferred from these loci under maximum likelihood have at least four internal branches with bootstrap support $> 70\%$ (Luay Nakhleh, personal communications). The priors and MCMC settings are the same as for the 106 loci. Using all the seven species, the species network with highest posterior probability (0.895) is shown in Figure 3.11f. Three hybridization events are recovered, two from an ancestor of *Saccharomyces sensu stricto* (Scer, Spar, Smik, Skud and Sbay) into Scas and Sklu. In addition, we found a hybridization event deriving from an ancestral species of sampled extant *Saccharomyces*. When using only five species by excluding Scas and Sklu in a separate run, the species tree with the same topology as the subtree in Figure 3.11f was inferred with highest posterior probability (0.98). This is different from the backbone tree using all 106 loci (cf. Figure 3.11d), indicating conflicting phylogenetic signal among loci.

3.5 DISCUSSION

We use species networks to model reticulate evolution. Although our method is motivated by species hybridization, species networks can also be applied to studies of migration and lateral (or horizontal) gene transfer. Rates of migration between taxa have been modeled previously using isolation-with-migration (IM) models (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007; Wilkinson-Herbots, 2008; Hey, 2010; Zhu and Yang, 2012; Dalquen *et al.*, 2017). Reticulation branches in species networks can also model migration, and may be a more natural fit when migration is not constant as in the case of secondary contact. The proportion of genetic material inherited through a reticulation event can come from a high rate of migration over a short period of time, or a lower rate of migration over a longer period of time. Lateral gene transfer has been modeled previously using gene duplication, transfer and loss (DTL) models (Tofigh *et al.*, 2011; Szöllősi *et al.*, 2012, 2013; Sjöstrand *et al.*, 2014; Szöllősi *et al.*, 2015). These models account for discordance between species tree and gene trees by any

discordance to DTL events, but ignore incomplete lineage sorting. Our implementation of species networks explicitly models the embedding of gene trees within a species network, and so can be used to infer lateral gene transfer events without confounding them with incomplete lineage sorting.

Methods to build a species network (e.g., Wu, 2010; Park *et al.*, 2010; Albrecht *et al.*, 2012) traditionally use inferred gene trees from each locus without accounting for their uncertainties, and employ nonparametric criteria such as parsimony. For population level data, the sequences are similar and the signal in gene tree topologies is typically low, so using fixed gene trees is assigning too much certainty to the data. These methods typically assume that gene tree discordance is solely due to reticulation, thus may suffer in the presence of incomplete lineage sorting (Yu *et al.*, 2011). The MSNC model (Yu *et al.*, 2014) provides a statistical framework to account for both incomplete lineage sorting and reticulate evolution. But properly analyzing genetic data to infer species networks under the MSNC model is a challenging task. There have been methods using only the gene tree topologies from multiple loci under MSNC (Yu *et al.*, 2012, 2014; Wen *et al.*, 2016). However, gene trees with branch lengths are more informative for inferring species tree or network topology than gene tree topologies alone. Accounting for branch lengths can improve distinguishability of species networks (Pardi and Scornavacca, 2015; Zhu and Degnan, 2017). Although methods using estimated gene trees (with branch lengths) from bootstrap or posterior samples as input take into account gene tree uncertainty (Yu *et al.*, 2014; Wen *et al.*, 2016), directly using sequence data to co-estimate species networks and gene trees in a Bayesian framework showed improved accuracy (Wen and Nakhleh, 2017). Pseudo-likelihood approaches (Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016) compute faster than full likelihood or Bayesian approaches, but have severe distinguishability issues and require more data to achieve good accuracy.

At the time of writing, another Bayesian method inferring species networks and gene trees

simultaneously from multilocus sequence data was released (Wen and Nakhleh, 2017). The general framework here is similar, but we highlight four major differences. We use a birth-hybridization prior for the species network which naturally models the process of speciation and hybridization. The prior is extendable to account for extinction, incomplete sampling, and rate variation over time, as we outline below. Wen and Nakhleh (2017) used a descriptive prior combining a Poisson distributed number of reticulations with exponential distributed branch lengths. Secondly, we allow parallel branches in the network. This is biologically possible. Even if the true species history has no parallel branches, the observed species network can still contain such features due to incomplete sampling. Note though that a very large number of individuals and loci are required to detect such parallel branches. To prevent the species network from growing arbitrarily big, such that it becomes indistinguishable by the gene trees (Pardi and Scornavacca, 2015; Zhu and Degnan, 2017), we typically assign an informative prior to ensure the hybridization rate is lower than the birth rate. A similar strategy was used in Wen *et al.* (2016); Wen and Nakhleh (2017) by restricting the rate of the Poisson distribution. Third, we account for the uncertainty in the embedding of a gene tree within a species network by estimating the MSNC probability conditional on a proposed embedding at each MCMC step. This provides a posterior distribution of gene trees and their embeddings within a species network, enabling analysis of which alleles are derived from which ancestral species. The cost instead is additional MCMC operations compared to integrating over all embeddings at each step (Wen *et al.*, 2016; Wen and Nakhleh, 2017). Last but not least, we implement analytical integration over population sizes in the species network (Eq. 3.4). This reduces the number of parameters for the rjMCMC operators to deal with, and should improve convergence and mixing. Finally, our implementation in `SpeciesNetwork` is an extension to BEAST 2 (Bouckaert *et al.*, 2014), to take advantage of many standard phylogenetic models, such as different substitution models, relaxed molecular clock models, and the

BEAUTi graphical interface.

In our approach, we employ a simple prior for the species network based on a birth-hybridization model. Analogous to birth-death priors for species trees (e.g., Stadler, 2010; Heath *et al.*, 2014), the birth-hybridization prior could be extended to account for extinction and incomplete sampling, to model networks containing both extant and fossil taxa. The rates could also be allowed to vary over time, to model the diversification patterns during speciation (the skyline model for trees, Stadler *et al.*, 2013). When considering networks instead of trees, techniques to derive the probability density of trees cannot be directly applied as the hybridization rate depends on pairs of lineages rather than individual lineages. This non-linearity necessitates solving differential equations to derive the species network probability densities, a task which we defer to a later study.

Our approach is limited in computational speed. The empirical analysis was done, e.g., on only three species with 50 individuals and 11 loci, or up to seven species and 106 loci but one individual per species. The main bottleneck is the MCMC operators. Due to hard constraints between the species network and embedded gene trees (Figure 3.2), MCMC operators changing them separately limit the ability to analyze genomic scale data from many individuals. More specifically, updating the species network will likely violate a gene tree embedding, resulting in very low acceptance rate of the operator. Thus it will be essential to design more efficient MCMC operators. There have been coordinated operators that can change species tree and gene trees simultaneously (Rannala and Yang, 2017; Jones, 2017). Such operators could possibly be extended to species networks, and will potentially improve efficiency of the MCMC algorithm. Proposing new embeddings of gene trees in species network is also costly. Thus it might be worthwhile to integrate over the embeddings (Wen *et al.*, 2016; Wen and Nakhleh, 2017) if they are not of interest. Moreover, there are methods to integrate out the gene trees under the multispecies coalescent model when analyzing biallelic genetic markers

(RoyChoudhury *et al.*, 2008; Bryant *et al.*, 2012; Zhu *et al.*, 2017). However, it is not yet feasible to apply this strategy to multilocus sequence alignment. Computationally, implementing Metropolis-coupled MCMC (MC³, Geyer, 1991) will help to overcome multiple local peaks in the posterior, and further parallelizing the cold and hot chains will gain speed.

In summary, we developed a Bayesian method for inferring species networks together with gene trees and evolutionary parameters from multilocus sequence data. The method is implemented within a general Bayesian framework, with potential future extensions to the theoretical model and to the practical implementation.

3.6 SUPPLEMENTARY MATERIAL

Supplementary material is available at Molecular Biology and Evolution online.¹

3.7 ACKNOWLEDGMENTS

This research was supported by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant number 335529 to T.S.). C.Z. acknowledges his salary as well as a visit covered by this grant to the Centre for Computational Evolution, University of Auckland, New Zealand in mid-2016. H.O. was supported by an Australian Laureate Fellowship awarded to Craig Moritz by the Australian Research Council (FL110100104). We sincerely thank Simone Linz for detailed discussion on modeling phylogenetic networks. We also thank three anonymous reviewers and the editors for many constructive comments leading to great improvement of the original manuscript.

¹<https://doi.org/10.1093/molbev/msx307> — accessed 15th December 2017

4

Inferring Species Trees Using Integrative Models of Species Evolution

ABSTRACT

Bayesian methods can be used to accurately estimate species tree topologies, times and other parameters, but only when the models of evolution which are available and utilized sufficiently account for the underlying evolutionary processes. Multispecies coalescent (MSC) models have been shown to accurately account for the evolution of genes within species in the absence of strong gene flow between lineages, and fossilized birth-death (FBD) models have been shown to estimate divergence times from fossil data in good agreement with expert opinion. Until now dating analyses using the MSC have been based on informally derived node priors instead of the FBD. On the other hand, dating analyses using an FBD process have con-

catenated all gene sequences and ignored coalescence processes. To address these mirror-image deficiencies in evolutionary models, we have developed an integrative model of evolution which combines both the FBD and MSC models. Using an exemplar data set consisting of molecular sequence data and morphological characters from the dog and fox subfamily *Caninae*, we show that concatenation causes predictable biases in estimated branch lengths, and the same biases are also observed when the FBD is used with concatenation. These biases can be avoided by using the FBD-MSC model, which we have implemented in a new version of StarBEAST2, a package developed for the BEAST2 phylogenetic software.

4.1 INTRODUCTION

We have vastly more data on biological organisms than at any point in the past; whole genome sequences, ancient DNA, morphological characters and fossil occurrences all contain a fingerprint of past evolutionary processes. With this wealth of data, we should expect coherent estimates of the pattern and timing of evolutionary events. Yet the story told by genomes and molecular clocks is often difficult to reconcile with morphological data (Paterson *et al.* 2014) and the fossil record (Meyer *et al.* 2012; O’Leary *et al.* 2013; dos Reis *et al.* 2014; Jarvis *et al.* 2014; Mitchell *et al.* 2015). These debates are often described as “rocks versus clocks” (Donoghue and Benton 2007) with famous examples including the timing of the origin of placental mammals (O’Leary *et al.* 2013; dos Reis *et al.* 2014), birds (Jarvis *et al.* 2014; Mitchell *et al.* 2015), flowering plants (Beaulieu *et al.* 2015), and the Cambrian Explosion (Lee *et al.* 2013). Most disturbingly, these debates persist even for evolutionarily recent and intensively studied questions like the timing of the human-chimp split, where fossils (Brunet *et al.* 2002; White *et al.* 2009; Wood and Harrison 2011; White *et al.* 2015) give different results than genomic data (Meyer *et al.* 2012; Patterson *et al.* 2006; Langergraber *et al.* 2012; Scally *et al.* 2012; Scally and Durbin 2012; Callaway 2015; Lipson *et al.* 2015).

Bayesian inference, the gold-standard in estimating evolutionary history (Huelsenbeck *et al.* 2001; Ronquist and Huelsenbeck 2003; Nylander *et al.* 2004; Drummond *et al.* 2012; Bouckaert *et al.* 2014; Höhna *et al.* 2016), provides a theoretical framework that supports the integration of multiple data sources. So called “total-evidence” analyses integrate molecular sequence and morphological character data. Where a fossil record is available, total-evidence data sets can be used with “tip-dating” methods to estimate time-calibrated species trees (Gavryushkina *et al.* 2017; Ronquist *et al.* 2012b; Zhang *et al.* 2016).

Tip-dating makes an advance over previous methods such as node-dating or a fixed clock by treating fossils as data. Node-dating, where researchers propose parametric prior distributions for the dates of particular nodes based on expert opinion and intuition, has been shown to infer misleading node ages (Gavryushkina *et al.* 2017). An alternative to tip- or node-dating is a fixed molecular clock. Fixing the molecular clock at 1 means that only relative divergence times can be estimated, while using a value from a previous study assumes that the *a priori* rate is accurate for the species and loci in the new study.

Previous implementations of tip-dating have so far made the assumption of a single phylogeny for all molecular loci and for the morphological characters. This assumption is known as “concatenation” because it is equivalent to concatenating several multiple sequence alignments into a single alignment, and it has been demonstrated to cause biases and overestimated precision when inferring species trees from molecular data (Liu *et al.* 2015, see also Chapters 1 and 2).

To enable the combined use of molecular, morphological and fossil data without the known problems of concatenation, we propose combining models of genealogical evolution, morphological evolution, and of speciation, extinction and fossilization.

THE FOSSILIZED BIRTH-DEATH PROCESS

Explicitly including fossils in stochastic models of phylogenies became possible with the birth-death-serial-sampling model (Stadler 2010), which added a fossil sampling rate (ψ) to speciation and extinction rates (λ and μ). This model was later re-named the fossilized birth-death (FBD) process (Heath *et al.* 2014). The “skyline” extension to the FBD (Stadler *et al.* 2013) allows all three parameters to vary through time in an arbitrary and independent fashion while the sampled ancestor extension (Gavryushkina *et al.* 2014) correctly treats fossil placement on the tree as a random variable where each fossil may be either a direct ancestor of other samples, or a tip branch if no descendants have been sampled.

THE MULTISPECIES COALESCENT

Modern phylogenetic inference distinguishes between high level phylogenetic relationships across species described by a species tree and relationships between individual specimens described by gene trees. It is now well understood that failure to take this into account can significantly bias results due to the effects of incomplete lineage sorting (ILS) and other processes (Liu *et al.* 2015; Linkem *et al.* 2016; Mendes and Hahn 2016, 2017).

StarBEAST2 (Chapter 2), BEST (Liu 2008) and BPP (Yang 2015; Rannala and Yang 2017) are all examples of Bayesian software that explicitly sample the joint posterior distribution over both species and gene trees under the multispecies coalescent (MSC) model, as described by Maddison 1997 and Degnan and Rosenberg 2009. These methods account for the hierarchical nature of the evolutionary process and explicitly model ILS. However none of these implementations allow fossils or other ancestral samples to be placed directly on the species tree.

INTEGRATIVE MODELS OF SPECIES EVOLUTION

Integrative models are desirable because they can integrate over uncertainty rather than assuming fixed parameters, and they can also directly utilise more sources of data than simpler models. In this paper we describe an integrative Bayesian phylogenetic model for estimating species trees and divergence times, capable of analyzing multilocus genetic data, fossil occurrence data and morphological data in a coherent probabilistic inference framework.

The model reconciles molecular and fossil evidence by explicitly distinguishing two evolutionary processes. The species tree is modeled using the FBD process, with the morphology of all species arising from a stochastic process of evolution that proceeds down the branches of this species tree. The molecular sequence data (sampled from extant individuals or as ancient DNA) are related by multiple independent gene trees, which may differ from each other due to processes such as ILS, but must be consistent with the shared species tree that they have all evolved within (Fig. 4.1).

The core of our work is combining two recent advances in phylogenetic modeling into a single coherent inference method. On the one hand we have the MSC, which has become the standard model for describing the relationship between molecular genealogies and species trees. On the other hand we have the FBD branching model of macroevolution, which describes speciation, extinction and fossilization processes. Conceptually these models have a natural hierarchical relationship, with the FBD model describing the distribution over species trees and the MSC model describing the probability distribution of molecular genealogies conditional on the species trees.

The BEAST2 phylogenetic software features “StarBEAST2” — a recent implementation of the MSC — and a implementation of the FBD prior with sampled ancestors (SA; Gavryushkina *et al.* 2014). We have updated StarBEAST2 to combine the MSC model with the FBD with SA process, henceforth “FBD-MSA”. To demonstrate the utility of the FBD-MSA model,

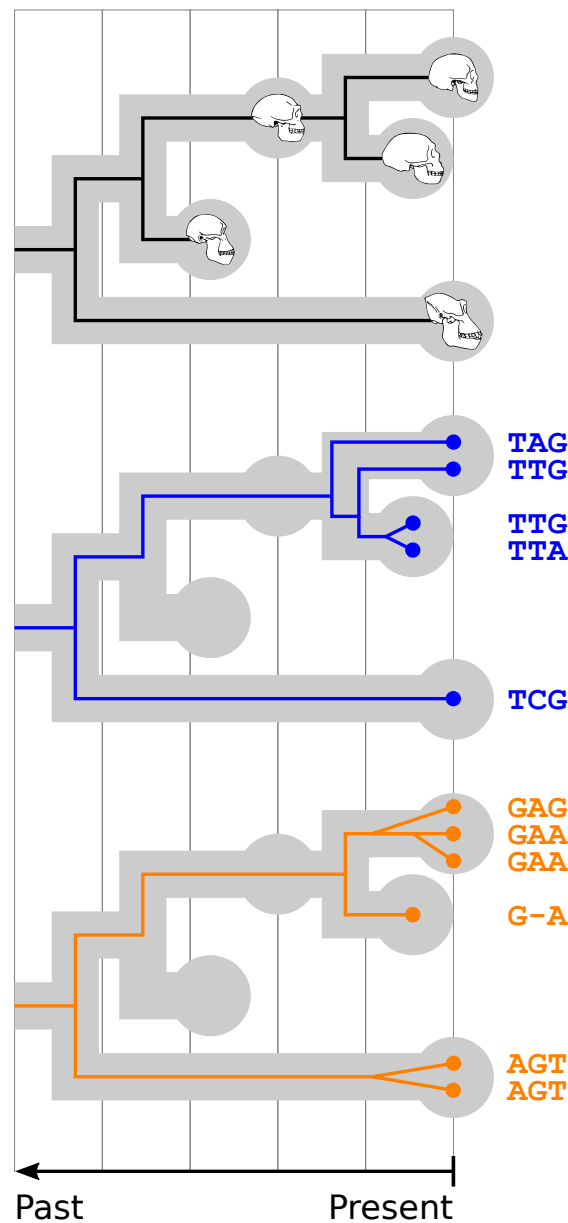


Figure 4.1: A species tree with a single sampled ancestor and its relationship to morphological data (top) and multilocus sequence alignments (middle and bottom) in a unified model.

we applied the latest version of StarBEAST2 to an exemplar data set of the dog and fox sub-family Caninae.

Estimates made under the FBD-MSM model are compared with estimates made using FBD with concatenation (henceforth “FBD-concatenation”), the MSM with a fixed molecular clock instead of an FBD prior, and concatenation with a fixed clock. FBD-MSM results were generally in agreement with fixed clock MSM estimates. Concatenation overestimated tip branch lengths, species divergence times, and the timing of diversification leading to extant Caninae, even when fossil data was incorporated using the FBD model.

4.2 METHODS

INTEGRATIVE MODEL PROBABILITY

The integrative model combining the MSC, the FBD process, and morphological evolution can be expressed by combining the component likelihoods. The likelihood of a gene is the phylogenetic likelihood (Felsenstein 1981) $P(D_i|G_i)$ where D_i is the multiple sequence alignment for the i th gene tree G_i . The MSC probability for that gene tree is $P(G_i|S)$ where S is the species tree. The likelihood of a morphological character is the phylogenetic likelihood $P(C_j|S)$ where C_j is the vector of states for the j th character. The prior probability of the species tree under the FBD process is $P(S|\theta)$, where θ is a vector of FBD parameters as described by Gavryushkina *et al.* 2014. Combining the likelihoods for the integrative model we get the probability of the species tree given the molecular, morphological and fossil data:

$$P(S|D) = \prod_i (P[D_i|G_i] \cdot P[G_i|S]) \cdot \prod_j P(C_j|S) \cdot P(S|\theta) \quad (4.1)$$

SAMPLING AND SIMULATING TREES FROM THE PRIOR

We tested our implementation of FBD-MSC by jointly sampling SA trees with a single embedded gene tree from the prior, and comparing those distributions with independently sampled SA trees and simulated gene trees.

Three and four-taxon SA trees were sampled from the prior using the SA package in BEAST2 (Gavryushkina *et al.* 2014). Following the parameterization in Gavryushkina *et al.* 2014, these trees were conditioned on an origin time t_{or} of 3, a birth rate λ of 1, a death rate μ of 0.5, a sampling rate ψ of 0.1, a removal probability r of 0 and a present-day sampling probability ρ of 0.1.

The sampled taxa for three-taxon SA trees were termed A, B and C, and had fixed ages of 0, 1, and 1.5 respectively. The sampled taxa for the four-taxon SA trees were termed A, B, C and

D, and had fixed ages of 0, 0, 0.5 and 2 respectively.

MCMC chains to sample SA trees were run for 100 million steps, and 100,000 trees sampled at a rate of 1 per 1,000 steps. One gene tree was simulated for each sample using custom Java code available as part of the StarBEAST2 package, assuming effective population sizes fixed at 1 for each branch.

When jointly sampling SA trees with an embedded gene tree from the prior using StarBEAST2, identical parameters were used but MCMC chains were run for 500 million steps. Species and gene trees were sampled at a rate of 1 per 5,000 steps for 100,000 species trees and the same number of gene trees.

COMPILING CANINAE DATA

Unphased molecular sequences were retrieved from NCBI GenBank. Sequences from Bardeleben *et al.* 2005a had accession numbers AY609082–AY609158. Sequences from Bardeleben *et al.* 2005b had accession numbers AY885308–AY885426. Sequences from Lindblad-Toh *et al.* 2005 had accession numbers DQ239439–DQ239486 and DQ240289–DQ240817. Outgroup (non-Caninae) and domestic dog sequences were discarded. *Canis aureus* was renamed *Canis anthus* following Koepfli *et al.* 2015. For each locus, we aligned those sequences to produce a multiple sequence alignment (MSA) using PRANK (Löytynoja and Goldman 2005). Phased MSAs were generated by duplicating each aligned sequence and randomly phasing heterozygous sites.

Coded morphological data, character names, character state names and tip dates from Slater 2015 were retrieved from Dryad¹. This data set built on monographs from Wang 1994, Wang *et al.* 1999 and Tedford *et al.* 2009.

Outgroup characters and characters invariable within Caninae were discarded. *Canis aureus* was again renamed *Canis anthus*, and *Cuon javanicus* was renamed *Cuon alpinus*, a synonym

¹<https://doi.org/10.5061/dryad.9qd51> — accessed 15th December 2017

used in the molecular sequence data. For species with molecular sequences but no morphological data, all characters were treated as missing data. An extant-only data set was produced by discarding fossil taxon characters, and characters invariable within extant Caninae. BEAST2-compatible NEXUS files were generated containing the coded data and names.

MSC AND CONCATENATION ANALYSES

The MSC (in practice, StarBEAST2) was configured to estimate a constant population size separately for each branch, with a maximum effective population size of 2, and a $1/X$ prior on the mean population size. Phased sequences were used with StarBEAST2, and unphased sequences with concatenation. For both StarBEAST2 and concatenation we set uniform priors on λ and μ .

The mean substitution rate was either fixed at 8×10^{-4} , or estimated with a lognormal prior which had a mean of 7.5×10^{-4} and a standard deviation of 0.6. Substitution rates among loci were allowed to vary with a flat Dirichlet prior. The HKY substitution model was used for molecular data (Hasegawa *et al.* 1985), and transition/transversion ratios estimated separately for each locus. The Mkv model (Lewis 2001) was used to model the evolution of morphological characters, assuming character state frequencies and transition rates are all equal. A morphological clock was estimated with a $1/X$ prior and a maximum rate of 1.

FBD analyses were conditioned on t_{or} which was estimated with a uniform prior. The sampling rate ψ was also estimated with a uniform prior. The other FBD parameters r and ρ were fixed at 0 and 1 respectively.

For each fixed clock analysis, we ran 20 independent MCMC chains of 400 million states each, sampling once every 200,000 states. For each fossilized birth-death analysis, we ran 20 independent MCMC chains of 10 billion states each, sampling once every 2 million states. After discarding the first 10% of samples from each chain as burnin, independent chains were concatenated and subsampled for a combined sample of 2,000 states.

POSTERIOR PREDICTIVE SIMULATIONS

For half (1,000) of the fixed clock StarBEAST2 posterior samples, we resimulated molecular and morphological data. For each locus a gene tree was simulated according to the MSC using DendroPy (Sukumaran and Holder 2010), embedded within the species tree (topology, times and per-branch population sizes) for that sample, with two individuals per extant species. An MSA was simulated for each gene tree using Seq-Gen (Rambaut and Grassly 1997), using the HKY model with the estimated κ ratio and substitution rate of the locus from the sample, and of the same length as the original locus. Unphased per-species sequences were generated using ambiguity codes for heterozygous sites.

Morphological data was resimulated by simulating a 1,000 character MSA along a sampled species tree with 20 states per character, again using Seq-Gen. Base frequencies and transition rates were all equal, and the substitution rate set to 0.03. Then for each morphological character in the original data set, we sampled without replacement one of the simulated characters with a matching number of observed states.

Each simulation was reanalyzed using concatenation with the same model and priors as for the original data set. However only one chain of 200 million states was run for each simulation, sampling once every 80,000 states, and 20% of samples were discarded as burnin.

CALCULATING SUMMARY STATISTICS

Summary statistics were calculated for each estimated distribution of trees using DendroPy. These included the maximum clade credibility (MCC) tree, branch lengths, node heights, branch support and node support. In this study, a node is defined as the root of a subtree containing all of, and only, a given set of extant taxa. A branch is defined as the direct connection between parent and child nodes as defined above. Lineages-through-time (LTT) curves for FBD analyses were calculated using a custom script. Summary statistics and LTT plots were

visualized using `ggplot2` (Wickham 2016) and `ggtree` (Yu *et al.* 2017).

4.3 RESULTS

FBD-MSC IMPLEMENTATION CORRECTNESS

To test the correctness of our FBD-MSC implementation, we first compared distributions of three and four-taxon FBD with SA trees drawn from the prior using BEAST2 without the MSC, to distributions drawn from the prior using the FBD-MSC model in StarBEAST2.

The marginal divergence time (Fig. B.1,B.2) and topology (Fig. B.3,B.4) distributions thus generated were found to be identical between implementations. As the BEAST2 implementation of the FBD model has been previously verified (Gavryushkina *et al.* 2014), this is strong evidence that the new implementation is also correct.

Gene trees were also sampled from the prior under the FBD-MSC model in StarBEAST2, and were compared to a distribution of gene trees simulated evolving within the FBD with SA trees that were drawn from the prior absent StarBEAST2. The distributions of gene tree coalescent times (Fig. B.5,B.6) and topologies (Fig. B.7,B.8) were identical for either method, further supporting the mathematical correctness of our implementation.

COMPILING AN EXEMPLAR DATASET

To demonstrate the effects of estimating species divergence times without accounting for coalescent processes, as when using concatenation, we compiled a data set by combining 19 previously published Caninae nuclear locus sequences from extant Caninae taxa (Table 4.1) with morphological characters and times from extant and fossil Caninae (Slater 2015).

The combined data set included 21 extant taxa with molecular data only, 9 extant taxa with molecular and morphological data, and 31 fossil taxa with tip dates and morphological data. After removing characters with no variation within Caninae, there were 72 morphological

Table 4.1: Nineteen nuclear loci used in this study.

Locus name	MSA length*	Publication
APOBS1	702	Lindblad-Toh <i>et al.</i> 2005
BDNF	489	Lindblad-Toh <i>et al.</i> 2005
BRCA1S2	741	Lindblad-Toh <i>et al.</i> 2005
Ch14	921	Lindblad-Toh <i>et al.</i> 2005
Ch21	601	Lindblad-Toh <i>et al.</i> 2005
Ch24	730	Lindblad-Toh <i>et al.</i> 2005
CHRNA1	376	Bardeleben <i>et al.</i> 2005a
CHST12	705	Lindblad-Toh <i>et al.</i> 2005
CMKOR1	735	Lindblad-Toh <i>et al.</i> 2005
CYPIA1	619	Bardeleben <i>et al.</i> 2005a
FES	483	Bardeleben <i>et al.</i> 2005a
FGFR3	503	Lindblad-Toh <i>et al.</i> 2005
GHR	821	Bardeleben <i>et al.</i> 2005a
RAG1	741	Lindblad-Toh <i>et al.</i> 2005
TMEM20	615	Lindblad-Toh <i>et al.</i> 2005
TRSP	722	Bardeleben <i>et al.</i> 2005b
VANGL2	546	Lindblad-Toh <i>et al.</i> 2005
VTN	487	Bardeleben <i>et al.</i> 2005a
VWF	732	Lindblad-Toh <i>et al.</i> 2005

*the number of sites in the multiple sequence alignment

characters remaining for FBD analyses. After further removing characters with no variation among the 9 taxa with both molecular and morphological data, there were 55 remaining for fixed clock analyses.

CALIBRATING SPECIES TREES USING A FIXED CLOCK

In the absence of a fossil record for a clade of interest, divergence times can be estimated using a fixed molecular clock. This scales the tree by an *a priori* chosen substitution rate, or a set of substitution rates for a set of genes. Substitution rates have been previously estimated for the nuclear RAG1 gene across multiple tetrapod clades, and for mammals the rate is approximately 1×10^{-3} substitutions per site per million years (Hugall *et al.* 2007). Exploratory analyses suggested that RAG1 evolves around 25% more quickly than the mean rate for all genes in our study, so we used a substitution rate fixed at 8×10^{-4} for analyses calibrated with a fixed clock.

We compared the posterior distribution of species trees inferred under the MSC and concatenation without any fossil data, including nuclear loci and morphological characters only from extant taxa, and using a birth-death prior for the species tree. The estimated lengths of all tip branches and some internal branches were longer when using concatenation (Fig. 4.2). A few internal branches were shorter, for example the 1–2 and 5–A branches.

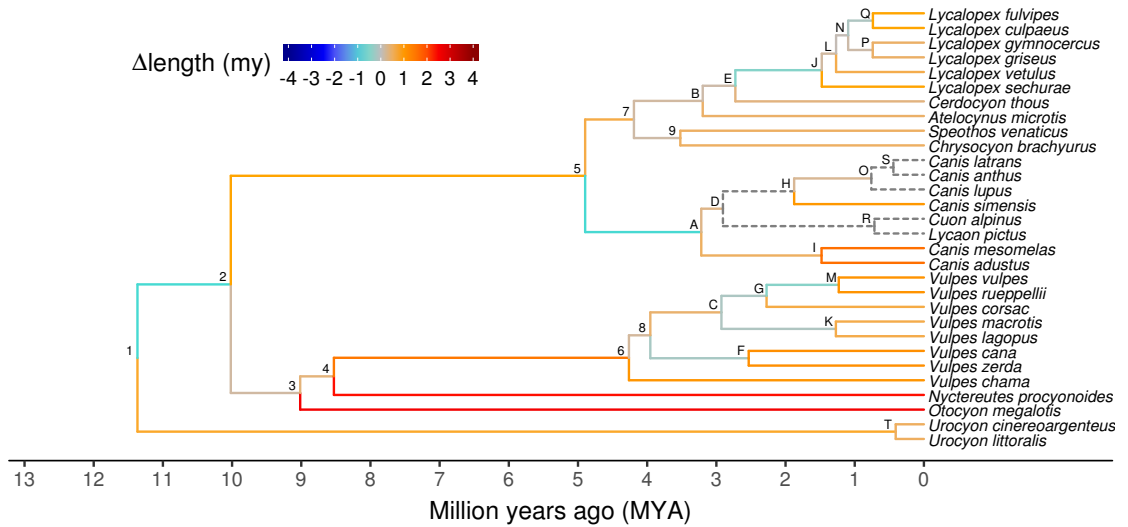


Figure 4.2: Branch length changes resulting from concatenation using a fixed clock. The color shows how branch length estimates differ when using concatenation rather than the multispecies coalescent (MSC). The tree is a maximum clade credibility (MCC) summary tree with mean node ages, generated from the MSC posterior distribution of species trees, inferred using molecular and extant morphological data with a fixed clock. The difference in branch lengths is the mean among concatenation samples including that branch, less the MSC mean. Dashed lines represent branches with less than 0.5% support using concatenation.

To understand whether failing to account for neutral coalescent processes could cause the observed branch length differences, we used posterior predictive simulations to model the expected differences. For 1,000 species tree samples in the fixed clock MSC posterior distribution, we resimulated gene trees according to the MSC. For each simulated gene tree, we simulated a multiple sequence alignment based on that sample’s substitution rates and transition/transversion ratios. A set of morphological characters were also simulated along the species tree for each sample. Posterior distributions of species trees using concatenation were then inferred from the simulated data.

For a given branch, we calculated the distribution of differences in branch length $\Delta(l_b - \bar{l}_b)$

between the assumed true length l of a branch b , and the concatenation estimate \bar{l}_b . This calculation was based on the replicates where the species tree used for simulation contained b . \bar{l}_b is the expectation marginalized over all samples containing b . In the case of phylogenetic cherries, only one branch was included, because their lengths are always equal in an ultrametric tree.

All observed differences in branch lengths fell within expectations (Fig. 4.3). This suggests that the failure to account for neutral coalescent processes, as modeled by the MSC, is responsible for the observed differences.

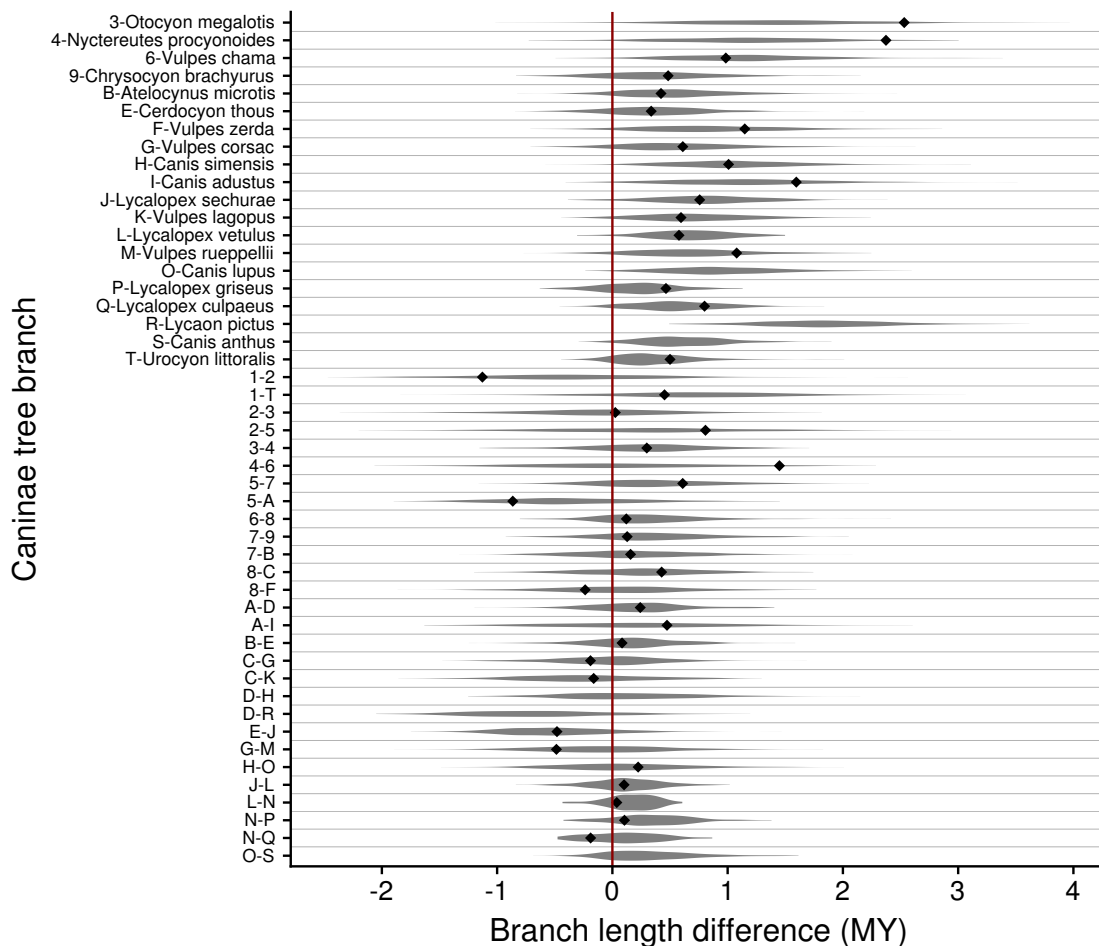


Figure 4.3: The expected and observed effects of concatenation on branch lengths. Branches correspond to those in Figure 4.2 and 4.4. Violin densities represent the distribution of differences between the posterior predictive simulations and corresponding concatenation estimates. Diamonds pinpoint the differences between multispecies coalescent and concatenation estimates using molecular and extant morphological data. Diamonds are missing where the support for a branch using concatenation is less than 0.5%. Estimates to the right of zero line (red) are longer when using concatenation, estimates to the left are shorter.

CALIBRATING SPECIES TREES USING FOSSIL DATA

Using a fixed molecular clock conditions the estimated divergence times on the accuracy of the *a priori* chosen substitution rate. The rate of molecular evolution is inversely associated with body size in mammals (Bromham 2011) so the substitution rate used for, say, baleen whales would likely be too slow when applied to Muridae. Instead the molecular substitution rate can be inferred jointly with the species tree topology and times by including fossil data and applying an FBD prior to the species tree.

We reran our concatenation and MSC analyses of Caninae after including morphological data with tip dates (fossils), and applied FBD with SA priors to the species trees. The placement of fossil taxa was very uncertain, so to make the FBD results interpretable we pruned the posterior distributions of species trees to include only extant taxa. This also enables direct comparisons of the FBD and fixed clock results. The MCC tree topology inferred by FBD-MSc was identical to fixed clock MSc (Figs. 4.2,4.4).

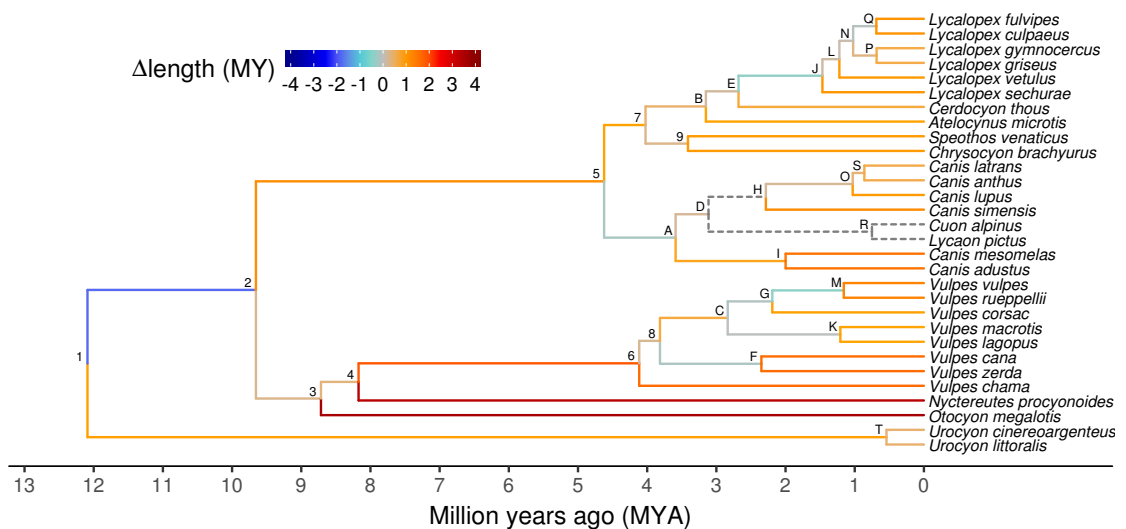


Figure 4.4: Branch length changes resulting from concatenation using the fossilized birth-death (FBD) model. The color shows how branch length estimates differ when using the concentration rather than the multispecies coalescent (MSC). The tree is a maximum clade credibility (MCC) summary tree with mean node ages, generated from the MSC posterior distribution of species trees, inferred using molecular and morphological data and fossil times. The difference in branch lengths is the mean among concatenation samples including that branch, less the MSC mean. Dashed lines represent branches with less than 0.5% support using concatenation.

The differences in branch lengths observed for FBD-concatenation compared to FBD-MSc

were very similar to those seen in the fixed clock scenario (Fig. 4.5). All branches with longer estimated lengths using concatenation and a fixed clock also had longer estimates using FBD-concatenation compared to the corresponding FBD-MSc estimates. The same applied to branches with shorter estimated lengths using concatenation (Figs. 4.2,4.4).

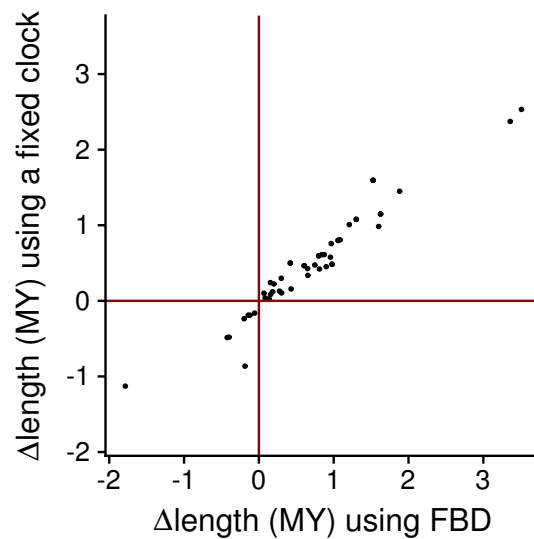


Figure 4.5: Consistency in how branch lengths change under concatenation rather than the multispecies coalescent (MSC). Differences are when using a fixed clock (y-axis) or when using the fossilized birth-death (FBD) process (x-axis). All branches present in MSC maximum clade credibility trees (Figs. 4.2, 4.4) with at least 0.5% support when using concatenation are included. Branch lengths in the top-right quadrant are overestimated by concatenation using a fixed clock or FBD, those in the bottom are underestimated by concatenation using either method.

Similar estimates were made of macroevolutionary parameters using the MSC and concatenation models, as long as the same species tree prior was used (Table 4.2). When using the FBD prior, the molecular clock rate highest posterior densities (HPDs) included the *a priori* rate of 8×10^{-4} with either the MSC or concatenation. The only non-overlapping HPDs were for the morphological clock rate, which was inferred to be slower when using the FBD compared to a fixed clock. The lower bound for turnover (extinction relative to speciation) was approximately zero when fossil data was not included, but was forced higher by the explicit inclusion of extinct species for FBD analyses.

Table 4.2: Macroevolutionary parameter estimates.

Clade	Multispecies coalescent		
	MO	FC	FBD
Molecular clock rate ($\times 10^{-4}$)	8	8	8.21 (6.16–9.81)
Morphological clock rate	NA	0.13 (0.09–0.17)	0.05 (0.04–0.06)
Mean $N_e g$	0.47 (0.37–0.58)	0.50 (0.40–0.62)	0.51 (0.37–0.67)
Diversification rate ($\lambda - \mu$)	0.23 (0.11–0.35)	0.23 (0.10–0.36)	0.14 (0.03–0.25)
Turnover ($\mu \div \lambda$)	0.26 (0.00–0.64)	0.29 (0.00–0.66)	0.71 (0.49–0.95)
Sampling proportion ($\psi \div (\psi + \mu)$)	NA	NA	0.31 (0.13–0.50)
	Concatenation		
Molecular clock rate ($\times 10^{-4}$)	8	8	7.34 (5.84–8.60)
Morphological clock rate	NA	0.09 (0.07–0.11)	0.05 (0.04–0.06)
Diversification rate ($\lambda - \mu$)	0.19 (0.10–0.28)	0.19 (0.10–0.29)	0.12 (0.03–0.21)
Turnover ($\mu \div \lambda$)	0.20 (0.00–0.52)	0.20 (0.00–0.51)	0.62 (0.30–0.88)
Sampling proportion ($\psi \div (\psi + \mu)$)	NA	NA	0.40 (0.18–0.67)

Values in brackets are 95% highest posterior densities.

Clock rates are in units of per-site or per-character per million years.

Diversification rate is in units of per million years.

Mean $N_e g$ refers to the mean of the effective population size N_e distribution, which is scaled by generation time g .

CLADE AGES AND UNCERTAINTY

For all clades in the FBD-MSc MCC tree with at least 0.5% support, the divergence time for the root node of that clade according to the FBD-MSc was younger than when estimated using FBD-concatenation (Fig. 4.6). While the HPD intervals of the two estimates often overlapped substantially, those for the A node (the MRCA of extant sampled *Canis*, *Cuon* and *Lycaon*) and the D node (nested within the A node and excluding *Canis mesomelas* and *C. adustus*) did not, and the FBD-concatenation estimates of those species divergence times were about 2MY older than FBD-MSc.

THE TEMPO OF CANINAE EVOLUTION

If species divergence times are always overestimated using concatenation, even when using fossil data and an FBD prior to calibrate the species trees, this is likely to affect macroevolutionary analyses. As an example, we present LTT curves of Caninae diversification estimated using FBD-MSc and FBD-concatenation (Fig. 4.7).

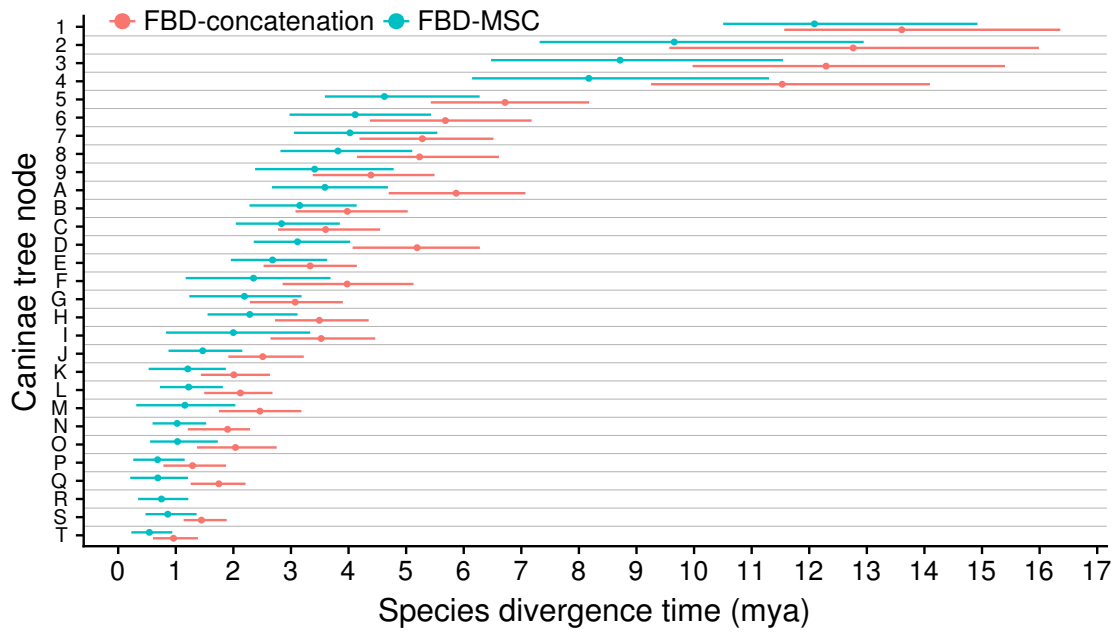


Figure 4.6: Speciation times estimated by fossilized birth-death with multispecies coalescent (FBD-MSC) and with concatenation (FBD-concatenation) models. FBD-MSC node ages (solid circles) and 95% highest posterior density (HPD) intervals (lines) are estimated from samples where that clade is present. FBD-concatenation ages and intervals are also conditioned on clade presence. Node labels correspond to those in Figure 4.2 and 4.4.

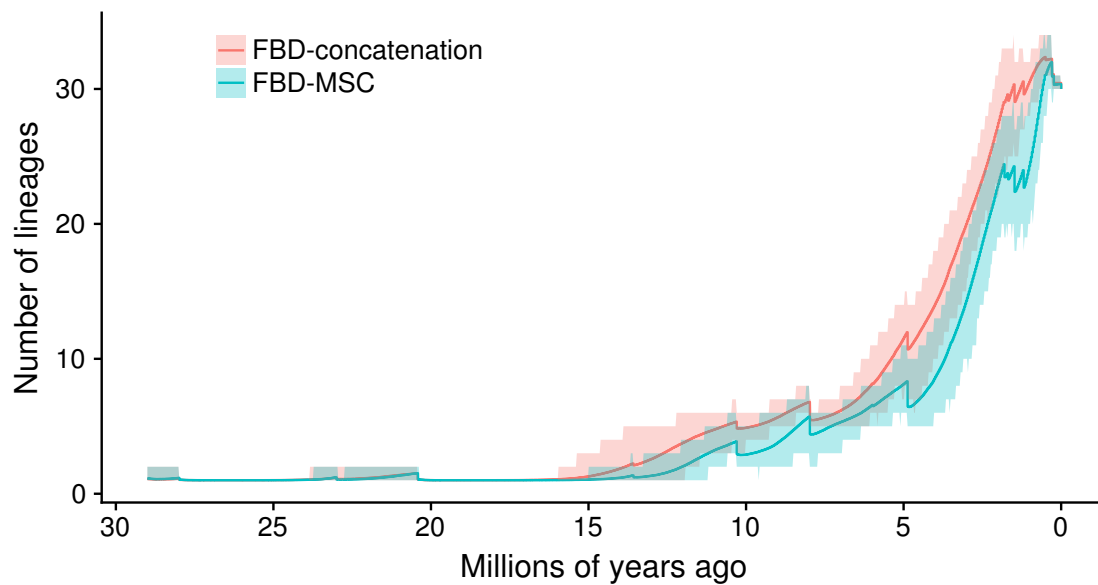


Figure 4.7: Lineages-through-time (LTT) plot of Caninae diversification. Mean estimates (solid lines) of LTT are calculated for 1,001 evenly spaced time steps spanning 0 to 1, and include extant, fossil and ancestral taxa, and sampled ancestors (which are both fossil and ancestral). 95% highest posterior density (HPD) intervals were also calculated for each step, and are shown as translucent ribbons.

For both methods the LTT curves are convex, as expected for a birth-death model of evolution with good taxon sampling (Stadler 2008). However the diversification leading to extant Caninae occurs earlier for the FBD-concatenation LTT curve compared to the FBD-MSC curve. The FBD-concatenation estimate also suggests a diversification slowdown during the

last \approx 2 million years, which is not suggested by the FBD-MSc curve. Diversification slowdown is a predicted spurious effect of concatenation (see Chapter 2).

SUPPORT FOR SPECIFIC CLADES

We considered clade support contradictory between analyses if that clade was highly supported ($> 95\%$) in any analysis and unsupported ($< 5\%$) in any other analysis. Only the R clade met this criterion, which is the clade that unites *Cuon* and *Lycaon* (Table 4.3).

The R clade was highly supported by the MSc regardless of whether a fixed clock or FBD was used to calibrate the species tree. To understand whether this support was driven by coalescent processes alone or by interactions with morphological data, we reran our fixed clock analyses with only molecular data. Without the morphological data there was no support for this clade even when using the MSc, suggesting that unmodeled processes such as selection for convergent morphological evolution might be increasing support for this clade.

4.4 DISCUSSION

CONCATENATED LIKELIHOOD METHODS ARE INACCURATE

Several recent studies have demonstrated that methods which use phylogenetic likelihood to estimate species trees from concatenated loci – “concatenated likelihood” for short – are inaccurate under realistic conditions. These studies have been based on simulation and analytical results, and have covered both maximum likelihood (ML) and Bayesian concatenation.

Mendes and Hahn 2016 showed that ML concatenation is systematically biased when estimating the lengths of particular branches on an asymmetric species tree. This is due to substitutions produced by ILS (SPILS), which are artificial substitutions on discordant species tree branches. Mendes and Hahn 2017 went on to show that SPILS is also responsible for the statistical inconsistency of ML concatenation when estimating species tree topologies, even

Table 4.3: Posterior probabilities of clades.

Clade	MSC			Concatenation		
	MO	FC	FBD	MO	FC	FBD
2	61%	68%	97%	8%	11%	31%
3	77%	82%	95%	69%	77%	92%
4	33%	36%	47%	41%	43%	44%
5	100%	100%	100%	100%	100%	100%
6	100%	100%	100%	100%	100%	100%
7	100%	100%	100%	100%	100%	100%
8	59%	59%	58%	72%	72%	74%
9	100%	100%	99%	100%	100%	100%
A	100%	100%	100%	100%	100%	100%
B	100%	100%	100%	100%	100%	100%
C	100%	100%	100%	100%	100%	100%
D	39%	59%	88%	100%	100%	100%
E	89%	87%	90%	100%	100%	100%
F	100%	100%	100%	100%	100%	100%
G	82%	82%	80%	100%	100%	100%
H	93%	100%	100%	100%	100%	100%
I	100%	100%	100%	100%	100%	100%
J	100%	100%	100%	100%	100%	100%
K	100%	100%	100%	100%	100%	100%
L	50%	47%	51%	100%	100%	100%
M	98%	98%	98%	100%	100%	100%
N	40%	40%	43%	1%	1%	1%
O	100%	100%	100%	100%	100%	100%
P	91%	90%	90%	12%	10%	13%
Q	86%	87%	86%	8%	7%	7%
R	0%	100%	100%	0%	0%	0%
S	16%	73%	46%	0%	0%	1%
T	100%	100%	100%	100%	100%	100%

Clades correspond to node labels used in Figure 4.2 and 4.4.

Probabilities were estimated using a fixed clock and molecular only data (MO), a fixed clock with molecular and extant morphological data (FC), or a fossilized birth-death process with molecular, extant morphological and fossil data (FBD).

outside of the so-called “anomaly zone” of short branch lengths where the most probable gene tree topology is discordant with the species tree.

Other studies have shown that Bayesian concatenation can be grossly inaccurate when estimating species trees. Bayesian concatenation can overestimate the lengths of tip branches by as much as 350%, and is less accurate than Bayesian MSC using the same number of loci (see Chapter 1). Bayesian concatenation is also less accurate at estimating the lengths of internal branches, and reports overly precise credible intervals and support values which can exclude

the true values and topologies a majority of the time (see Chapter 2).

We have built on previous results by studying the effect of concatenation on an empirical data set of Caninae. Using posterior predictive simulations, we have shown that the observed differences in species tree branch lengths between the MSC and concatenation are expected and caused by a failure to account for coalescent processes. Consistent with previous studies, tip branch lengths were always overestimated, and internal branch lengths were sometimes inaccurate in either direction (Figs. 4.2,4.3).

FBD-MSC RESULTS ARE MORE PLAUSIBLE

Researchers may wonder if the known problems of concatenation are relevant to dated trees inferred using an FBD process. Our study showed that for Caninae, dated species trees inferred using a fixed clock are very similar to dated species trees inferred using an FBD process. We further demonstrated that the differences between MSC and concatenation estimates made under a birth-death process without fossil data are very similar to those made under a FBD process with fossil data (Fig. 4.5).

Considering coalescent theory and the totality of our results, the FBD-MSC results are more plausible than the FBD-concatenation results. The posterior predictive simulations show that the observed differences in branch lengths between the MSC and concatenation are expected due to a failure to account for coalescent processes.

This has important implications for downstream analyses, as seen in the LTT plots (Fig. 4.7) where the FBD-concatenation LTT curve suggests a slowdown in Caninae diversification during the past ≈ 2 million years. In contrast, the FBD-MSC LTT curve shows a burst of diversification in the same time frame.

In this study the estimated clock rate of Caninae using the FBD was consistent with the rate inferred by Hugall *et al.* 2007. Despite this consistency, FBD models are still necessary to account for the correct amount of uncertainty in clock rates, and because the *a priori* clock

rate will not always be accurate. If we had studied a different mammalian clade, it would not necessarily have a mean substitution rate consistent with Hugall *et al.* 2007.

Some unexplored possibilities are that FBD-concatenation would approach FBD-MSM given a morphological matrix covering more taxa and/or when using a relaxed clock. These are hypothetically interesting questions but in practice morphological data sets are usually quite limited in the number of taxa and characters. Concatenation with a relaxed clock is much slower than StarBEAST2 with a strict clock, without any evidence of improved error rates (see Chapter 2).

MORPHOLOGICAL AND MOLECULAR DISCORDANCE

We observed that the inclusion or omission of morphological data completely changes the support of the *Lycaon+Cuon* clade from 100% to 0% respectively when using MSM models (Table 4.3). Support for this clade is ubiquitous in morphological phylogenetic studies of Caninae (Tedford *et al.* 2009; Prevosti 2010) and probably is due to their specialized dentitions. A previous study of Caninae which combined morphological characters and mitochondrial sequence alignments found that support for this clade came only from the morphological data, and proposed that the responsible characters are likely convergent due to the hypercarnivory of these two species (Zrzavý and Řičánková 2004).

Molecular phylogeneticists should be aware of the potential for morphological model violations when conducting total-evidence studies, and be appropriately cautious when interpreting results. A potential avenue for future research is the development of improved models of morphological evolution, which allow for convergence across many characters at once due to selection. New models could either rule in or out support for *Lycaon+Cuon* by ascribing their similar morphology to convergent evolution. Alternatively, support for this putative clade could be further scrutinized through expanded sampling of fossil representatives of these lineages.

The molecular signal could also be potentially misleading due to unmodeled processes, for example introgression. This could be addressed by integrating the FBD with the multispecies network coalescent, which unlike the MSC does allow for introgression and hybridization (Wen and Nakhleh 2017, see also Chapter 3).

INTEGRATIVE MODELS ARE THE FUTURE

The development and implementation of the integrative FBD-MSc model demonstrates how integrative models are made possible within a Bayesian framework. Unlike previous Bayesian implementations of the MSC which are ultrametric and hence limited to contemporary sources of data, using the FBD-MSc we can incorporate morphological and timing information from excavated fossils. The FBD-MSc is a first step, and the future will see further development of integrative models in theory, and the development and use of new implementations in practice.

4.5 ACKNOWLEDGMENTS

This research was funded a Royal Society of New Zealand Marsden award granted to AJD, DW, NJM, TGV and TS (16-UOA-277). HAO was supported by an Australian Laureate Fellowship awarded to Craig Moritz by the Australian Research Council (FL110100104). This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We thank Craig Moritz for his advice on preparing the manuscript, and the late Colin Groves for his insight into Caninae fossil record.

5

Conclusion

The results and new methods described in my thesis make an original contribution to the field of phylogenetic inference. By evaluating the multispecies coalescent (MSC) model, we have informed researchers of the circumstances where the MSC should be used instead of concatenation. It is clear that estimates of branch lengths and divergence times made within eukaryotic genera or families should be based on fully Bayesian methods like StarBEAST2 (Chapter 1,2,4).

Because we developed StarBEAST2 and accelerated the MSC, it is now practical to use about twice as many loci as *BEAST when inferring species trees, leading to more accurate inferences of taxonomic relationships and divergence times (Chapter 2). By extending the MSC to support fossil data, species trees can be dated using a process-based model of fossilisation rather than informally derived node priors, also benefiting accuracy (Chapter 4).

Some research questions were previously unanswerable using the MSC. The evolution of molecular clock rates (Thorne *et al.*, 1998) has been previously studied using mitochondrial DNA (e.g. Nabholz *et al.* 2009), a single nuclear locus (e.g. Hugall *et al.* 2007), and concatenation (e.g. Dornburg *et al.* 2012). However concatenation has systematic biases leading to apparent variation in substitution rates where none exists (Mendes and Hahn, 2016, 2017). By developing and implementing species tree relaxed clocks in StarBEAST2, we have enabled the accurate and precise reconstruction of past and present molecular clock rates from multiple nuclear loci (Chapter 2).

By developing a fully Bayesian implementation of the multispecies-network coalescent (MSNC), we have enabled the discovery and characterisation of introgression and hybrid species (Chapter 3), the prevalence of which is becoming increasingly obvious from genomic data. Because our method “SpeciesNetwork” is fully Bayesian, it can be used to infer the timing, direction, and proportion of introgression in a joint analysis.

Like any study, the results of my thesis raise additional questions, and suggest pathways for future research. To conclude my thesis, I will discuss two of what I believe are the most promising directions.

5.1 BEYOND MARKOV CHAIN MONTE CARLO

Starting about ten years ago, most of the increase in computer power has come from increasing the number of CPU and GPU cores (Geer, 2005). To take advantage of more than one core, algorithms must be able to execute operations simultaneously on different cores. The result of one of these operations cannot depend on the result of another, otherwise they cannot run in parallel.

StarBEAST2 and SpeciesNetwork are both Markov chain Monte Carlo (MCMC) algorithms. MCMC is inherently serial because the results of one step of the chain is needed be-

fore the next step can execute. Concatenation MCMC algorithms can still run effectively in parallel because the large concatenated MSA can be split into several segments, and the phylogenetic likelihood of each segment is computed on a separate core.

However for most MCMC steps when running StarBEAST2, only one gene tree will be changed, so the phylogenetic and MSC/MSNC likelihood only needs to be computed for a single MSA and gene tree. These methods are typically used with short nucleotide sequences, making the phylogenetic likelihood calculation even faster.

The very short time taken by each MCMC step leaves little room to accelerate those calculations, and the overhead of parallelization will typically be larger than the performance gain, resulting in worse performance than running StarBEAST2 on a single core.

Besides MCMC, there are other types of algorithms for Bayesian inference that naturally run in parallel on many cores. These include sequential Monte Carlo (SMC), also known as particle filtering, which has been applied to phylogenetic inference with promising results (Bouchard-Côté *et al.*, 2012). Even if SMC turns out to be less efficient than MCMC, because it is naturally parallel it can still be faster if run on many cores. An SMC implementation of the MSC that could run on a supercomputer would be useful to many researchers.

Another algorithm for Bayesian inference is variational Bayes (VB), which is many times faster than MCMC (Kucukelbir *et al.*, 2015). VB works with continuous parameters as opposed to phylogenetic trees which are discrete, but one possibility is to pursue a hybrid approach using MCMC for species and gene tree topologies, and VB for branch lengths.

5.2 MODELLING MORPHOLOGICAL EVOLUTION

In Chapter 4 we found disagreement between morphological characters and molecular sequences in a dataset of the dog and fox subfamily Caninae. The morphological data supported a clade containing *Cuon alpinus* and *Lycaon pictus*, which are both hypercarnivores.

When using both sources of data with our combined fossilized birth-death (FBD) MSC model, the species tree reflected the signal in the morphological characters. However this signal is likely misleading, as many of the identical morphological characters of *Cuon* and *Lyacon* are probably convergent due to selection for hypercarnivory, rather than identical through common descent (Zrzavý and Řičánková, 2004).

The Mkv model of morphological evolution assumes that different characters are evolving independently on a common tree (Lewis, 2001), but in cases like Caninae this assumption appears to be violated. Future morphological models that relax this assumption would more accurately gauge the uncertainty in morphological data sets.

Another idea is to couple character evolution with the ecological niche of each species. The state frequencies of a discrete character could be estimated separately for each ecological niche, so for a given niche particular states would be more favoured. Such a model would naturally fit convergent morphological evolution. Unlike common substitution models it would violate the assumption of state frequency stationarity, but non-stationary models have been developed for nucleotide data (Galtier and Gouy, 1998; Kaehler, 2017).

5.3 FINAL REMARKS

New integrative models like the FBD-MS, and new probabilistic methods like StarBEAST2 and SpeciesNetwork, and are a step towards full consideration of the complexity present in biological systems, and of the uncertainty inherent to most scientific study. With more capable implementations of accurate models, future phylogenetic research will be both more accurate and more precise, without the false precision of estimates made using bad approximations. I hope to continue developing better phylogenetic methods and models, and to make new discoveries with StarBEAST2, SpeciesNetwork, and future methods.

A

Algorithms used in Chapter 3

A.0.1 REPRESENTATION OF PHYLOGENETIC NETWORKS

Species networks are outputted in extended Newick format (Cardona *et al.*, 2008), which is also used in the software PhyloNet (Than *et al.*, 2008).

For example, the species network in Figure 3.1a is written as

```
((A:0.02,(B:0.01)#H1[&gamma=0.3]:0.01)S1:0.03,  
(#H1:0.02,C:0.03)S2:0.02)R:0.03;
```

where the hash sign indicates a reticulation node, and the inheritance probability is in the brackets as metadata. Such extended Newick string can be read into IcyTree (Vaughan, 2017)¹ and be displayed nicely.

¹<https://icytree.org/> — accessed 15th December 2017

A.0.2 NUMBERING AND LABELING SUBNETWORKS ACROSS A SAMPLE

We describe an algorithm by pseudocode to enumerate all unique subnetwork topologies within a sample of phylogenetic networks. Apart from subnetwork topologies consisting of a single node (i.e. leaf nodes), each topology label has a corresponding set of child subnetwork topology numbers. The algorithm works by recursively constructing the mapping of parent to child subnetwork topology numbers, beginning at the root or origin node of each phylogenetic network.

Initialize the counter i to 0

Initialize the (node label set to node label) map m

For each taxon t :

 Assign i as the label of t

 Increment i

For each phylogenetic network s :

 Begin Recursion from the oldest node of s

Recursion:

 Input: A network node n

 Output: A label l to identify the subnetwork topology of n

 If n is a leaf node:

 Get the label l of the taxon t of n

 Else:

 Initialize the node label set d

 For each child node n_c of n :

```

    Get  $l_c$  by continuing Recursion from  $n_c$ 

    Add  $l_c$  to  $d$ 

    If  $d$  is in  $m$ :

        Get the label  $l$  of  $d$ 

    Else:

        Set  $l$  to the value of  $i$ 

        Link  $d$  to  $l$  in  $m$ 

        Increment  $i$ 

    Return  $l$ 

```

A.0.3 PROPOSING EMBEDDINGS PROPORTIONAL TO THEIR LIKELIHOODS

We describe an algorithm by pseudocode to propose compatible gene tree embeddings, given a species network and a set of gene trees, in proportion to their embedding likelihoods. The algorithm works by stochastically constructing an embedding during a depth-first search of a gene tree. When a gene tree lineage traverses a bifurcation node, there is a set of compatible embedding histories (for the subtree defined by the gene tree lineage) where the lineage descends through the left child branch of the bifurcation node, and another set for the right child branch. A left or right embedding is chosen at random weighted by the sum total of embedding likelihoods for each child branch of the bifurcation node, to ensure that embeddings are proposed in proportion to their likelihoods.

The likelihood for the proposed embedding is also computed during the depth-first search; when a gene tree lineage traverses a reticulation node, its likelihood is multiplied by the γ_h or the (alternative) $1 - \gamma_h$ probability. When a coalescent event occurs, the likelihoods of the left and right subtrees are multiplied. Because embeddings are proposed in proportion to their

likelihoods, the MCMC proposal probability is the embedding likelihood normalized by the sum total of compatible embedding likelihoods.

Given the species network s :

For each gene tree g :

Get the root gene tree node gtn_r from g

Get the root species network branch snb_r from s

Try to get e , l , and t by Recursion from gtn_r and snb_r

If there is no compatible embedding:

Reject the proposal

Else:

Propose e as the new embedding

Multiply the proposal probability by $(l \div t)$

Recursion:

Input 1: A gene tree node gtn

Input 2: A species network branch snb

Output 1: An embedding e

Output 2: Its likelihood l

Output 3: The total likelihood t

If gtn traverses through the tipward node of snb :

For each child branch snb_c of snb :

If there is any compatible embedding of gtn through snb_c :

Get e_c , l_c , and t_c by Recursion from gtn and snb_c

Add the traversal of gtn through snb_c to e_c

If the tipward node of snb is a reticulation:

Multiply l_c by γ_h or $1 - \gamma_h$

Multiply t_c by γ_h or $1 - \gamma_h$

Pick one snb_c at random weighted by t_c

Set the embedding e to the value of e_c for the chosen snb_c

Set the likelihood l to the value of l_c for the chosen snb_c

Calculate the total likelihood t as the sum of all t_c

Else:

If gtn is a leaf:

Initialize an embedding e

Initialize the likelihood l to 1

Initialize the total likelihood t to 1

Else:

For each child node gtn_c of gtn :

Get e_c , l_c , and t_c by Recursion from gtn_c and snb

Construct the embedding e by merging both e_c

Calculate the likelihood l as the product of both l_c

Calculate the total likelihood t as the product of both t_c

Return e , l , and t

B

Supplementary figures for Chapter 4

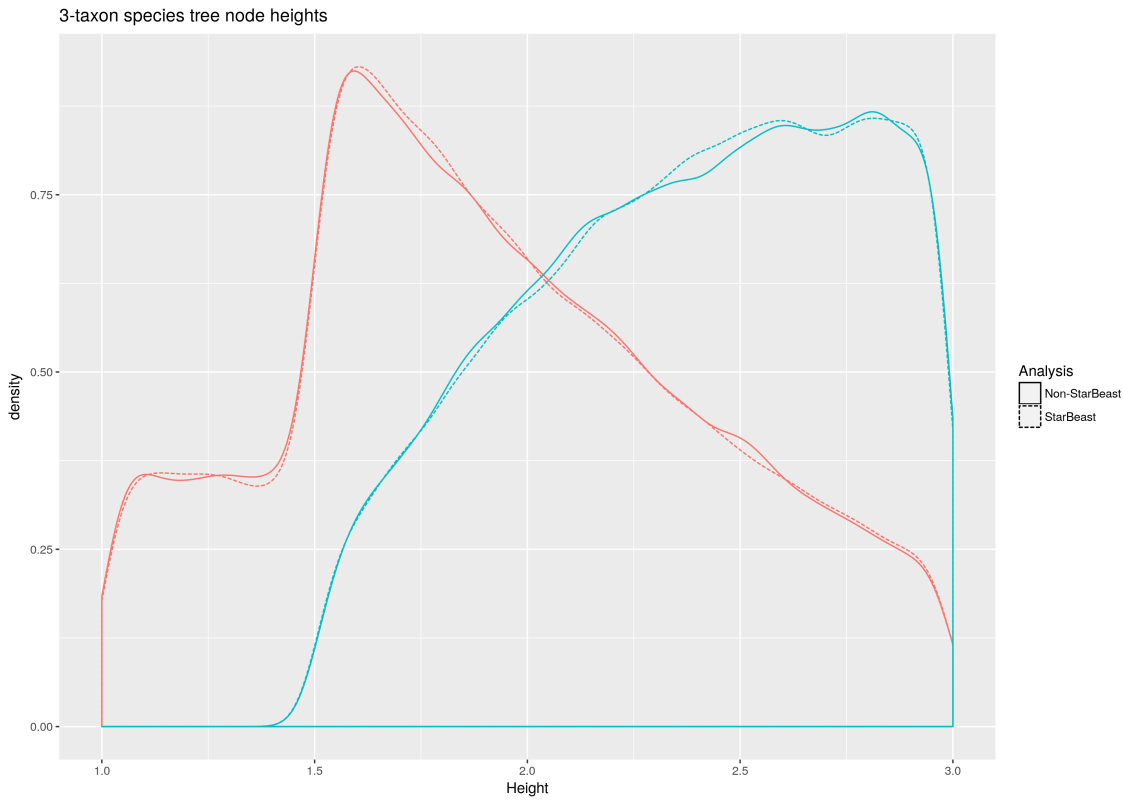


Figure B.1: Marginal distribution of species divergence times in three taxon sampled ancestor species trees. Heights are plotted for each divergence time; the first divergence time is red and the second is cyan. Solid lines are for trees sampled from the prior without StarBEAST2, dashed lines are for species trees sampled from the prior jointly with an embedded gene tree using StarBEAST2.

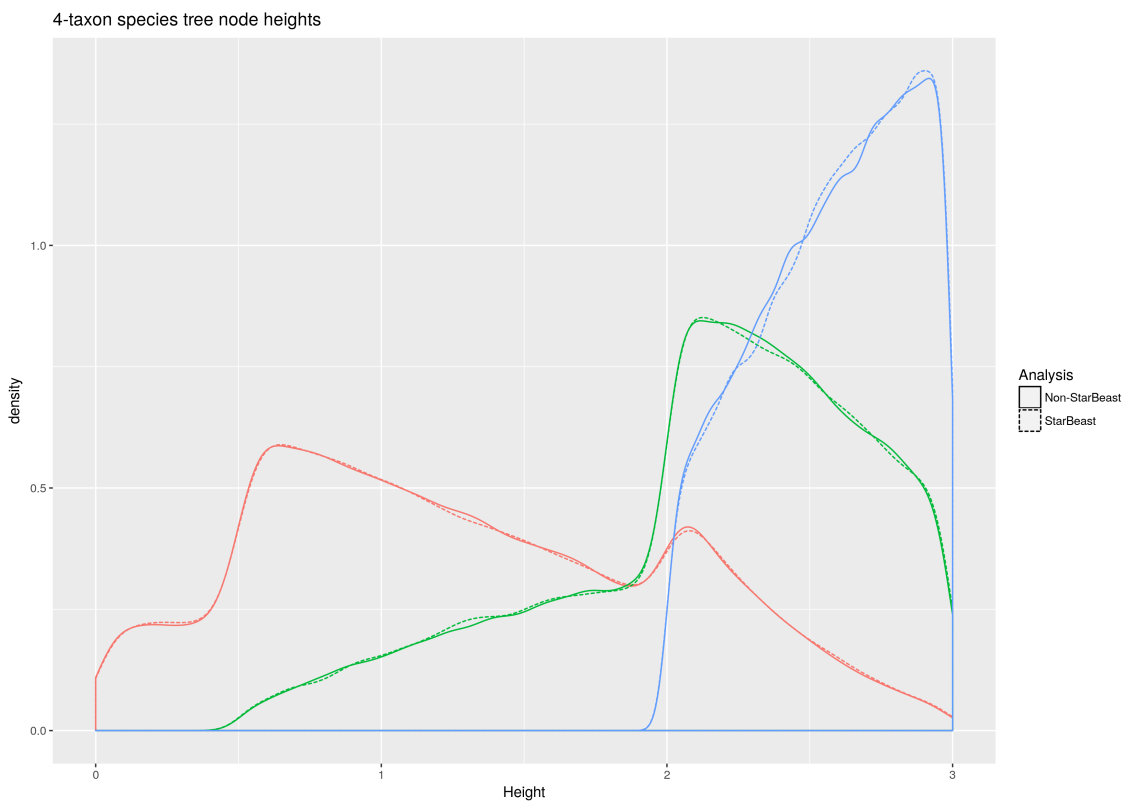


Figure B.2: Marginal distribution of species divergence times in four taxon sampled ancestor species trees. Heights are plotted for each divergence time; the first divergence time is red, then green, then blue. Solid lines are for trees sampled from the prior without StarBEAST2, dashed lines are for species trees sampled from the prior jointly with an embedded gene tree using StarBEAST2.

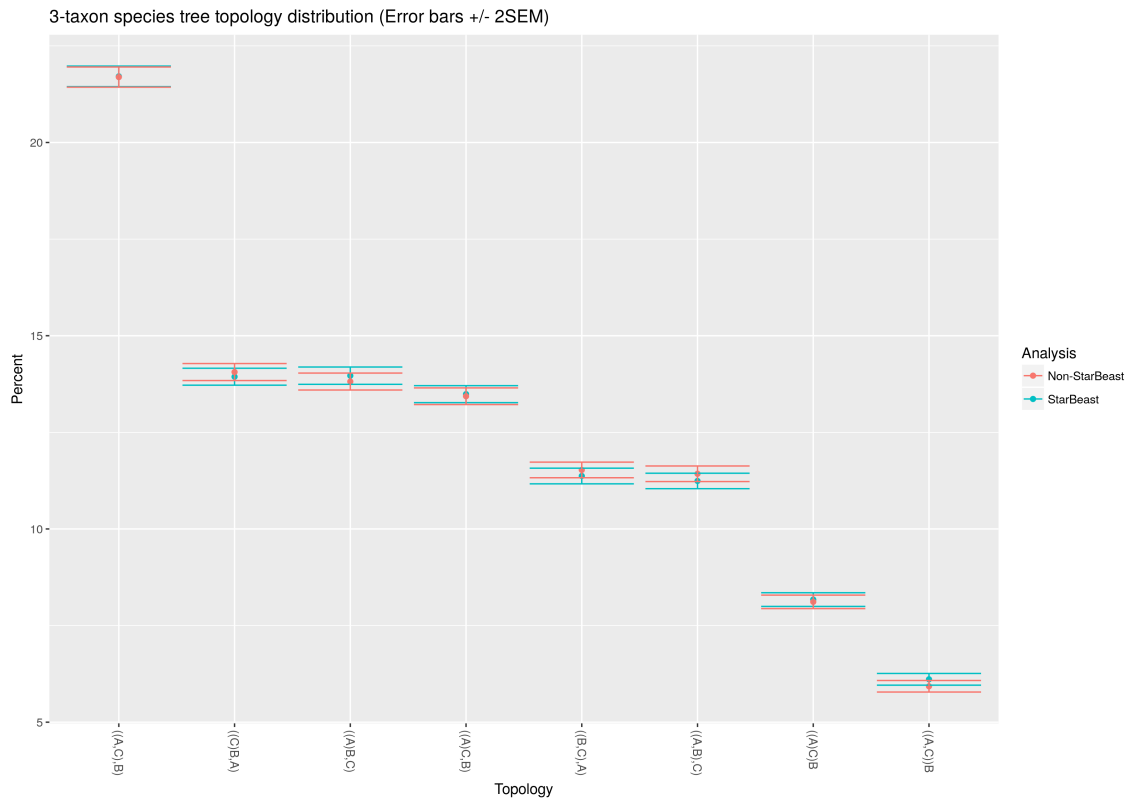


Figure B.3: Marginal distribution of three taxon sampled ancestor species tree topologies. Red points and error bars are for trees sampled from the prior without StarBEAST2, cyan is for species trees sampled from the prior jointly with an embedded gene tree using StarBEAST2.

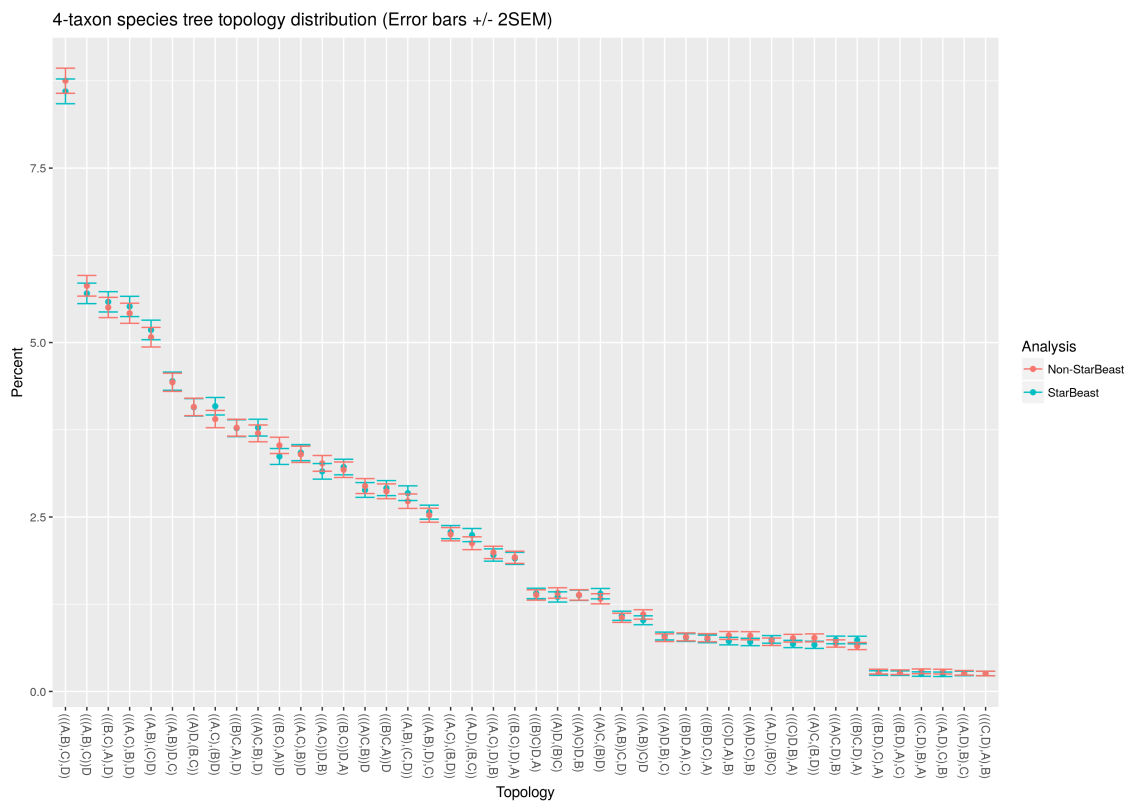


Figure B.4: Marginal distribution of four taxon sampled ancestor species tree topologies. Red points and error bars are for trees sampled from the prior without StarBEAST2, cyan is for species trees sampled from the prior jointly with an embedded gene tree using StarBEAST2.

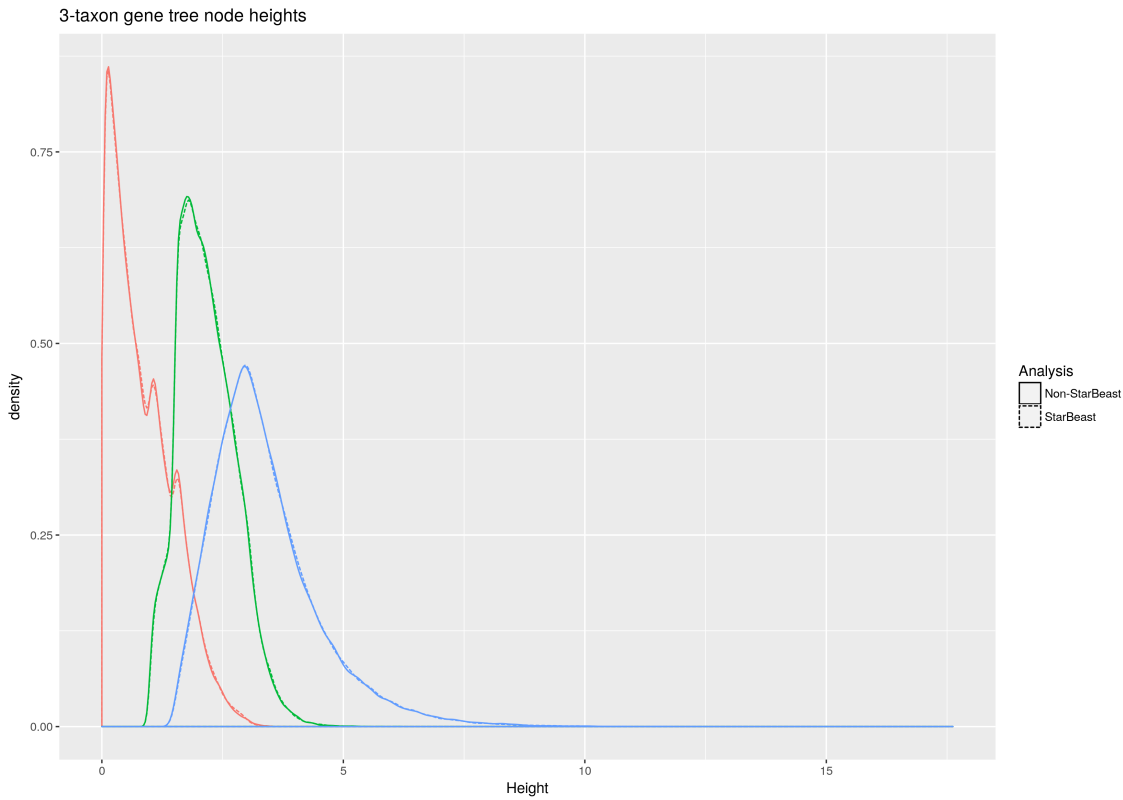


Figure B.5: Marginal distribution of gene coalescence times in three taxon/four sample gene trees. Heights are plotted for each coalescence time; the first coalescence time is red, then green, then blue. Solid lines are for trees simulated within a sampled ancestor tree sampled from the prior without StarBEAST2, dashed lines are for gene trees sampled from the prior jointly and embedded within a sampled ancestor species tree using StarBEAST2.

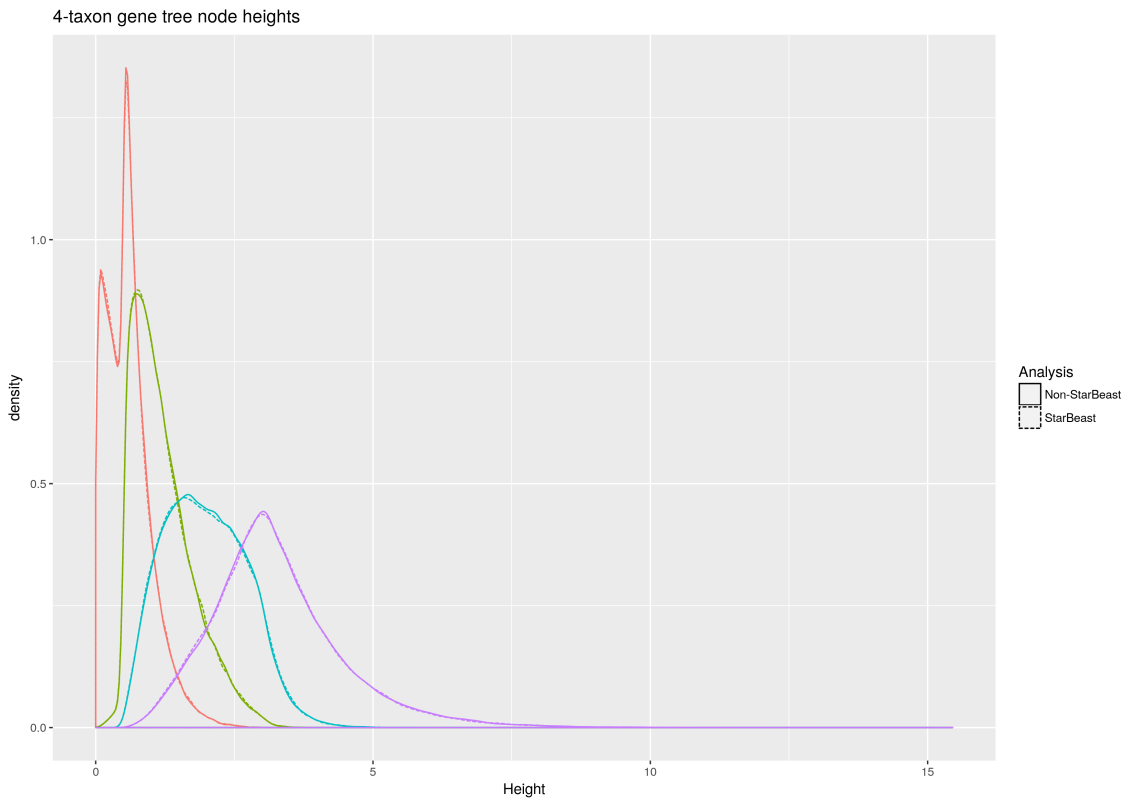


Figure B.6: Marginal distribution of gene coalescence times in four taxon/five sample gene trees. Heights are plotted for each coalescence time; the first coalescence time is red, then green, then cyan, then purple. Solid lines are for trees simulated within a sampled ancestor tree sampled from the prior without StarBEAST2, dashed lines are for gene trees sampled from the prior jointly and embedded within a sampled ancestor species tree using StarBEAST2.

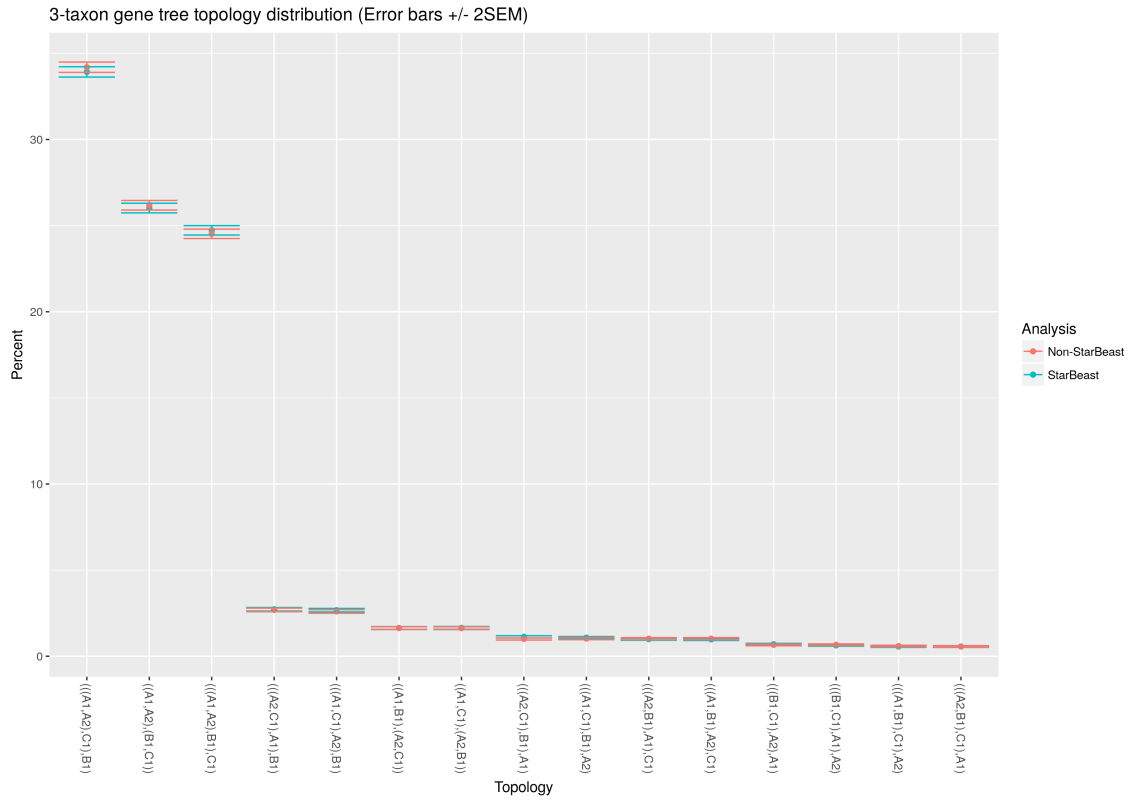


Figure B.7: Marginal distribution of three taxon/four sample gene tree topologies. Red points and error bars are for trees simulated within a sampled ancestor tree sampled from the prior without StarBEAST2, cyan is for gene trees sampled from the prior jointly and embedded within a sampled ancestor species tree using StarBEAST2.

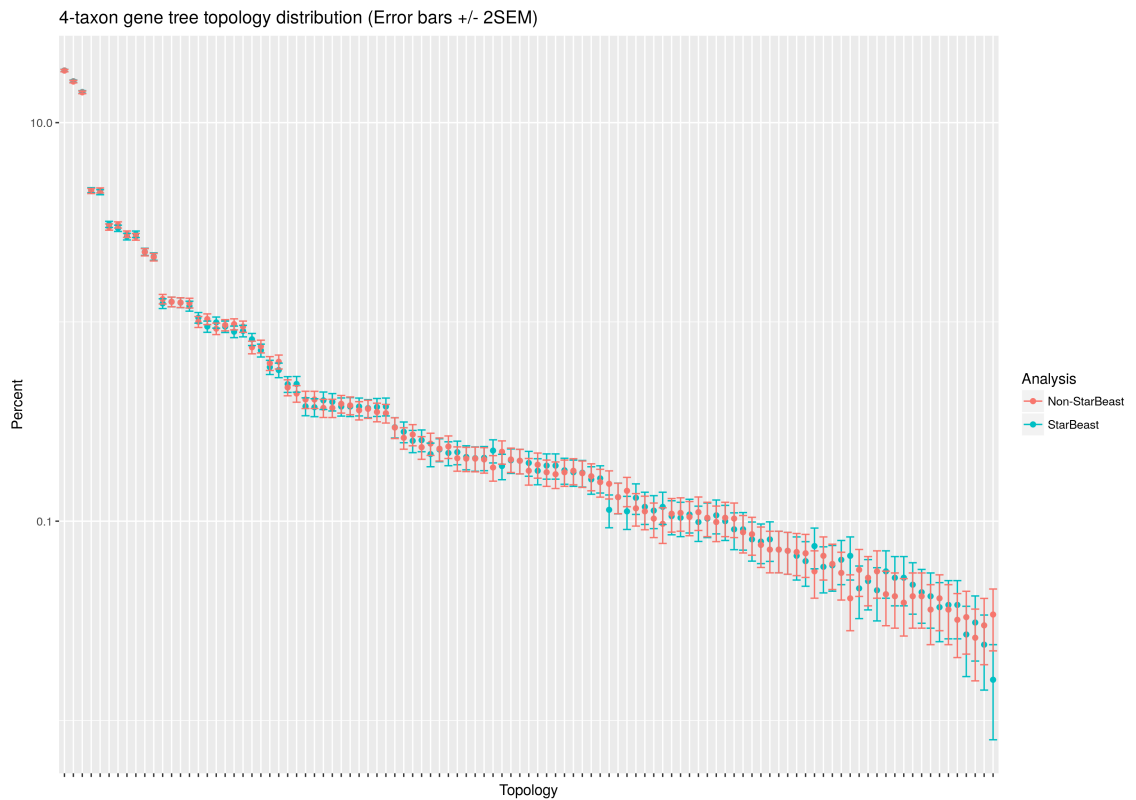


Figure B.8: Marginal distribution of four taxon/five sample gene tree topologies. Red points and error bars are for trees simulated within a sampled ancestor tree sampled from the prior without StarBEAST2, cyan is for gene trees sampled from the prior jointly and embedded within a sampled ancestor species tree using StarBEAST2. The topology newick string is omitted from this figure for clarity, but topologies are plotted in order of their observed frequencies.

Bibliography

- Aberer, A. J., Kobert, K., and Stamatakis, A. 2014. ExaBayes: Massively parallel Bayesian tree inference for the whole-genome era. *Molecular Biology and Evolution*, 31(10): 2553–2556.
- Afonso Silva, A. C., Santos, N., Ogilvie, H. A., and Moritz, C. 2017. Validation and description of two new north-western Australian Rainbow skinks with multispecies coalescent methods and morphology. *PeerJ*, 5: e3724.
- Albrecht, B., Scornavacca, C., Cenci, A., and Huson, D. H. 2012. Fast computation of minimum hybridization networks. *Bioinformatics*, 28(2): 191–197.
- Andrieu, C. and Thoms, J. 2008. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4): 343–373.
- Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P., and Slowinski, J. B. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics*, 33: 707–740.
- Baer, C. F., Miyamoto, M. M., and Denver, D. R. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, 8(8): 619–631.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, 3(10): e3376.

- Bardeleben, C., Moore, R. L., and Wayne, R. K. 2005a. Isolation and molecular evolution of the selenocysteine tRNA (Cf TRSP) and RNase P RNA (Cf RPPH1) genes in the dog family, Canidae. *Molecular Biology and Evolution*, 22(2): 347–359.
- Bardeleben, C., Moore, R. L., and Wayne, R. K. 2005b. A molecular phylogeny of the Canidae based on six nuclear loci. *Molecular Phylogenetics and Evolution*, 37(3): 815–831.
- Barr, C. M., Neiman, M., and Taylor, D. R. 2005. Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytologist*, 168(1): 39–50.
- Barrow, L. N., Ralicki, H. F., Emme, S. A., and Lemmon, E. M. 2014. Species tree estimation of North American chorus frogs (Hylidae: *Pseudacris*) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution*, 75: 78–90.
- Bayzid, M. S. and Warnow, T. 2013. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18): 2277–2284.
- BBC and Warner Bros. 1979. Life on Earth. Presented by David Attenborough.
- Beaulieu, J. M., O’Meara, B. C., Crane, P., and Donoghue, M. J. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Systematic Biology*, 64(5): 869–878.
- Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.
- Berv, J. S. and Prum, R. O. 2014. A comprehensive multilocus phylogeny of the Neotropical cotingas (Cotingidae, Aves) with a comparative evolutionary analysis of breeding system and plumage dimorphism and a revised phylogenetic classification. *Molecular Phylogenetics and Evolution*, 81: 120–136.

- Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., and Good, J. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, 13(1): 403.
- Blom, M. P. K., Horner, P., and Moritz, C. 2016. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1832).
- Blum, M. G. B., François, O., and Janson, S. 2006. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, 16(4): 2195–2214.
- Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. 2012. Phylogenetic inference via sequential Monte Carlo. *Systematic Biology*, 61(4): 579–593.
- Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097): 957–960.
- Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, 10(4): e1003537.
- Bromham, L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1577): 2503–2513.
- Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.-R., De Bonis, L., Coppens, Y., Dejax, J., Denys, C., Durringer, P., Eisenmann, V., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau,

- F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Pelaez Campomanes, P., Ponce De Leon, M., Rage, J.-C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A., and Zollikofer, C. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*, 418(6894): 145–51.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8): 1917–1932.
- Cadotte, M. W., Cardinale, B. J., and Oakley, T. H. 2008. Evolutionary history and the effect of biodiversity on plant productivity. *Proceedings of the National Academy of Sciences*, 105(44): 17012–17017.
- Caldwell, J. P. 1987. Demography and life history of two species of chorus frogs (Anura: Hyliidae) in South Carolina. *Copeia*, 1987(1): 114–127.
- Callaway, E. 2015. DNA clock proves tough to set. *Nature*, 519(7542): 139–40.
- Camargo, A., Avila, L. J., Morando, M., and Sites, J. W. 2012. Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the *Liolaemus darwini* group (Squamata, Liolaemidae). *Systematic Biology*, 61(2): 272–288.
- Cardona, G., Rosselló, F., and Valiente, G. 2008. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, 9(1): 532.
- Cavender, J. A. and Felsenstein, J. 1987. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4(1): 57–71.
- Chifman, J. and Kubatko, L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23): 3317–3324.

- Chung, Y. and Ané, C. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: Simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology*, 60(3): 261–275.
- Dalquen, D. A., Zhu, T., and Yang, Z. 2017. Maximum likelihood implementation of an Isolation-with-Migration model for three species. *Systematic Biology*, 66(3): 379–398.
- Darwin, C. 1839. *Journal of researches into the geology and natural history of the various countries visited by H.M.S. Beagle*. Henry Colburn, London.
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7): 499–510.
- de Magalhães, R. F., Lemes, P., Camargo, A., Oliveira, U., Brandão, R. A., Thomassen, H., Garcia, P. C. d. A., Leite, F. S. F., and Santos, F. R. 2017. Evolutionarily significant units of the critically endangered leaf frog *Pithecopus ayeaye* (Anura, Phyllomedusidae) are not effectively preserved by the Brazilian protected areas network. *Ecology and Evolution*, 7(21): 8812–8828.
- de Queiroz, A. 2014. *The Monkey's Voyage: How Improbable Journeys Shaped the History of Life*. Basic Books, New York.
- DeGiorgio, M. and Degnan, J. H. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular Biology and Evolution*, 27(3): 552–569.
- Degnan, J. H. and Rosenberg, N. A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5): e68.
- Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6): 332–340.

- Donoghue, P. C. and Benton, M. J. 2007. Rocks and clocks: calibrating the tree of life using fossils and molecules. *Trends in Ecology & Evolution*, 22(8): 424–431.
- Dornburg, A., Brandley, M. C., McGowen, M. R., and Near, T. J. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). *Molecular Biology and Evolution*, 29(2): 721–736.
- dos Reis, M., Donoghue, P. C. J., and Yang, Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. *Biology Letters*, 10(1).
- Doyle, J. J. 2011. Phylogenetic perspectives on the origins of nodulation. *Molecular Plant-Microbe Interactions*, 24(11): 1289–1295.
- Drummond, A. J. and Bouckaert, R. R. 2015. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press.
- Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7: 214.
- Drummond, A. J. and Suchard, M. A. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8: 114.
- Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5): e88.
- Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8): 1969–73.
- Eaton, D. A. R. and Ree, R. H. 2013. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, 62(5): 689–706.

- Edwards, S. V., Liu, L., and Pearl, D. K. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14): 5936–5941.
- Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61(5): 717–726.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4): 401–410.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.
- Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Biology*, 20(4): 406–416.
- Galtier, N. and Gouy, M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 15(7): 871–879.
- Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7): 685–695.
- Gatesy, J. and Springer, M. S. 2013. Concatenation versus coalescence versus “concordance”. *Proceedings of the National Academy of Sciences*, 110(13): E1179.
- Gatesy, J. and Springer, M. S. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concordance conundrum. *Molecular Phylogenetics and Evolution*, 80: 231–266.

- Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12): e1003919.
- Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1): 57–73.
- Geer, D. 2005. Chip makers turn to multicore processors. *Computer*, 38(5): 11–13.
- Gernhard, T. 2008. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4): 769–778.
- Geurts, R., Lillo, A., and Bisseling, T. 2012. Exploiting an ancient signalling machinery to enjoy a nitrogen fixing symbiosis. *Current Opinion in Plant Biology*, 15(4): 438–443.
- Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163.
- Giarla, T. C. and Esselstyn, J. A. 2015. The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology*, 64(5): 727–740.
- Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C. P. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution*, 9(3): 585–598.
- Gossmann, T. I., Schmid, M. W., Grossniklaus, U., and Schmid, K. J. 2014. Selection-driven evolution of sex-biased genes is consistent with sexual selection in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 31(3): 574–583.

- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.
- Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2): 160–174.
- Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97–109.
- Hayden, E. C. 2014. The \$1,000 genome. *Nature*, 507(7492): 294–295.
- Heath, T. A., Huelsenbeck, J. P., and Stadler, T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29): E2957–E2966.
- Heled, J. 2013. biopy – a library for phylogenetic exploration. https://figshare.com/articles/biopy_a_Library_for_Phylogenetic_Exploration/761224. Accessed 15th December 2017.
- Heled, J. and Bouckaert, R. R. 2013. Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology*, 13: 221.
- Heled, J. and Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3): 570–580.
- Heled, J., Bryant, D., and Drummond, A. J. 2013. Simulating gene trees under the multi-species coalescent and time-dependent migration. *BMC Evolutionary Biology*, 13: 44.
- Hey, J. 2010. Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, 27(4): 905–920.

- Hey, J. and Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2): 747–760.
- Hey, J. and Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8): 2785–2790.
- Höhna, S. and Drummond, A. J. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. *Systematic Biology*, 61(1): 1–11.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550): 2310–2314.
- Hugall, A. F., Foster, R., Lee, M. S. Y., and Hedin, M. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Systematic Biology*, 56(4): 543–563.
- Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4): 726–736.
- Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M.

- P. C., Prosdocimi, F., Samaniego, J. A., Vargas Velazquez, A. M., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jönsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215): 1320–1331.
- Joly, S., McLenachan, P. A., and Lockhart, P. J. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, 174(2): E54–E70.
- Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, 74(1): 447–467.
- Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York.
- Kaehler, B. D. 2017. Full reconstruction of non-stationary strand-symmetric models on rooted phylogenies. *Journal of Theoretical Biology*, 420(Supplement C): 144–151.
- Kainer, D. and Lanfear, R. 2015. The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution*, 32(6): 1611–1627.
- Kendall, D. G. 1948. On the generalized “birth-and-death” process. *Annals of Mathematical Statistics*, 19(1): 1–15.

- Kingman, J. F. 1982. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248.
- Kirkpatrick, M. and Slatkin, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4): 1171–1181.
- Koepfli, K.-P., Pollinger, J., Godinho, R., Robinson, J., Lea, A., Hendricks, S., Schweizer, R. M., Thalmann, O., Silva, P., Fan, Z., *et al.* 2015. Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. *Current Biology*, 25(16): 2158–2165.
- Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, 58(5): 478–488.
- Kubatko, L. S. and Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1): 17–24.
- Kubatko, L. S., Carstens, B. C., and Knowles, L. L. 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7): 971–973.
- Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. 2015. Automatic variational inference in Stan. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 568–576.
- Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4(2): 167–191.
- Lambert, S. M., Reeder, T. W., and Wiens, J. J. 2015. When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. *Molecular Phylogenetics and Evolution*, 82, Part A: 146–155.

- Lanfear, R., Ho, S. Y. W., Jonathan Davies, T., Moles, A. T., Aarssen, L., Swenson, N. G., Warman, L., Zanne, A. E., and Allen, A. P. 2013. Taller plants have lower rates of molecular evolution. *Nature Communications*, 4: 1879.
- Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoiniski, T. S., Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbühler, K., Pääbo, S., and Vigilant, L. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, 109(39): 15716–15721.
- Lanier, H. C., Huang, H., and Knowles, L. L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. *Molecular Phylogenetics and Evolution*, 70: 112–119.
- Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22): 2910–2911.
- Laver, R. J., Doughty, P., and Oliver, P. M. 2017a. Origins and patterns of endemic diversity in two specialized lizard lineages from the Australian Monsoonal Tropics (*Oedura* spp.). *Journal of Biogeography*. Early View.
- Laver, R. J., Nielsen, S. V., Rosauer, D. F., and Oliver, P. M. 2017b. Trans-biome diversity in Australian grass-specialist lizards (Diplodactylidae: Strophurus). *Molecular Phylogenetics and Evolution*, 115: 62–70.
- Leaché, A. D. and Rannala, B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2): 126–137.

- Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. 2014. The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, 63(1): 17–30.
- Lee, M. S. Y., Soubrier, J., and Edgecombe, G. D. 2013. Rates of phenotypic and genomic evolution during the Cambrian Explosion. *Current Biology*, 23(19): 1889–1895.
- Lemmon, A. R., Emme, S. A., and Lemmon, E. M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5): 727–744.
- Lemmon, E. M. and Lemmon, A. R. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44(1): 99–121.
- Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, 24(12): 2669–2680.
- Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6): 913–925.
- Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., *et al.* 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069): 803–819.
- Linkem, C. W., Minin, V. N., and Leaché, A. D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). *Systematic Biology*, 65(3): 465–477.
- Lipson, M., Loh, P.-R., Sankararaman, S., Patterson, N., Berger, B., and Reich, D. 2015. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLOS Genetics*, 11(11): e1005550.

- Liu, L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21): 2542–2543.
- Liu, L. and Edwards, S. V. 2015. Comment on “Statistical binning enables an accurate coalescent-based estimation of the avian tree”. *Science*, 350(6257): 171.
- Liu, L., Pearl, D. K., Brumfield, R. T., and Edwards, S. V. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution*, 62(8): 2080–2091.
- Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. 2009a. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1): 320–328.
- Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. 2009b. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5): 468–477.
- Liu, L., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10: 302.
- Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. 2015. Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, 1360(1): 36–53.
- Long, J. C. 1991. The genetic structure of admixed populations. *Genetics*, 127(2): 417–428.
- Löytynoja, A. and Goldman, N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences*, 102(30): 10557–10562.
- Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology*, 46(3): 523–536.
- Maddison, W. P., Midford, P. E., and Otto, S. P. 2007. Estimating a binary character’s effect on speciation and extinction. *Systematic Biology*, 56(5): 701–710.

- Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5): 229–237.
- Mallet, J. 2007. Hybrid speciation. *Nature*, 446(7133): 279–283.
- Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2): 111–118.
- Mandel, J. R., Dikow, R. B., and Funk, V. A. 2015. Using phylogenomics to resolve megafamilies: An example from Compositae. *Journal of Systematics and Evolution*, 53(5): 391–402.
- Matzke, N. J. and Wright, A. 2016. Inferring node dates from tip dates in fossil Canidae: the importance of tree priors. *Biology Letters*, 12(8).
- McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66(2): 526–538.
- Mendes, F. K. and Hahn, M. W. 2016. Gene tree discordance causes apparent substitution rate variation. *Systematic Biology*, 65(4): 711–721.
- Mendes, F. K. and Hahn, M. W. 2017. Why concatenation fails near the anomaly zone. *Systematic Biology*. Advance article.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemistry and Physics*, 21: 1087–1092.
- Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N.,

- Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104): 222–226.
- Mirarab, S. and Warnow, T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12): i44–i52.
- Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. 2014a. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–i548.
- Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. 2014b. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215): 1337.
- Mirarab, S., Bayzid, M. S., and Warnow, T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3): 366–380.
- Mitchell, K. J., Cooper, A., and Phillips, M. J. 2015. Comment on “Whole-genome analyses resolve early branches in the tree of life of modern birds”. *Science*, 349(6255): 1460.
- Moen, D. and Morlon, H. 2014. Why does diversification slow down? *Trends in Ecology & Evolution*, 29(4): 190–197.
- Moritz, C. C., Pratt, R. C., Bank, S., Bourke, G., Bragg, J. G., Doughty, P., Keogh, J. S., Laver, R. J., Potter, S., Teasdale, L. C., Tedeschi, L. G., and Oliver, P. M. 2017. Cryptic lineage diversity, body size divergence, and sympatry in a species complex of Australian lizards (Gehyra). *Evolution*. Early View.

- Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., and Springer, M. S. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294(5550): 2348–2351.
- Nabholz, B., Glémin, S., and Galtier, N. 2009. The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. *BMC Evolutionary Biology*, 9(1): 54.
- Nee, S., Holmes, E. C., May, R. M., and Harvey, P. H. 1994. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1307): 77–82.
- Nielsen, R. and Wakeley, J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158(2): 885–896.
- Nylander, J. A., Ronquist, F., Huelsenbeck, J. P., and Nieves-Aldrey, J. 2004. Bayesian phylogenetic analysis of combined data. *Systematic Biology*, 53(1): 47–67.
- O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z., Meng, J., Ni, X., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R., and Cirranello, A. L. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science*, 339(6120): 662–667.
- O'Neill, E. M., Schwartz, R., Bullock, C. T., Williams, J. S., Shaffer, H. B., Aguilar-Miguel, X., Parra-Olea, G., and Weisrock, D. W. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, 22(1): 111–129.

- Ottenburghs, J., Megens, H.-J., Kraus, R. H. S., van Hooft, P., van Wieren, S. E., Crooijmans, R. P. M. A., Ydenberg, R. C., Groenen, M. A. M., and Prins, H. H. T. 2017. A history of hybrids? Genomic patterns of introgression in the true geese. *BMC Evolutionary Biology*, 17(1): 201.
- Paajanen, P., Kettleborough, G., Lopez-Girona, E., Giolai, M., Heavens, D., Baker, D., Lister, A., Wilde, G., Hein, I., Macaulay, I., Bryan, G. J., and Clark, M. D. 2017. A critical comparison of technologies for a plant genome sequencing project. *bioRxiv*. doi:10.1101/201830.
- Page, R. D. M. and Charleston, M. A. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7(2): 231–240.
- Pamilo, P. and Nei, M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5): 568–583.
- Pardi, F. and Scornavacca, C. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Computational Biology*, 11(4): e1004135.
- Park, H., Jin, G., and Nakhleh, L. 2010. Algorithmic strategies for estimating the amount of reticulation from a collection of gene trees. In *Proceedings of the 9th Annual International Conference on Computational Systems Biology*, pages 114–123.
- Paterson, A. M., Wallis, G. P., Kennedy, M., and Gray, R. D. 2014. Behavioural evolution in penguins does not reflect phylogeny. *Cladistics*, 30(3): 243–259.
- Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097): 1103–1108.
- Perrot-Minnot, M.-J., Špakulová, M., Wattier, R., Kotlík, P., Düşen, S., Aydoğdu, A., and

- Tougaard, C. 2017. Contrasting phylogeography of two Western Palaearctic fish parasites despite similar life cycles. *Journal of Biogeography*. Early View.
- Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M., Pritchard, J. K., and Gilad, Y. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, 22(4): 602–610.
- Phillips, M. J., Bennett, T. H., and Lee, M. S. Y. 2009. Molecules, morphology, and ecology indicate a recent, amphibious ancestry for echidnas. *Proceedings of the National Academy of Sciences*, 106(40): 17089–17094.
- Prevosti, F. J. 2010. Phylogeny of the large extinct South American Canids (Mammalia, Carnivora, Canidae) using a “total evidence” approach. *Cladistics*, 26(5): 456–481.
- Pybus, O. G. and Harvey, P. H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, 267(1459): 2267–2272.
- Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmon, A. R., and Burbrink, F. T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution*, 81: 221–231.
- Rambaut, A. and Grassly, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13(3): 235–238.
- Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

- Rannala, B. and Yang, Z. 2007. Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56(3): 453–466.
- Rannala, B. and Yang, Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, 66(5): 823–842.
- Rasmussen, M. D. and Kellis, M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Research*, 17(12): 1932–1942.
- Reaz, R., Bayzid, M. S., and Rahman, M. S. 2014. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLOS ONE*, 9(8): e104008.
- Rieseberg, L. H. 1991. Homoploid reticulate evolution in *Helianthus* (Asteraceae): Evidence from ribosomal genes. *American Journal of Botany*, 78(9): 1218–1237.
- Robinson, D. and Foulds, L. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53: 131–147.
- Roch, S. and Steel, M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100: 56–62.
- Rokas, A., Williams, B. L., King, N., and Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960): 798–804.
- Ronquist, F. and Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12): 1572–1574.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012a. MrBayes 3.2: efficient Bayesian

- phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3): 539–542.
- Ronquist, F., Klopstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. 2012b. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, 61: 973–999.
- Rosauer, D. F., Pollock, L. J., Linke, S., and Jetz, W. 2017. Phylogenetically informed spatial planning is required to conserve the mammalian tree of life. *Proceedings of the Royal Society of London B: Biological Sciences*, 284(1865).
- RoyChoudhury, A., Felsenstein, J., and Thompson, E. A. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics*, 180(2): 1095–1105.
- Rusinko, J. and McPartlon, M. 2017. Species tree estimation using neighbor joining. *Journal of Theoretical Biology*, 414(Supplement C): 5–7.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425.
- Scally, A. and Durbin, R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13: 745–753.
- Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O’Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T.,

- Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C., and Durbin, R. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388): 169–175.
- Scornavacca, C. and Galtier, N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1): 112–120.
- Shapiro, B. and Hofreiter, M. 2014. A paleogenomic perspective on evolution and gene function: New insights from ancient DNA. *Science*, 343(6169): 1236573.
- Sibley, C. G. and Ahlquist, J. E. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *Journal of Molecular Evolution*, 20(1): 2–15.
- Sjöstrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. 2014. A Bayesian method for analyzing lateral gene transfer. *Systematic Biology*, 63(3): 409–420.
- Slater, G. J. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. *Proceedings of the National Academy of Sciences*, 112(16): 4897–4902.
- Slowinski, J. B. and Page, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48(4): 814–825.
- Solís-Lemus, C. and Ané, C. 2016. Inferring phylogenetic networks with maximum pseudo-likelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3): e1005896.
- Solís-Lemus, C., Bastide, P., and Ané, C. 2017. PhyloNetworks: A package for phylogenetic networks. *Molecular Biology and Evolution*, 34(12): 3292–3298.

- Springer, M. S. and Gatesy, J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94, Part A: 1–33.
- Stadler, T. 2008. Lineages-through-time plots of neutral models for speciation. *Mathematical Biosciences*, 216(2): 163–171.
- Stadler, T. 2010. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267(3): 396–404.
- Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H. F., Drummond, A. J., Bonhoeffer, S., and the Swiss HIV Cohort Study 2012. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, 29(1): 347–357.
- Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1): 228–233.
- Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.
- Steel, M. and Mooers, A. 2010. The expected length of pendant and interior edges of a Yule tree. *Applied Mathematics Letters*, 23(11): 1315–1319.
- Steiper, M. E. and Young, N. M. 2006. Primate molecular divergence dates. *Molecular Phylogenetics and Evolution*, 41(2): 384–394.
- Streicher, J. W., Schulte, II, J. A., and Wiens, J. J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Systematic Biology*, 65(1): 128–145.

- Sukumaran, J. and Holder, M. T. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12): 1569–1571.
- Sun, Y., Abbott, R. J., Li, L., Li, L., Zou, J., and Liu, J. 2014. Evolutionary history of purple cone spruce (*Picea purpurea*) in the Qinghai-Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Molecular Ecology*, 23(2): 343–359.
- Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. 2015. The inference of gene trees with species trees. *Systematic Biology*, 64(1): e42–e62.
- Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43): 17513–17518.
- Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. 2013. Lateral gene transfer from the dead. *Systematic Biology*, 62(3): 386–397.
- Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. 2015. The Inference of Gene Trees with Species Trees. *Systematic Biology*, 64(1): e42–e62.
- Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In R. Miura, editor, *Some Mathematical Questions in Biology: DNA Sequence Analysis*, volume 17 of *Lectures on mathematics in the life sciences*, pages 57–86. American Mathematical Society, Providence, Rhode Island.
- Tavaré, S., Marshall, C. R., Will, O., Soligo, C., and Martin, R. D. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, 416(6882): 726–729.
- Tedford, R. H., Wang, X., and Taylor, B. E. 2009. Phylogenetic systematics of the North

- American fossil caninae (Carnivora: Canidae). *Bulletin of the American Museum of Natural History*, 325.
- Than, C., Ruths, D., and Nakhleh, L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(1): 322.
- Thorne, J. L. and Kishino, H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*, 51(5): 689–702.
- Thorne, J. L., Kishino, H., and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12): 1647–1657.
- Tofigh, A., Hallett, M., and Lagergren, J. 2011. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2): 517–535.
- Tougaard, C., García Dávila, C. R., Römer, U., Duponchelle, F., Cerqueira, F., Paradis, E., Guinand, B., Angulo Chávez, C., Salas, V., Quérrouil, S., Sirvas, S., and Renno, J.-F. 2017. Tempo and rates of diversification in the South American cichlid genus *Apistogramma* (Teleostei: Perciformes: Cichlidae). *PLOS ONE*, 12(9): e0182618.
- Vaughan, T. G. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15): 2392–2394.
- vonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., Shapiro, B., Wall, J., and Wayne, R. K. 2016. Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, 2(7): e1501714.
- Wang, X. 1994. Phylogenetic systematics of the Hesperocyoninae (Carnivora, Canidae). *Bulletin of the American Museum of Natural History*, 221.

- Wang, X., Tedford, R. H., and Taylor, B. E. 1999. Phylogenetic systematics of the Borophaginae (Carnivora, Canidae). *Bulletin of the American Museum of Natural History*, 243.
- Wen, D. and Nakhleh, L. 2017. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*. Advance article.
- Wen, D., Yu, Y., and Nakhleh, L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS genetics*, 12(5): e1006006.
- White, T. D., Asfaw, B., Beyene, Y., Haile-Selassie, Y., Lovejoy, C. O., Suwa, G., and Wolde-Gabriel, G. 2009. *Ardipithecus ramidus* and the paleobiology of early hominids. *Science*, 326(5949): 75–86.
- White, T. D., Lovejoy, C. O., Asfaw, B., Carlson, J. P., and Suwa, G. 2015. Neither chimpanzee nor human, *Ardipithecus* reveals the surprising ancestry of both. *Proceedings of the National Academy of Sciences*, 112(16): 4877–4884.
- Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company, 2nd edition.
- Wiens, J. J. and Morrill, M. C. 2011. Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. *Systematic Biology*, 60(5): 719–731.
- Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z., and Tavaré, S. 2011. Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology*, 60(1): 16–31.
- Wilkinson-Herbots, H. M. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the “isolation with migration” model. *Theoretical Population Biology*, 73(2): 277–288.

- Wilson, I. J. and Balding, D. J. 1998. Genealogical inference from microsatellite data. *Genetics*, 150(1): 499–510.
- Wood, B. and Harrison, T. 2011. The evolutionary context of the first hominins. *Nature*, 470(7334): 347–352.
- Wu, Y. 2010. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics*, 26(12): i140–i148.
- Xi, Z., Liu, L., and Davis, C. C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, 92: 63–71.
- Xu, B. and Yang, Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204(4): 1353–1368.
- Yang, F. S. and Wang, X. Q. 2007. Extensive length variation in the cpDNA *trnT-trnF* region of hemiparasitic *Pedicularis* and its phylogenetic implications. *Plant Systematics and Evolution*, 264(3-4): 251–264.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3): 306–314.
- Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.
- Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Current Zoology*, 61(5): 854–865.
- Yang, Z. and Rannala, B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution*, 31(12): 3125–3135.

- Yang, Z., Goldman, N., and Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, 11(2): 316–324.
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1): 28–36.
- Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(Suppl 10): S10.
- Yu, Y., Than, C., Degnan, J. H., and Nakhleh, L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2): 138–149.
- Yu, Y., Degnan, J. H., and Nakhleh, L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS genetics*, 8(4): e1002660.
- Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46): 16448–16453.
- Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(403): 21–87.
- Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. 2016. Total-evidence dating under the fossilized birth-death process. *Systematic Biology*, 65(2): 228–249.

- Zhang, C., Sayyari, E., and Mirarab, S. 2017. ASTRAL-III: Increased scalability and impacts of contracting low support branches. In J. Meidanis and L. Nakhleh, editors, *15th International Workshop on Comparative Genomics*, pages 53–75.
- Zhu, J., Wen, D., Yu, Y., Meudt, H., and Nakhleh, L. 2017. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *bioRxiv*. doi:10.1101/143545.
- Zhu, S. and Degnan, J. H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Systematic Biology*, 66(2): 283–298.
- Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, 29(10): 3131–3142.
- Zimmermann, T., Mirarab, S., and Warnow, T. 2014. BBCA: Improving the scalability of *BEAST using random binning. *BMC Genomics*, 15(Suppl 6): S11.
- Zrzavý, J. and Řičánková, V. 2004. Phylogeny of recent Canidae (Mammalia, Carnivora): relative reliability and utility of morphological and molecular datasets. *Zoologica Scripta*, 33(4): 311–333.
- Zuckerlandl, E. and Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, pages 97–166.