Title

Phylogenetic Models Linking Speciation and Extinction to Chromosome and Mating System Evolution

Permalink

https://escholarship.org/uc/item/29n8r0nm

Author

Freyman, William Allen

Publication Date

2017

Peer reviewed|Thesis/dissertation

Phylogenetic Models Linking Speciation and Extinction to Chromosome and Mating System Evolution

by

William Allen Freyman

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

Doctor of Philosophy
in
Integrative Biology
and the Designated Emphasis
in
Computational and Genomic Biology

in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Dr. Bruce G. Baldwin, Chair
Dr. John P. Huelsenbeck
Dr. Brent D. Mishler
Dr. Kipling W. Will

Fall 2017

Phylogenetic Models Linking Speciation and Extinction to Chromosome and Mating System Evolution

Abstract

Phylogenetic Models Linking Speciation and Extinction to Chromosome and Mating System Evolution

by
William Allen Freyman
Doctor of Philosophy in Integrative Biology
and the Designated Emphasis in Computational and Genomic Biology
University of California, Berkeley
Dr. Bruce G. Baldwin, Chair

Key evolutionary transitions have shaped the tree of life by driving the processes of speciation and extinction. This dissertation aims to advance statistical and computational approaches that model the timing and nature of these transitions over evolutionary trees. These methodological developments in phylogenetic comparative biology enable formal, model-based, statistical examinations of the macroevolutionary consequences of trait evolution. Chapter 1 presents computational tools for data mining the large-scale molecular sequence datasets needed for comparative phylogenetic analyses. I describe a novel metric, the missing sequence decisiveness score (MSDS), which assesses the phylogenetic decisiveness of a matrix given the pattern of missing sequence data. In Chapter 2, I introduce a class of phylogenetic models of chromosome number evolution that accommodate both anagenetic and cladogenetic change. The models reveal the mode of chromosome number evolution; is chromosome evolution occurring primarily within lineages, primarily at lineage splitting, or in clade-specific combinations of both? Furthermore, these models permit estimation of the location and timing of possible chromosome speciation events over the phylogeny. Finally, Chapter 3 demonstrates a new method of stochastic character mapping for state-dependent speciation and extinction models and applies it to test the impact of plant mating systems on the extinction of lineages. This approach estimates the timing of both character state transitions and shifts in diversification rates over the phylogeny. Confirming long standing theory, I found that self-compatible lineages have higher extinction rates and lower net diversification rates compared to self-incompatible lineages. Additionally, the method shows that the loss of self-incompatibility is followed by a short-term spike in speciation rates, which declines after a time lag of several million years resulting in evolutionary decline.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I want to deeply thank my wife, Sophia, and daughter, Aliya, for not only tolerating my impractical career choices but also for their love and support as I find my way through the arcane procedures of academia. I am also profoundly grateful to my parents, Bill and Sarah, for inspiring my awe of the diversity of the natural world and my appreciation for the abstractions of thought that link music, literature, and science.

None of this work would have been possible if not for my doctoral advisor, Bruce Baldwin. I thank him for taking a chance with a nontraditional student and encouraging me to pursue my own research directions. He always made himself available to provide guidance and I would have been adrift without his expertise and mentorship. Additionally, I thank Brent Mishler, John Huelsenbeck, and Kip Will for serving on my dissertation committee and teaching me a great deal. I had the tremendous opportunity to teach phylogenetics with all three professors and their diverse perspectives and experience have enriched my understanding. In addition to my committee, I have benefited immensely from interactions with other professors both at UC Berkeley and beyond, in particular David Ackerly, Paul Fine, Emma Goldberg, Charles Marshall, Brian Moore, Kevin Padian, and Carl Rothfels. Their exceptional insights and scholarship have made me a better scientist.

I want to especially thank my unofficial mentor Sebastian Höhna, who was a patient guide and collaborator. Sebastian is a major inspiration who pushed me to challenge myself, and I am honored by the contributions he made as coauthor of two of the publications that resulted from this dissertation. I also want to thank all other RevBayes developers whose code I have learned from, particularly Tracy Heath and Michael Landis who have shared advice and encouragement.

Through botanical adventures in the wilds of California and shared time in our windowless dungeon below the Valley Life Sciences Building, my labmates in the Baldwin lab taught me a ton and became good friends: Matt Guilliams, Isaac Lichter-Marck, Mike Park, Adam Schneider, and Genevieve Walden. Bridget Wessa in the Molecular Phylogenetics Lab and the entire staff of the University and Jepson Herbaria helped make my research possible. Many other graduate students became important friends and colleagues: all the members of the Rothfels, Mishler, and Specht labs and those who participated in our lively Berkeley, Beer, and Biosystematics discussion group. I'd like to specially thank Ingrid Jordan-Thaden, Seema Sheth, and Andrew Thornhill who have provided advice and comradery.

I came to evolutionary biology through a rather convoluted career path; but a major inspiration was my time working with Stephen Packard and the amazing stewards of Somme Prairie Grove and other Chicago area forest preserves. I additionally want to thank Linda Masters at Openlands, the staff of Audubon Chicago, and Joel Olfelt at Northeastern Illinois

University for guiding me on my first tentative steps in science.

Finally, the opportunity to devote years of one's life learning and researching is a great privilege that I deeply appreciate. I am grateful for the financial support I received from UC Berkeley and the National Science Foundation during this time. These institutions have supported my research despite the fact that it does not have any obvious economic application. Too few people have a similar privilege with the access to support and guidance that I have had; I hope that in my career moving forward I can help make this opportunity possible for others and strengthen our society's investment in basic research.

Introduction

The evolution of life has long been understood to be a branching process through time, with all extant species representing twigs on the great tree of life. In the past 20 years, evolutionary biology has made enormous progress by mathematically modeling this branching process using the framework of phylogenetic theory. A major challenge for phylogenetic theory has been modeling the key evolutionary transitions that drive the diversification of life; transitions in trait evolution that may be associated with the speciation process and/or may drive shifts in the macroevolutionary rates of speciation and extinction. This dissertation aims to advance statistical and computational approaches that model the timing and nature of these transitions over evolutionary trees. These methodological developments in phylogenetic comparative biology enable formal, model-based, statistical examinations of the macroevolutionary consequences of trait evolution. I apply these methods to test long-standing hypotheses about the role of chromosome changes in the speciation process and the impact of plant mating systems on the extinction of lineages.

Comparative phylogenetic analyses require increasingly massive datasets to achieve statistical power. Fortunately the amount of phylogenetically informative sequence data available in online databases is growing at an exponential rate. However, computational techniques are needed to extract this data from online repositories and automate construction of large-scale molecular sequence matrices. These approaches frequently produce vast matrices with sparse taxon coverage that underscore the need for methods to evaluate the effect of missing data. Chapter 1 presents the software Supermatrix Constructor (SUMAC), a tool to data mine GenBank, construct phylogenetic supermatrices, and assess the phylogenetic decisiveness of a matrix given the pattern of missing sequence data. I develop a novel metric, the missing sequence decisiveness score (MSDS), which measures how much each individual missing sequence contributes to the phylogenetic decisiveness of the matrix. MSDS can be used to compare supermatrices and prioritize the acquisition of new sequence data. This approach is then used to construct datasets for the following chapters.

Chapter 2 introduces a class of phylogenetic models of chromosome number evolution that accommodate both anagenetic and cladogenetic change. Chromosome number is a key feature of the higher-order organization of the genome, and changes in chromosome number play a fundamental role in evolution and possibly the speciation process itself. Dysploid gains and losses in chromosome number, as well as polyploidization events, may drive reproductive isolation and lineage diversification. The models developed here reveal the mode of chromosome number evolution; is chromosome evolution occurring primarily within lineages, primarily at lineage splitting, or in clade-specific combinations of both? Furthermore, these models permit estimation of the location and timing of possible chromosome speciation events over the phylogeny. I test the models' accuracy with simulations and re-examine chromosomal evolution in *Aristolochia*, *Carex* section *Spirostachyae*, *Helianthus*, *Mimulus* sensu lato, and *Primula* section *Aleuritia*, finding evidence for clade-specific combinations of anagenetic and cladogenetic dysploid and polyploid modes of chromosome evolution.

Chapter 3 demonstrates a new method of stochastic character mapping for state-dependent speciation and extinction (SSE) models. This approach estimates the timing and nature of both character state transitions and shifts in diversification rates over the phylogeny. I apply it to study mating system evolution over a densely sampled fossil-calibrated phylogeny of the plant family Onagraceae. Utilizing a hidden state SSE model I tested the association of the loss of self-incompatibility with shifts in diversification rates. Confirming long standing theory, I found that self-compatible lineages have higher extinction rates and lower net diversification rates compared to self-incompatible lineages. Further, my mapped character histories show that the loss of self-incompatibility is followed by a short-term spike in speciation rates, which declines after a time lag of several million years resulting in negative net diversification. Lineages that have long been self-compatible such as *Fuchsia* and *Clarkia* are in a previously unrecognized and ongoing evolutionary decline.

Chapter 1

Data mining for large-scale phylogenetic analyses

Abstract

The amount of phylogenetically informative sequence data in GenBank is growing at an exponential rate, and large phylogenetic trees are increasingly used in research. Tools are needed to to construct phylogenetic sequence matrices from GenBank data and evaluate the effect of missing data. `Supermatrix Constructor` (`SUMAC`) is a tool to data mine GenBank, construct phylogenetic supermatrices, and assess the phylogenetic decisiveness of a matrix given the pattern of missing sequence data. `SUMAC` calculates a novel metric, missing sequence decisiveness scores (MSDS), which measure how much each individual missing sequence contributes to the decisiveness of the matrix. MSDS can be used to compare supermatrices and prioritize the acquisition of new sequence data. `SUMAC` constructs supermatrices either through an exploratory clustering of all GenBank sequences within a taxonomic group, or by using guide sequences to build homologous clusters in a more targeted manner. `SUMAC` assembles supermatrices for any taxonomic group recognized in GenBank, and is optimized to run on multicore computer systems by parallelizing multiple stages of operation. `SUMAC` is implemented as a `Python` package that can run as a stand-alone command line program, or its modules and objects can be incorporated within other programs. `SUMAC` is released under the open source GPLv3 license and is available at `https://github.com/wf8/sumac`.

1.1 Introduction

In pursuit of large-scale evolutionary questions, biologists are increasingly using massive phylogenetic datasets to reconstruct ever-growing portions of the tree of life. These large phylogenetic trees are commonly inferred using a supermatrix approach, in which multiple datasets are combined and analyzed simultaneously (de Queiroz and Gatesy 2007). However, assembling and utilizing supermatrices is challenging due to difficulties such as determining homology of molecular sequences, assembling chimeric operational taxonomic units, and managing the amount of missing data. Despite these challenges, considerable bioinformatic advances have made large supermatrix based phylogenetic analyses more common.

Multiple software tools for building supermatrices are already available to evolutionary biologists. The PhyLoTA Browser (Sanderson et al. 2008) provides a web interface to view all GenBank sequences within taxonomic groups clustered into homologs. A different approach is implemented in the programs PHLAWD (Smith et al. 2009) and NCBIminer (Xu et al. 2015), which mine GenBank for sequence clusters homologous to guide sequences provided by the user. The method implemented in SUMAC (Supermatrix Constructor) combines elements of both approaches; the user can perform an exploratory clustering of all GenBank sequences within a taxonomic group or provide guide sequences to build homologous sequence clusters in a more targeted manner. Furthermore, by calculating supermatrix assessment metrics derived from the concept of phylogenetic decisiveness (Steel and Sanderson 2010) SUMAC provides a unique toolkit with which GenBank can be repeatedly mined using different settings and the resulting data matrices can be compared. In this paper my objectives are to (1) introduce the SUMAC software, (2) describe a novel metric that assesses the effect of missing data in phylogenetic supermatrices, and (3) illustrate the use of SUMAC with a case study.

1.2 Implementation

1.2.1 Overview

SUMAC is a Python package designed to run as a stand-alone command-line program, though the modules can also be imported and used in other Python scripts. When run from the command-line, SUMAC will perform a number of steps to construct a supermatrix. First, SUMAC creates a local SQLite3 (Hipp and Kennedy 2007) database of the specified GenBank division (e.g. PLN or MAM), automatically downloading sequences from NCBI if necessary. Using NCBI taxonomy, SUMAC searches the local database for all sequences in the user-specified ingroup and outgroup. Found sequences are then clustered as putative homologs in one of two ways: (1) performing exhaustive all-by-all BLASTn (Camacho et al. 2009) comparisons of each ingroup and outgroup sequence and using a single-linkage hierarchical clustering algorithm, or (2) user-provided guide sequences that typify each cluster are BLASTed against all ingroup and outgroup sequences.

1.2.2 Hierarchical clustering algorithms

By default, SUMAC clusters sequences using the SLINK (Sibson 1973) single-linkage hierarchical clustering algorithm. This achieves $0(n^2)$ time complexity by representing the dendrogram of hierarchical sequence clusters in pointer representation. Given n sequences and the dendrogram c, pointer representation consists of two functions:

$$\Pi(i) = \max\{j : (i,j) \in c(\Lambda(i)) \wedge i,j \in [0, n-1]\}$$
$$\Lambda(i) = \inf\{h : \exists j > i \wedge (i,j) \in c(h) \wedge i,j \in [0, n-1]\}$$

The function $\Pi(i)$ is the last sequence that sequence i clusters with, and $\Lambda(i)$ is the distance h (the BLAST e-value) between sequence $\Pi(i)$ and sequence i. SUMAC's default clustering depth is an e-value threshold of $1.0e{-}10$ and a sequence length percent similarity threshold of 0.5, though both thresholds can be modified by the user with optional command-line arguments. If run with the command line flag `--hac`, `SUMAC` will instead cluster sequences using a naive hierarchical agglomerative clustering (HAC) algorithm. Proposed by Sneath (Sneath 1957), this single-linkage clustering algorithm uses an agglomerative scheme that merges the closest sequence clusters into consecutively larger clusters. However, with $O(n^3)$ time complexity the HAC algorithm is considerably less efficient than the SLINK algorithm.

1.2.3 Alignments

Once clustering is complete, `SUMAC` discards clusters that are not phylogenetically informative (<4 taxa), and aligns each cluster of sequences using `MAFFT` (Katoh et al. 2002) with the `--adjustdirection` flag to ensure correct sequence polarity. The individual locus alignments are saved to enable gene tree inference, and then the alignments are concatenated by species binomial (based on the NCBI taxonomy) to create the final supermatrix. Finally, a number of metrics are reported, a graph indicating taxon coverage density is generated, and spreadsheets (in CSV format) are produced with information about each DNA region and GenBank accession used in the supermatrix.

1.2.4 Parallelization

`SUMAC` utilizes `Python`'s multiprocessing module (Python Software Foundation 2008) to parallelize `BLAST` comparisons and `MAFFT` alignments on multicore computer systems. `SUMAC` also depends on the `BioPython` (Cock et al. 2009) library for sequence manipulation.

1.3 Missing Sequence Decisiveness Scores

Large-scale sequence matrices may contain a great deal of missing data, and quantifying the effect of that missing data can be difficult. When run with the `--decisiveness` command-line flag, `SUMAC` will calculate the *fraction of triples*, a metric of the partial decisiveness (PD) of the sequence matrix (Sanderson et al. 2010). PD measures how the arrangement of missing data in a multi-locus sequence matrix limits the number of trees out of all possible trees that can be inferred. The fraction of triples is the easiest PD metric to compute and applies to the set of all rooted trees; it is the percentage of each possible set of three taxa which all have sequence data for at least one of the same gene regions.

Here I extend the fraction of triples concept by introducing missing sequence decisiveness scores (MSDS). MSDS measure the contribution of each individual missing sequence to the overall PD of the matrix. MSDS values are in the range $[0, 1]$ and are only assigned to missing sequences. When the MSDS of a missing sequence is close to 1 the addition of new data will increase the PD of the matrix more than where MSDS is low. In this way MSDS prioritize which sequences to add to the matrix, and identifies taxa or loci that contribute disproportionately to the lack of decisiveness in the matrix. `SUMAC` produces a graph that portrays the distribution of MSDS across the supermatrix (Figure 1.1). PD metrics and MSDS can be applied to any multi-locus phylogenetic matrix, thus `SUMAC` can calculate these metrics for user provided sequence alignments as well as those mined by `SUMAC` from GenBank.

Given a set of n taxa X and a collection $S = \{Y_1, ..., Y_k\}$ of subsets of X with an overall fraction of triples ϵ, the MSDS M_{ij} of taxon i and locus j is:

$$M_{ij} = \left[\frac{\Theta_i - \min\{\Theta_l : l \in X\}}{\max\{\Theta_m : m \in X\} - \min\{\Theta_l : l \in X\}} + \frac{\Upsilon_j - \min\{\Upsilon_s : s \in S\}}{\max\{\Upsilon_t : t \in S\} - \min\{\Upsilon_s : s \in S\}} \right] \Big/ 2,$$

$$\text{where } \Theta_i = \frac{\epsilon}{\epsilon_i}, \Upsilon_j = \frac{\epsilon}{\epsilon_j}.$$

ϵ_i is the fraction of triples of S with taxon i removed, and ϵ_j is the fraction of triples of S with locus j removed. For the case $\epsilon_i = 0$ or $\epsilon_j = 0$:

$$\Theta_i = \epsilon \binom{n}{3}, \Upsilon_j = \epsilon \binom{n}{3}.$$

The calculations above are performed after values for ϵ, ϵ_i for all $i \in X$, and ϵ_j for all $j \in S$ are computed using a modified version of Fischer's phylogenetic decisiveness decision problem algorithm for rooted trees (Fischer 2012). This algorithm has an $O(k \cdot n^3)$ time complexity.

1.4 Case Study: Onagraceae

1.4.1 Overview

To demonstrate the utility of `SUMAC` for discovering phylogenetically informative sequences within GenBank, I compared the construction of a phylogenetic supermatrix using both the `PhyLoTA Browser` (Sanderson et al. 2008) and `SUMAC`. I did not use `PHLAWD` (Smith et al. 2009) or `NCBIminer` (Xu et al. 2015) since they only target genes already known to be of interest. The goal of this example was to build a supermatrix of the plant families Onagraceae (as an ingroup) and Lythraceae (as an outgroup) with as many informative loci as possible.

1.4.2 `PhyLoTa` data mining

I searched the `PhyLoTa` database for the taxon names Onagraceae and Lythraceae, retrieving 5504 and 2547 sequences respectively. `PhyLoTa` constructed supermatrics for each of the two groups separately, resulting in an Onagraceae supermatrix with 325 species and 43 phylogenetically informative sequence clusters. The Lythraceae supermatrix had 172 species and 77 phylogenetically informative clusters. Upon inspection many of the sequence clusters should have been combined; for example, 8 of the 43 Onagraceae clusters were fragments of the 18S ribosomal gene. To use these data for a phylogenetic analysis the 120 Onagraceae and Lythraceae clusters would need to be reviewed and manually combined.

1.4.3 SUMAC data mining

I ran SUMAC with the command `python -m sumac -d pln -i Onagraceae -o Lythraceae`. SUMAC retrieved 5764 Onagraceae sequences and 3133 Lythraceae sequences. SUMAC found 846 more sequences than PhyLoTa because SUMAC always uses the latest available release of GenBank (release 205 in this case), whereas PhyLoTa was developed using GenBank release 194. SUMAC constructed an initial supermatrix of 599 Onagraceae and Lythraceae species consisting of 108 phylogenetically informative sequence clusters.

Like the results from PhyLoTa, some of the 108 sequence clusters should have been combined (again 8 of the clusters were fragments of 18S ribosomal DNA). With SUMAC, however, the user has options to produce a more satisfactory data matrix. One option is to repeat the data mining process using less stringent thresholds for clustering. These can be configured by the user with the `--evalue` and `--length` flags. Another option, and the one demonstrated here, is to select sequences from the recovered clusters to act as guide sequences and build homologous clusters in a targeted manner similar to the approach used in PHLAWD. This option combines the strengths of both the PhyLoTa and PHLAWD methods.

Table 1.1: Missing sequence decisiveness scores (MSDS) for some of the 2857 missing sequences in the data matrix shown in Figure 1.1. The scores are shown in descending order, prioritizing which holes in the data matrix should be filled to increase the phylogenetic decisiveness of the sequence matrix. SUMAC outputs the entire list as a CSV spreadsheet.

MSDS Rank	MSDS	OTU	Gene Region	Gene Name
1	0.862	*Ludwigia peploides*	1	ITS
2	0.857	*Ludwigia hyssopifolia*	1	ITS
3	0.775	*Epilobium brachycarpum*	1	ITS
4	0.772	*Clarkia lewisii*	1	ITS
5	0.772	*Epilobium macropus*	1	ITS
...				
...				
2855	0.001	*Sonneratia ovata*	2	matK
2856	0.001	*Sonneratia ovata*	9	pgiC
2857	<0.001	*Sonneratia ovata*	3	ndhF

Out of the 108 sequence clusters, I selected guide sequences from the 10 clusters with the highest taxon coverage. SUMAC was then run a second time using the `--guide` flag to produce a final supermatrix of 10 gene regions and 384 species (Figure 1.1). The final taxon coverage density was 0.26 and the partial decisiveness was 0.31. If necessary, this last step could be repeated using different gene regions to try to increase the decisiveness of the sequence matrix. Furthermore, SUMAC prioritized the acquisition of new sequence data by calculating MSDS scores for each missing sequence (Table 1.1).

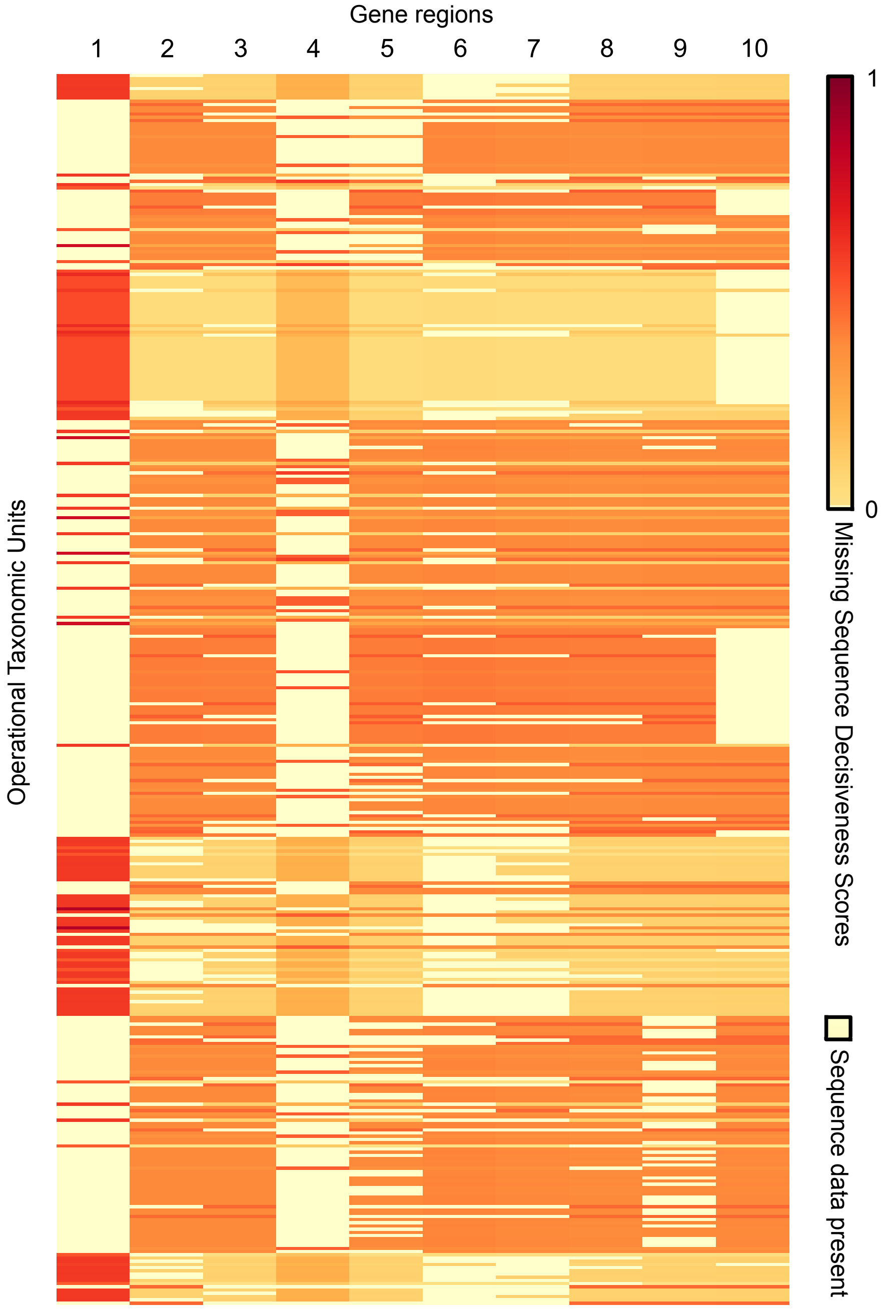

Figure 1.1: Missing sequence decisiveness scores (MSDS) for a sequence matrix with 10 genes, 384 OTUs, taxon coverage density of 0.26, and partial decisiveness of 0.31. Pale yellow represents sequence data present, shades of orange represent missing sequences with low to intermediate MSDS (0-0.75), and red to maroon represents missing sequences with high MSDS (0.75-1.0). MSDS measures how much the individual missing sequence contributes to the decisiveness of the matrix given the overall pattern of missing data. MSDS prioritizes which sequences to add to the matrix; where MSDS is high the addition of new data will increase the decisiveness of the matrix more than where MSDS is low.

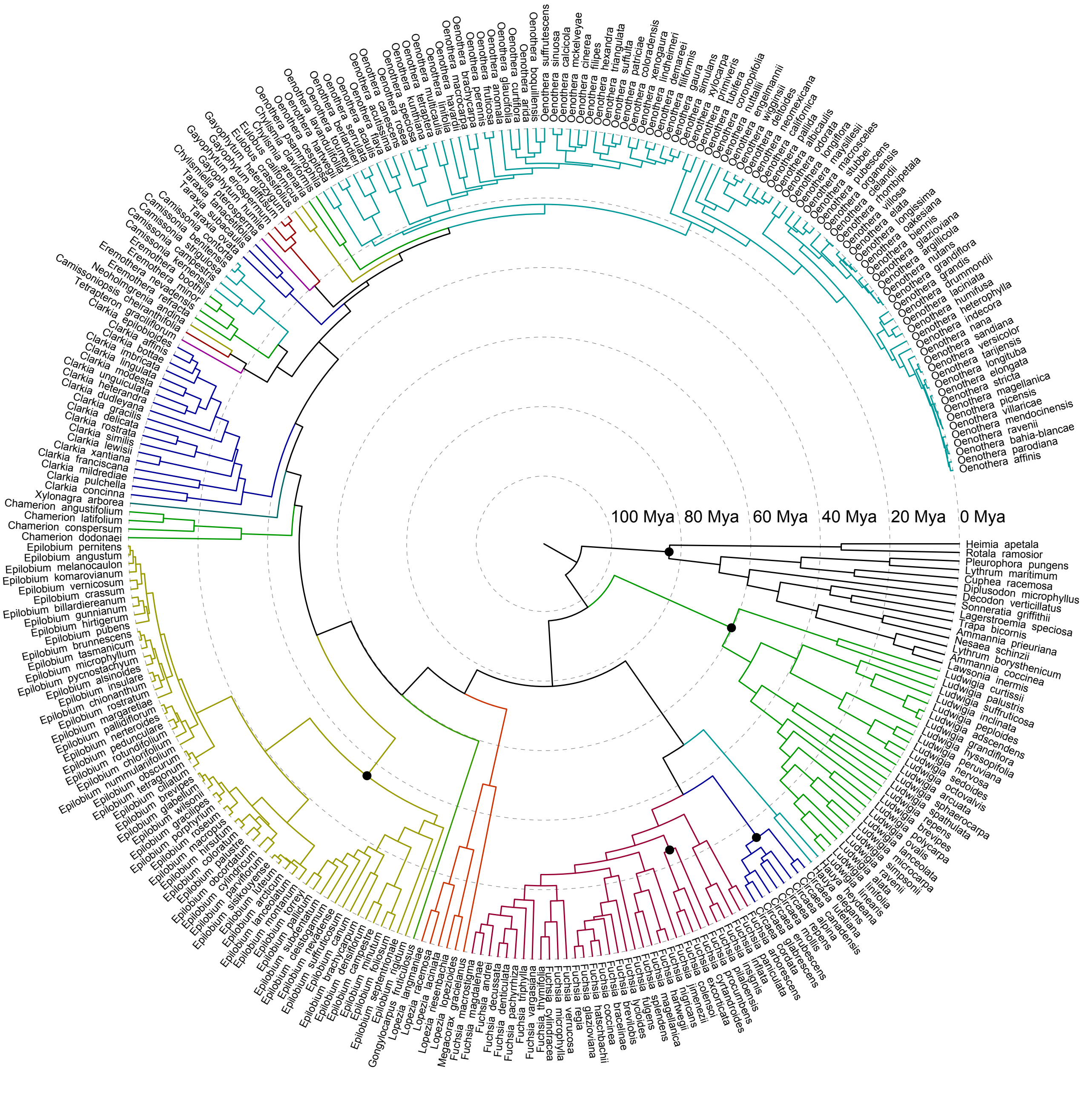

Figure 1.2: Fossil-calibrated Onagraceae phylogeny estimated using data mined by SUMAC. Bayesian chronogram of Onagraceae estimated to demonstrate the utility of SUMAC. Approximate positions of fossil calibration points are shown as black circles. All genera described in Wagner et al. (2007) are colored. Final results shown in Figure 3.10; for the methods used to infer the phylogeny see Section 3.5.2. Detailed divergence time estimates with 95% HPD intervals are shown in Table 3.3.

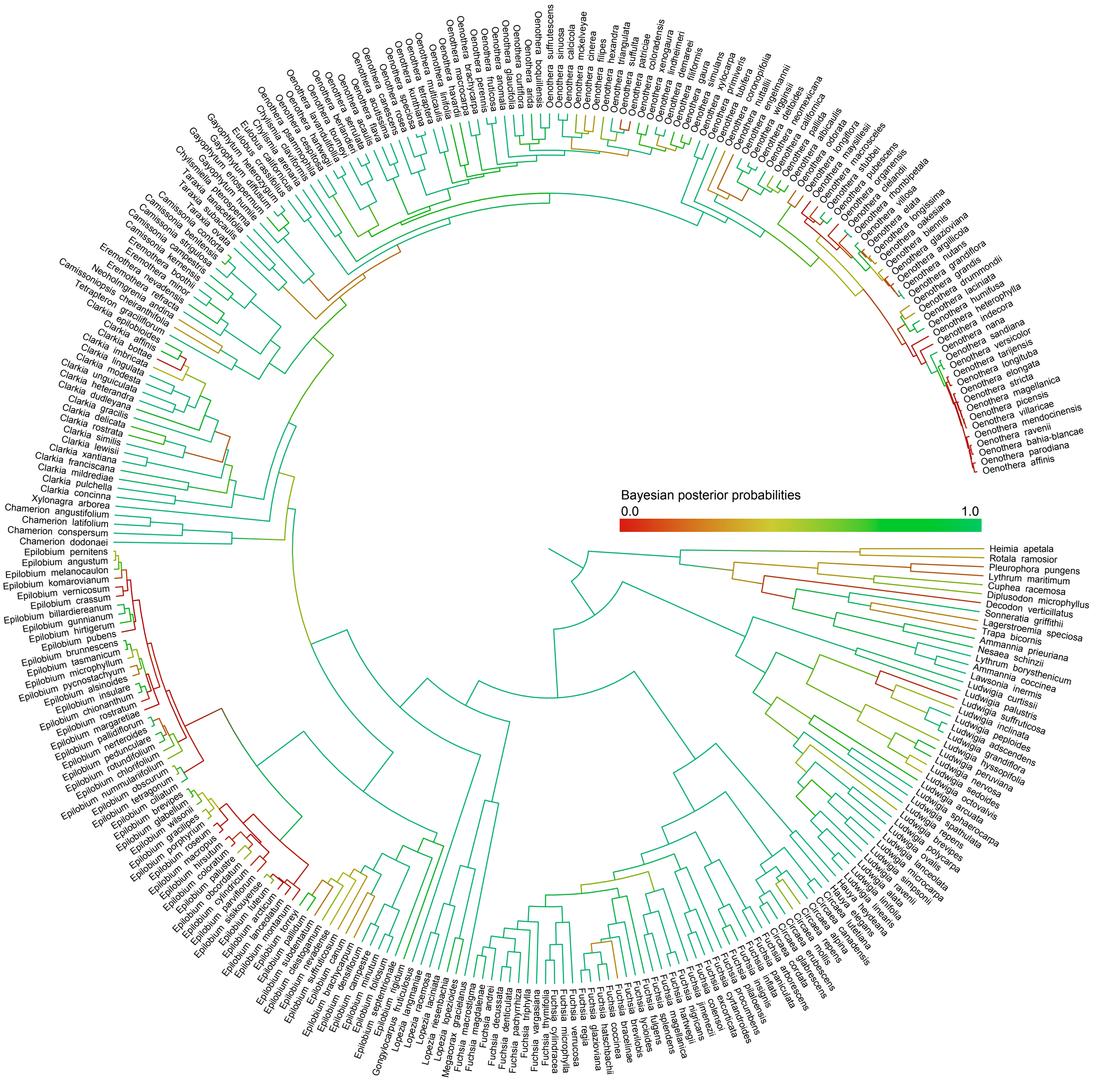

Figure 1.3: Posterior probabilities of Onagraceae phylogeny estimated using data mined by SUMAC. Bayesian chronogram of Onagraceae estimated to demonstrate the utility of SUMAC. Estimated posterior probabilities close to 1.0 are shown in green. Final results shown in Figure 3.10; for the methods used to infer the phylogeny see Section 3.5.2. Detailed divergence time estimates with 95% HPD intervals are shown in Table 3.3.

1.5 Discussion

The advantage of the supermatrix approach to phylogenetic estimation is that it combines data from diverse sources into one large analysis. Using guide sequences makes supermatrix construction much faster, however it requires a priori knowledge of which DNA regions will be used in the supermatrix. Performing all-by-all BLAST comparisons is computationally more expensive, but it effectively data mines GenBank in an exploratory fashion, so that sequence data not necessarily used in previous systematic studies can also be incorporated into the supermatrix. SUMAC enables both options to be pursued, and provides metrics to compare the resulting supermatrices. Additionally, GenBank can be repeatedly mined using different clustering threshold values to optimize the resulting sequence matrix for the taxonomic group being analyzed and the sequence data available.

Missing sequence decisiveness scores (MSDS) quantify the distribution of phylogenetic partial decisiveness over a given multi-locus sequence matrix (Figure 1.1). Multiple properties of MSDS are worth exploring in an expanded simulation study. For example, sequences could be selectively removed from a complete dataset to examine how MSDS are related to phylogenetic uncertainty during tree inference. MSDS could be mapped onto the branches of phylogenies to determine the impact missing data has on the posterior probabilities and/or bootstrap values of clades.

With methodological refinements such as those presented here, supermatrix methods will continue to be widely used for large-scale phylogenetic studies. However, alternative approaches such as supertrees (Von Haeseler 2012) and coalescent-based gene tree/species tree methods (Maddison 1997) are increasingly used. SUMAC outputs both a concatenated supermatrix and individual gene alignments, enabling the application of multiple phylogenetic inference methods. Many of the methodological advances developed for supermatrix approaches apply equally well to gene tree/species tree approaches, thus utilities like SUMAC will continue to be indispensable as researchers aggregate increasingly large phylogenetic datasets and assess the effect of missing data.

Chapter 2

Cladogenetic and anagenetic models of chromosome number evolution

Abstract

Chromosome number is a key feature of the higher-order organization of the genome, and changes in chromosome number play a fundamental role in evolution. Dysploid gains and losses in chromosome number, as well as polyploidization events, may drive reproductive isolation and lineage diversification. The recent development of probabilistic models of chromosome number evolution in the groundbreaking work by Mayrose et al. (2010, ChromEvol) have enabled the inference of ancestral chromosome numbers over molecular phylogenies and generated new interest in studying the role of chromosome changes in evolution. However, the ChromEvol approach assumes all changes occur anagenetically (along branches), and does not model events that are specifically cladogenetic. Cladogenetic changes may be expected if chromosome changes result in reproductive isolation. Here we present a new class of models of chromosome number evolution (called ChromoSSE) that incorporate both anagenetic and cladogenetic change. The ChromoSSE models allow us to determine the mode of chromosome number evolution; is chromosome evolution occurring primarily within lineages, primarily at lineage splitting, or in clade-specific combinations of both? Furthermore, we can estimate the location and timing of possible chromosome speciation events over the phylogeny. We implemented ChromoSSE in a Bayesian statistical framework, specifically in the software `RevBayes`, to accommodate uncertainty in parameter estimates while leveraging the full power of likelihood based methods. We tested ChromoSSE's accuracy with simulations and re-examined chromosomal evolution in *Aristolochia*, *Carex* section *Spirostachyae*, *Helianthus*, *Mimulus* sensu lato (s.l.), and *Primula* section *Aleuritia*, finding evidence for clade-specific combinations of anagenetic and cladogenetic dysploid and polyploid modes of chromosome evolution.

2.1 Introduction

A central organizing component of the higher-order architecture of the genome is chromosome number, and changes in chromosome number have long been understood to play a fundamental role in evolution. In the seminal work *Genetics and the Origin of Species* (1937), Dobzhansky identified "the raw materials for evolution", the sources of natural variation, as two evolutionary processes: mutations and chromosome changes. "Chromosomal changes are one of the mainsprings of evolution," Dobzhansky asserted, and changes in chromosome number such as the gain or loss of a single chromosome (dysploidy), or the doubling of the entire genome (polyploidy), can have phenotypic consequences, affect the rates of recombination, and increase reproductive isolation among lineages and thus drive diversification (Stebbins 1971). Recently, evolutionary biologists have studied the macroevolutionary consequences of chromosome changes within a molecular phylogenetic framework, mostly due to the groundbreaking work of Mayrose et al. (2010, ChromEvol) which introduced likelihood-based models of chromosome number evolution. The ChromEvol models have permitted phylogenetic studies of ancient whole genome duplication events, rapid "catastrophic" chromosome speciation, major reevaluations of the evolution of angiosperms, and new insights into the fate of polyploid lineages (e.g. Pires and Hertweck 2008; Mayrose et al. 2011; Tank et al. 2015).

One aspect of chromosome evolution that has not been thoroughly studied in a probabilistic framework is cladogenetic change in chromosome number. Cladogenetic changes occur solely at speciation events, as opposed to anagenetic changes that occur within lineages and are not associated with speciation events. Studying cladogenetic chromosome changes in a phylogenetic framework has been difficult since the approach used by ChromEvol models only anagenetic changes and ignores the changes that occur specifically at speciation events and may be expected if chromosome changes result in reproductive isolation. Reproductive incompatibilities caused by chromosome changes may play an important role in the speciation process, and led White (1978) to propose that chromosome changes perform "the primary role in the majority of speciation events." Indeed, chromosome fusions and fissions may have played a role in the formation of reproductive isolation and speciation in the great apes (Ayala and Coluzzi 2005), and the importance of polyploidization in plant speciation has long been appreciated (Coyne et al. 2004; Rieseberg and Willis 2007). Recent work by Zhan et al. (2016) revealed phylogenetic evidence that polyploidization is frequently cladogenetic in land plants. However, their approach did not examine the role dysploid changes may play in speciation, and it required a two step analysis in which one first used ChromEvol to infer ploidy levels, and then a second modeling step to infer the proportion of ploidy shifts that were cladogenetic. Since ChromEvol only models anagenetic polyploidization events these two modeling steps are inconsistent with one another.

Here we present models of chromosome number evolution that simultaneously account for both cladogenetic and anagenetic polyploid as well as dysploid changes in chromosome number over a phylogeny. These models reconstruct an explicit history of cladogenetic and anagenetic changes in a clade, enabling estimation of ancestral chromosome numbers. Our approach also identifies different modes of chromosome number evolution among clades; we can detect primarily anagenetic, primarily cladogenetic, or clade-specific combinations of both modes of chromosome changes. Furthermore, these models allow us to infer the timing and location of possible polyploid and dysploid speciation events over the phylogeny. Since these models only account for changes in chromosome number, they ignore speciation that may accompany other types of chromosome rearrangements such as inversions. Our models cannot determine that changes in chromosome number "caused" the speciation event, but they do reveal that speciation and chromosome change are temporally correlated. Thus, these models can give us evidence that the chromosome number change coincided with cladogenesis and so may have played a significant role in the speciation process.

A major challenge for all phylogenetic models of cladogenetic character change is accounting for unobserved speciation events due to lineages going extinct and not leaving any extant descendants (Bokma 2002), or due to incomplete sampling of lineages in the present. Teasing apart the phylogenetic signal for cladogenetic and anagenetic processes given unobserved speciation events is a major difficulty. The Cladogenetic State change Speciation and Extinction (ClaSSE) model (Goldberg and Igić 2012) accounts for unobserved speciation events by jointly modeling both character evolution and the phylogenetic birth-death process. Our class of chromosome evolution models uses the ClaSSE approach, and could be considered a special case of ClaSSE. We implemented our models (called ChromoSSE) in a Bayesian framework and use Markov chain Monte Carlo algorithms to estimate posterior probabilities of the model's parameters. However, compared to most character evolution models, SSE models require additional complexity since they must model extinction and speciation processes. Using simulations, we examined the impact of this additional complexity on our chromosome evolution models' performance. Note that ChromoSSE uses the SSE approach to integrate over all unobserved speciation events and in this work we do not investigate how chromosome number affects diversification rates. Nonetheless, our implementation enables chromosome number dependent speciation and extinction rates to be estimated and this will be explored in future work.

Out of the class of ChromoSSE models described here, it is possible that no single model will adequately describe the chromosome evolution of a given clade. The most parameter-rich ChromoSSE model has at least 12 independent rate parameters, however the models that best describe a given dataset (a phylogeny and a set of observed chromosome counts) may be special cases of the full model. For example, there may be a clade for which the best fitting models have no anagenetic rate of polyploidization (the rate = 0.0) and for which all polyploidization events are cladogenetic. To explore the entire space of all possible models of chromosome number evolution we constructed a reversible jump Markov chain Monte Carlo (Green 1995) that samples across models of different dimensionality, drawing samples from chromosome evolution models in proportion to their posterior probability and enabling Bayes factors for each model to be calculated. This approach incorporates model uncertainty by permitting model-averaged inferences that do not condition on a single model; we draw estimates of ancestral chromosome numbers and rates of chromosome evolution from all possible models weighted by their posterior probability. For general reviews of this approach to model averaging see Madigan and Raftery (1994), Hoeting et al. (1999), Kass and Raftery (1995), and for its use in phylogenetics see Posada and Buckley (2004). Averaging over all models has been shown to provide a better average predictive ability than conditioning on a single model (Madigan and Raftery 1994). Conditioning on a single model ignores model uncertainty, which can lead to an underestimation in the uncertainty of inferences made from that model (Hoeting et al. 1999). In our case, this can lead to overconfidence in estimates of ancestral chromosome numbers and chromosome evolution parameter value estimates.

Our motivation in developing these phylogenetic models of chromosome evolution is to determine the mode of chromosome number evolution; is chromosome evolution occurring primarily within lineages, primarily at lineage splitting, or in clade-specific combinations of both? By identifying how much of the pattern of chromosome number evolution is explained by anagenetic versus cladogenetic change, and by identifying the timing and location of possible chromosome speciation events over the phylogeny, the ChromoSSE models can help uncover how much of a role chromosome changes play in speciation. In this paper we first describe the ChromoSSE models of chromosome evolution and our Bayesian method of model selection, then we assess the models' efficacy by testing them with simulated datasets, particularly focusing on the impact of unobserved speciation events on inferences, and finally we apply the models to five empirical datasets that have been previously examined using other models of chromosome number evolution.

2.2 Methods

2.2.1 Models of Chromosome Evolution

In this section we introduce our class of probabilistic models of chromosome number evolution. We are interested in modeling the changes in chromosome number both within lineages (anagenetic evolution) and at speciation events (cladogenetic evolution). The anagenetic component of the model is a continuous-time Markov process similar to Mayrose et al. (2010) as described below. The cladogenetic changes are accounted for by a birth-death process similar to Maddison et al. (2007) and Goldberg and Igić (2012), except each type of cladogenetic chromosome event is given its own rate. Thus, the birth-death process has multiple speciation rates (one for each type of cladogenetic change) and a single constant extinction rate. Our models of chromosome number evolution can therefore be understood as a specific case of the Cladogenetic State change Speciation and Extinction (ClaSSE) model (Goldberg and Igić 2012), which integrates over all possible unobserved speciation events (due to lineages that were unsampled or have gone extinct) directly in the likelihood calculation of the observed chromosome counts and tree shape. To test the importance of accounting for unobserved speciation events we also briefly describe a version of the model that handles different cladogenetic event types as transition probabilities at each observed speciation event and ignores unobserved speciation events, similar to the dispersal-extinction-cladogenesis (DEC) models of geographic range evolution (Ree and Smith 2008).

Our implementation assumes chromosome numbers can take the value of any positive integer, however to limit the transition matrices to a reasonable size for likelihood calculations we follow Mayrose et al. (2010) in setting the maximum chromosome number C_m to $n + 10$, where n is the highest chromosome number in the observed data. Note that we allow this parameter to be set in our implementation. Hence, it is easily possible to test the impact of setting a specific value for the maximum chromosome count.

Our models contain a set of 6 free parameters for anagenetic chromosome number evolution, a set of 5 free parameters for cladogenetic chromosome number evolution, an extinction rate parameter, and a vector of C_m root frequencies of chromosome numbers, for a total of $12 + C_m$ free parameters. All of the 11 chromosome rate parameters can be removed (fixed to 0.0) except the cladogenetic no-change rate parameter. Thus, the class of chromosome number evolution models described here has a total of $2^{10} = 1024$ nested models of chromosome evolution.

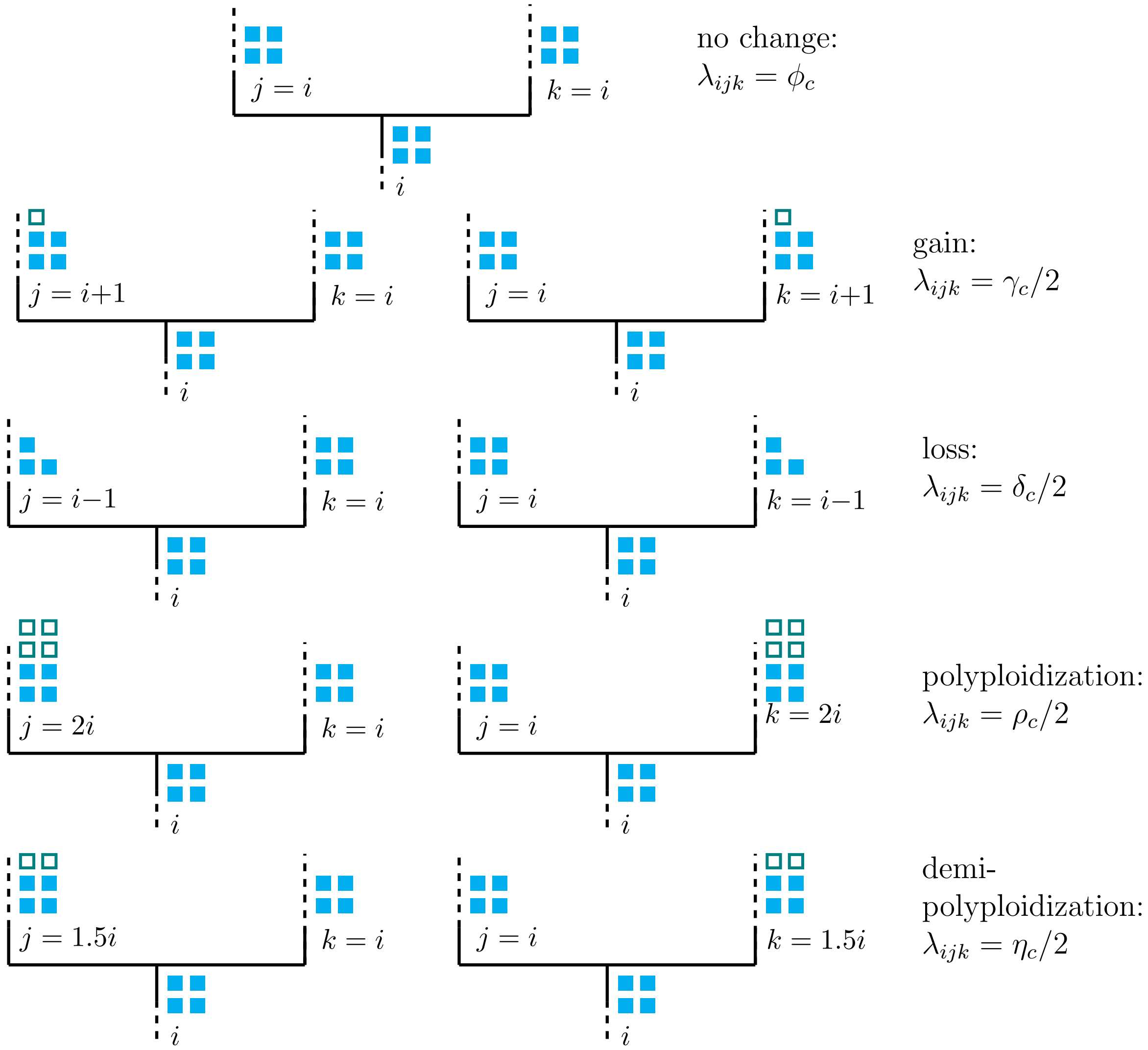

Figure 2.1: Modeled cladogenetic chromosome evolution events. At each speciation event 9 different cladogenetic events are possible. The rate of each type of speciation event is λ_{ijk} where i is the chromosome number before cladogenesis and j and k are the states of each daughter lineage immediately after cladogenesis. The dashed lines represent possible chromosomal changes within lineages that are modeled by the anagenetic rate matrix Q.

Chromosome evolution within lineages

Chromosome number evolution within lineages (anagenetic change) is modeled as a continuous-time Markov process similar to Mayrose et al. (2010). The continuous-time Markov process is described by an instantaneous rate matrix Q where the value of each element represents the instantaneous rate of change within a lineage from a genome of i chromosomes to a genome of j chromosomes. For all elements of Q in which either $i = 0$ or $j = 0$ we define $Q_{ij} = 0$. For the off-diagonal elements $i \neq j$ with positive values of i and j, Q is determined by:

$$Q_{ij} = \begin{cases} \gamma_a e^{\gamma_m (i-1)} & j = i+1, \\ \delta_a e^{\delta_m (i-1)} & j = i-1, \\ \rho_a & j = 2i, \\ \eta_a & j = 1.5i, \\ 0 & \text{otherwise}, \end{cases} \tag{2.1}$$

where γ_a, δ_a, ρ_a, and η_a are the rates of chromosome gains, losses, polyploidizations, and demi-polyploidizations. γ_m and δ_m are rate modifiers of chromosome gain and loss, respectively, that allow the rates of chromosome gain and loss to depend on the current number of chromosomes. This enables modeling scenarios in which the probability of fusion or fission events is positively or negatively correlated with the number of chromosomes. If the rate modifier $\gamma_m = 0$, then $\gamma_a e^{0(i-1)} = \gamma_a$. If the rate modifier $\gamma_m > 0$, then $\gamma_a e^{\gamma_m (i-1)} \geq \gamma_a$, and if $\gamma_m < 0$ then $\gamma_a e^{\gamma_m (i-1)} \leq \gamma_a$. These two rate modifiers replace the parameters λ_l and δ_l in Mayrose et al. (2010), which in their parameterization may result in negative transition rates. Here we chose to exponentiate γ_m and δ_m to ensure positive transition rates, and avoid ad hoc restrictions on negative transition rates that may induce unintended priors. Note that this assumes the rates of chromosome change can vary exponentially as a function of the current chromosome number, whereas Mayrose et al. (2010) assumes a linear function.

For odd values of i, we set $Q_{ij} = \eta/2$ for the two integer values of j resulting when $j = 1.5i$ was rounded up and down. We define the diagonal elements $i = j$ of Q as:

$$Q_{ii} = -\sum_{i \neq j}^{C_m} Q_{ij}. \tag{2.2}$$

The probability of anagenetically transitioning from chromosome number i to j along a branch of length t is then calculated by exponentiation of the instantaneous rate matrix:

$$P_{ij}(t) = e^{-Qt}. \tag{2.3}$$

Chromosome evolution at cladogenesis events

At each lineage divergence event over the phylogeny, nine different cladogenetic changes in chromosome number are possible (Figure 2.1). Each type of cladogenetic event occurs with the rate $\phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$, representing the cladogenesis rates of no change, chromosome gain, chromosome loss, polyploidization, and demi-polyploidization, respectively. The speciation rates λ for the birth-death process generating the tree are given in the form of a 3-dimensional matrix between the ancestral state i and the states of the two daughter lineages j and k. For all positive values of i, j, and k, we define:

$$\lambda_{ijk} = \begin{cases} \phi_c & j = k = i \\ \gamma_c/2 & j = i+1 \text{ and } k = i, \\ \gamma_c/2 & j = i \text{ and } k = i+1, \\ \delta_c/2 & j = i-1 \text{ and } k = i, \\ \delta_c/2 & j = i \text{ and } k = i-1, \\ \rho_c/2 & j = 2i \text{ and } k = i, \\ \rho_c/2 & j = i \text{ and } k = 2i, \\ \eta_c/2 & j = 1.5i \text{ and } k = i, \\ \eta_c/2 & j = i \text{ and } k = 1.5i, \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

so that the total speciation rate of the birth-death process λ_t is given by:

$$\lambda_t = \phi_c + \gamma_c + \delta_c + \rho_c + \eta_c. \tag{2.5}$$

Similar to the anagenetic instantaneous rate matrix described above, for odd values of i, we set $\lambda_{ijk} = \eta_c/4$ for the integer values of j and k resulting when $1.5i$ is rounded up and down. The extinction rate μ is constant over the tree and for all chromosome numbers.

Note that this model allows only a single chromosome number change event on a maximum of one of the daughter lineages at each cladogenesis event. Changes in both daughter lineages at cladogenesis are not allowed; at least one of the daughter lineages must inherit the chromosome number of the ancestor. The model also assumes that cladogenesis events are always strictly bifurcating and that there are no hard polytomies.

Likelihood Calculation Accounting for Unobserved Speciation

The likelihood of cladogenetic and anagenetic chromosome number evolution over a phylogeny is calculated using a set of ordinary differential equations similar to the Binary State Speciation and Extinction (BiSSE) model (Maddison et al. 2007). The BiSSE model was extended to incorporate cladogenetic changes by Goldberg and Igić (2012). Following Goldberg and Igić (2012), we define $D_{Ni}(t)$ as the probability that a lineage with chromosome number i at time t evolves into the observed clade N. We let $E_i(t)$ be the probability that a lineage with chromosome number i at time t goes extinct before the present, or is not sampled at the present. However, unlike the full ClaSSE model the extinction rate μ does not depend on the chromosome number i of the lineage. The differential equations for these two probabilities is given by:

$$\begin{aligned} \frac{dD_{Ni}(t)}{dt} = &-\left(\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu \right) D_{Ni}(t) \\ &+ \sum_{j=1}^{C_m} Q_{ij} D_{Nj}(t) + \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} \left(D_{Nk}(t) E_j(t) + D_{Nj}(t) E_k(t) \right) \end{aligned} \tag{2.6}$$

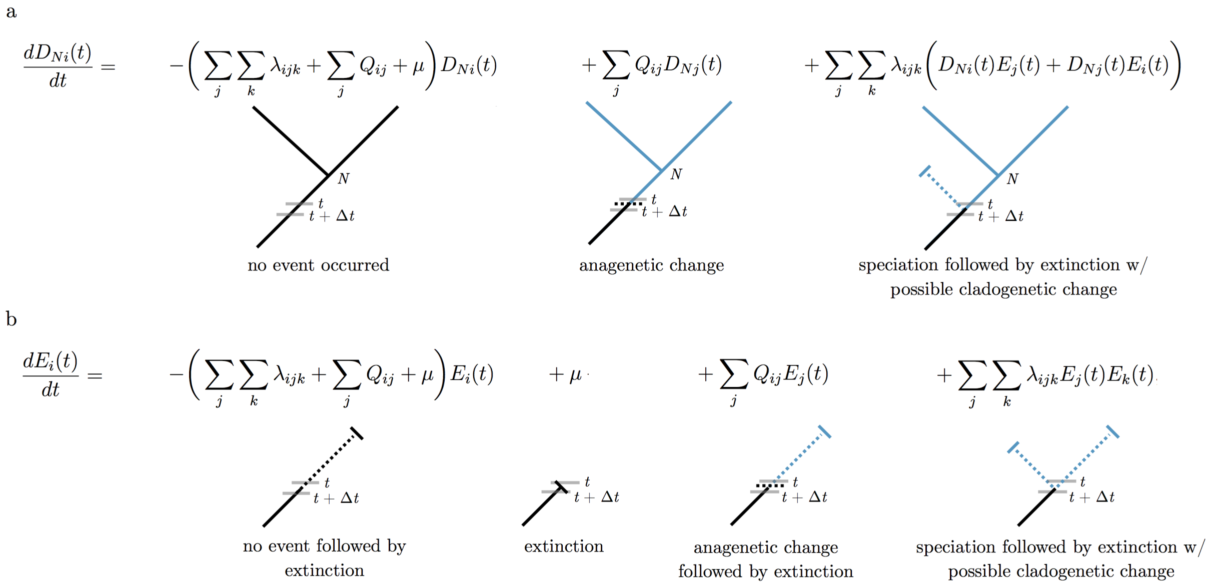
a
$\frac{dD_{Ni}(t)}{dt} =$
$-\left(\sum_j\sum_k \lambda_{ijk} + \sum_j Q_{ij} + \mu\right) D_{Ni}(t)$
$+\sum_j Q_{ij} D_{Nj}(t)$
$+\sum_j\sum_k \lambda_{ijk}\left(D_{Ni}(t)E_j(t) + D_{Nj}(t)E_i(t)\right)$
N
t
$t + \Delta t$
no event occurred
anagenetic change
speciation followed by extinction w/
possible cladogenetic change
b
$\frac{dE_i(t)}{dt} =$
$-\left(\sum_j\sum_k \lambda_{ijk} + \sum_j Q_{ij} + \mu\right) E_i(t)$
$+\mu$
$+\sum_j Q_{ij} E_j(t)$
$+\sum_j\sum_k \lambda_{ijk} E_j(t) E_k(t)$
no event followed by
extinction
extinction
anagenetic change
followed by extinction
speciation followed by extinction w/
possible cladogenetic change

Figure 2.2: Chromosome evolution through time. An illustration of chromosome evolution events that could occur during each time interval Δt along the branches of a phylogeny. Equations 2.6 and 2.7 (subfigures a and b, respectively) sum over each possible chromosome evolution event and are numerically integrated backwards through time over the phylogeny to calculate the likelihood. a) $D_{Ni}(t)$ is the probability that the lineage at time t evolves into the observed clade N. To calculate the change in this probability over Δt we sum over three possibilities: no event occurred, an anagenetic change in chromosome number occurred, or a speciation event with a possible cladogenetic chromosome change occurred followed by an extinction event on one of the two daughter lineages. b) $E_i(t)$ is the probability that the lineage goes extinct or is not sampled at the present. To calculate the change in this probability over Δt we sum over four possibilities: no event occurred followed eventually by extinction, extinction occurred, an anagenetic change occurred followed by extinction, or a speciation event with a possible cladogenetic change occurred followed by extinction of both daughter lineages.

$$\frac{dE_i(t)}{dt} = -\left(\sum_{j=1}^{C_m}\sum_{k=1}^{C_m} \lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu\right) E_i(t) + \mu + \sum_{j=1}^{C_m} Q_{ij} E_j(t) + \sum_{j=1}^{C_m}\sum_{k=1}^{C_m} \lambda_{ijk} E_j(t) E_k(t), \quad (2.7)$$

where λ_{ijk} for each possible cladogenetic event is given by equation 2.4, and the rates of anagenetic changes Q_{ij} are given by equation 2.1. See Figure 2.2 for an explanation of equations 2.6 and 2.7.

The differential equations above have no known analytical solution. Therefore, we numerically integrate the equations for every arbitrarily small time interval moving along each branch from the tip of the tree towards the root. When a node l is reached, the probability of it being in state i is calculated by combining the probabilities of its descendant nodes m

and n as such:

$$D_{li}(t) = \sum_{j=1}^{C_m} \sum_{k=1}^{C_m} \lambda_{ijk} D_{mj}(t) D_{nk}(t), \tag{2.8}$$

where again λ_{ijk} for each possible cladogenetic event is given by equation 2.4. Letting D denote a set of observed chromosome counts, Ψ an observed phylogeny, and θ_q a particular set of chromosome evolution model parameters, then the likelihood for the model parameters θ_q is given by:

$$P(D, \Psi | \theta_q) = \sum_{i=1}^{C_m} \pi_i D_{0i}(t), \tag{2.9}$$

where π_i is the root frequency of chromosome number i and $D_{0i}(t)$ is the likelihood of the root node being in state i conditional on having given rise to the observed tree Ψ and the observed chromosome counts D.

Initial Conditions

The initial conditions for each observed lineage at time $t = 0$ for the extinction probabilities described by equation 2.7 are $E_i(0) = 1 - \rho_s$ for all i where ρ_s is the sampling probability of including that lineage. For lineages with an observed chromosome number of i, the initial condition is $D_{Ni}(0) = \rho_s$. The initial condition for all other chromosome numbers j is $D_{Nj}(0) = 0$.

Likelihood Calculation Ignoring Unobserved Speciation

To test the effect of unobserved speciation events on inferences of chromosome number evolution we also implemented a version of the model described above that only accounts for observed speciation events. At each lineage divergence event over the phylogeny, the probabilities of cladogenetic chromosome number evolution $P(\{j, k\}|i)$ are given by the simplex $\{\phi_p, \gamma_p, \delta_p, \rho_p, \eta_p\}$, where $\phi_p, \gamma_p, \delta_p, \rho_p$, and η_p represent the probabilities of no change, chromosome gain, chromosome loss, polyploidization, and demi-polyploidization, respectively. This approach does not require estimating speciation or extinction rates.

Here, we calculate the likelihood of chromosome number evolution over a phylogeny using Felsenstein's pruning algorithm (Felsenstein 1981) modified to include cladogenetic probabilities similar to models of biogeographic range evolution (Landis et al. 2013; Landis 2017). Let D again denote a set of observed chromosome counts and Ψ represent an observed phylogeny where node l has descendant nodes m and n. The likelihood of chromosome number evolution at node l conditional on node l being in state i and θ_q being a particular set of chromosome evolution model parameter values is given by:

$$P_l(D, \Psi | i, \theta_q) = \underbrace{\sum_{j=1}^{C_m} \sum_{k=1}^{C_m} P(\{j, k\}|i)}_{\text{cladogenetic}} \underbrace{\left[\sum_{j_e=1}^{C_m} P_{jj_e}(t_m) P_m(D, \Psi | j_e, \theta_q) \right] \left[\sum_{k_e=1}^{C_m} P_{kk_e}(t_n) P_n(D, \Psi | k_e, \theta_q) \right]}_{\text{anagenetic}}, \tag{2.10}$$

where the length of the branches between l and m is t_m and between l and n is t_n. The state at the end of these branches near nodes m and n is j_e and k_e, respectively. The state at the beginning of these branches, where they meet at node l, is j and k respectively. The cladogenetic term sums over the probabilities $P(\{j,k\}|i)$ of all possible cladogenetic changes from state i to the states j and k at the beginning of each daughter lineage. The anagenetic term of the equation is the product of the probability of changes along the branches from state j to state j_e and state k to state k_e (given by equation 2.3) and the likelihood of the tree above node l recursively computed from the tips.

The likelihood for the model parameters θ_q is given by:

$$P(D,\Psi|\theta_q) = \sum_{i=1}^{C_m} \pi_i P_0(D,\Psi|i,\theta_q), \tag{2.11}$$

where $P_0(D,\Psi|i,\theta_q)$ is the conditional likelihood of the root node being in state i and π_i is the root frequency of chromosome number i.

Estimating Parameter Values and Ancestral States

For any given tree with a set of observed chromosome counts, there exists a posterior distribution of model parameter values and a set of probabilities for the ancestral chromosome numbers at each internal node of the tree. Let $P(s_i,\theta_q|D,\Psi)$ denote the joint posterior probability of θ_q and a vector of specific ancestral chromosome numbers s_i given a set of observed chromosome counts D and an observed tree Ψ. The posterior is given by Bayes' rule:

$$P(s_i,\theta_q,|D,\Psi) = \frac{P(D,\Psi|s_i,\theta_q)P(s_i|\theta_q)P(\theta_q)}{\int\limits_{\theta}\sum\limits_{s=1}^{C_m} P(D,\Psi|s,\theta)P(s|\theta)P(\theta)d\theta}. \tag{2.12}$$

Here, $P(s_i|\theta_q)$ is the prior probability of the ancestral states s conditioned on the model parameters θ_q, and $P(\theta_q)$ is the joint prior probability of the model parameters.

In the denominator of equation 2.12 we integrate over all possible values of θ and sum over all possible ancestral chromosome numbers s. Since θ is a vector of $12 + C_m$ parameters and s is a vector of $n-1$ ancestral states where n is the number of observed tips in the phylogeny, the denominator of equation 2.12 requires a high dimensional integral and an extremely large summation that is impossible to calculate analytically. Instead we use Markov chain Monte Carlo methods (Metropolis et al. 1953; Hastings 1970) to estimate the posterior probability distribution in a computationally efficient manner.

Ancestral states are inferred using a two-pass tree traversal procedure as described in Pupko et al. (2000), and previously implemented in a Bayesian framework by Huelsenbeck and Bollback (2001) and Pagel et al. (2004). First, partial likelihoods are calculated during the backwards-time post-order tree traversal in equations 2.6 and 2.7. Joint ancestral states are then sampled during a pre-order tree traversal in which the root state is first drawn from the marginal likelihoods at the root, and then states are drawn for each descendant node conditioned on the state at the parent node until the tips are reached. Again, we must numerically integrate over a system of differential equations during this root-to-tip tree traversal. This integration, however, is performed in forward-time, thus the set of ordinary differential equations must be slightly altered since our models of chromosome number evolution are not time reversible. Accordingly, we calculate:

$$\begin{aligned}\frac{dD_{Ni}(t)}{dt} = &-\left(\sum_{j=1}^{C_m}\sum_{k=1}^{C_m}\lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu\right) D_{Ni}(t) \\ &+ \sum_{j=1}^{C_m} Q_{ji} D_{Nj}(t) + D_{Nj}(t)E_k(t)\left(\sum_{j=1}^{C_m}\sum_{k=1}^{C_m}\lambda_{jik} + \sum_{j=1}^{C_m}\sum_{k=1}^{C_m}\lambda_{jki}\right) \quad (2.13)\end{aligned}$$

$$\begin{aligned}\frac{dE_i(t)}{dt} = &\left(\sum_{j=1}^{C_m}\sum_{k=1}^{C_m}\lambda_{ijk} + \sum_{j=1}^{C_m} Q_{ij} + \mu\right) E_i(t) \\ &- \mu - \sum_{j=1}^{C_m} Q_{ij}E_j(t) - \sum_{j=1}^{C_m}\sum_{k=1}^{C_m}\lambda_{ijk}E_j(t)E_k(t), \quad (2.14)\end{aligned}$$

during the forward-time root-to-tip pass to draw ancestral states from their joint distribution conditioned on the model parameters and observed chromosome counts. See Freyman and Höhna (2017b) for a derivation of these equations. For more details and validation of our method to estimate ancestral states, please see Section 2.5.1.

Priors

Model parameter priors are listed in Table 2.1. Our implementation allows all priors to be easily modified so that their impact on results can be effectively assessed. Priors for anagenetic rate parameters are given an exponential distribution with a mean of $2/\Psi_l$ where Ψ_l is the length of the tree Ψ. This corresponds to a mean rate of two events over the observed tree. The priors for the rate modifiers γ_m and δ_m are assigned a uniform distribution with the range $-3/C_M$ to $3/C_m$. This sets minimum and maximum bounds on the amount the rate modifiers can affect the rates of gain and loss at the maximum chromosome number to $\gamma_a e^{-3} = \gamma_a 0.050$ and $\gamma_a e^3 = \gamma_a 20.1$, and $\delta_a e^{-3} = \delta_a 0.050$ and $\delta_a e^3 = \delta_a 20.1$, respectively.

The speciation rates are drawn from an exponential prior with a mean equal to an estimate of the net diversification rate $\hat{d}$. Under a constant rate birth-death process not conditioning on survival of the process, the expected number of lineages at time t is given by:

$$E(N_t) = N_0 e^{td}, \quad (2.15)$$

where N_0 is the number of lineages at time 0 and d is the net diversification rate $\lambda - \mu$ (Nee et al. 1994b; Höhna 2015). Therefore, we estimate $\hat{d}$ as:

$$\hat{d} = (\ln N_t - \ln N_0)/t, \quad (2.16)$$

where N_t is the number of lineages in the observed tree that survived to the present, t is the age of the root, and $N_0 = 2$.

The extinction rate μ is given by:

$$\mu = r \times \lambda_t = r \times (\phi_c + \gamma_c + \delta_c + \rho_c + \eta_c), \quad (2.17)$$

where λ_t is the total speciation rate and r is the relative extinction rate. The relative extinction rate r is assigned a uniform (0,1) prior distribution, thus forcing the extinction rate to be smaller than the total speciation rate. The root frequencies of chromosome numbers π are drawn from a flat Dirichlet distribution.

Table 2.1: Model parameter names and prior distributions. See the main text for complete description of model parameters and prior distributions. Ψ_l represents the length of tree Ψ and C_m is the maximum chromosome number allowed.

	Parameter	X	$f(X)$
Anagenetic	Chromosome gain rate	γ_a	Exponential($\lambda = \Psi_l/2$)
	Chromosome loss rate	δ_a	Exponential($\lambda = \Psi_l/2$)
	Polyploidization rate	ρ_a	Exponential($\lambda = \Psi_l/2$)
	Demi-polyploidization rate	η_a	Exponential($\lambda = \Psi_l/2$)
	Linear component of chromosome gain rate	γ_m	Uniform($-3/C_m, 3/C_m$)
	Linear component of chromosome loss rate	δ_m	Uniform($-3/C_m, 3/C_m$)
Cladogenetic	No change	ϕ_c	Exponential($\lambda = 1/\hat{d}$)
	Chromosome gain	γ_c	Exponential($\lambda = 1/\hat{d}$)
	Chromosome loss	δ_c	Exponential($\lambda = 1/\hat{d}$)
	Polyploidization	ρ_c	Exponential($\lambda = 1/\hat{d}$)
	Demi-polyploidization	η_c	Exponential($\lambda = 1/\hat{d}$)
Other	Root frequencies	π	Dirichlet(1,...,1)
	Relative-extinction	r	Uniform(0, 1)

2.2.2 Model Uncertainty and Selection

Model Averaging

To account for model uncertainty we calculate the posterior density of chromosome evolution model parameters θ without conditioning on any single model of chromosome evolution. For each of the 1024 chromosome models M_k, where $k = 1, 2, \ldots, 1024$, the posterior distribution of θ is

$$P(\theta|D) = \sum_{k=1}^{K} P(\theta|D, M_k) P(M_k|D). \tag{2.18}$$

Here we average over the posterior distributions conditioned on each model weighted by the model's posterior probability. We assume an equal prior probability for each model $P(M_k) = 2^{-10}$.

Reversible Jump Markov Chain Monte Carlo

To sample from the space of all possible chromosome evolution models, we employ reversible jump MCMC (Green 1995). This algorithm draws samples from parameter spaces of differing dimensions, and in stationarity samples each model in proportion to its posterior probability. This permits inference of each model's fit to the data while simultaneously accounting for model uncertainty.

Our reversible jump MCMC moves between models of different dimensions using augment and reduce moves (Huelsenbeck et al. 2000; Pagel and Meade 2006; May et al. 2016). The reduce move proposes that a parameter should be removed from the current model by setting its value to 0.0, effectively disallowing that class of evolutionary event. Augment moves reverse reduce moves by allowing the parameter to once again have a non-zero value. Both augment and reduce moves operate on all chromosome rate parameters except for ϕ_c the rate of no cladogenetic change. Thus the least complex model the MCMC can sample from is one in which $\phi_c > 0.0$ and all other chromosome rate parameters are set to 0.0, corresponding to a model of no chromosomal changes over the phylogeny. The prior probability of reducing or augmenting model M_k is $P_r(M_k) = P_a(M_k) = 0.5$.

Bayes Factors

In some cases we wish to compare the fit of models to summarize the mode of evolution within a clade. Bayes factors (Kass and Raftery 1995) compare the evidence between two competing models M_i and M_j

$$B_{ij} = \frac{P(D|M_i)}{P(D|M_j)} = \frac{P(M_i|D)}{P(M_j|D)} / \frac{P(M_i)}{P(M_j)}. \tag{2.19}$$

In words, the Bayes factor B_{ij} is given by the ratio of the posterior odds to the prior odds of the two models. Unlike other methods of model selection such as Akaike Information Criterion (AIC; Akaike 1974) and the Bayesian Information Criterion (BIC; Schwarz 1978), Bayes factors take into account the full posterior densities of the model parameters and do not rely on point estimates. Furthermore AIC and BIC ignore the priors assigned to parameters, whereas Bayes factors penalizes parameters based on the informativeness of the prior. If the prior is informative but overlaps little with the likelihood it is penalized more than a diffuse uninformative prior that allows the parameter to take on whatever value is informed by the data (Xie et al. 2011).

2.2.3 Implementation

The model and MCMC analyses described here are implemented in `C++` in the software `RevBayes` (Höhna et al. 2016). In Section 2.5.1 we validated our SSE likelihood calculations and ancestral state estimates against those of the `R` package `diversitree` (FitzJohn 2012). `Rev` scripts that specify the chromosome number evolution model (ChromoSSE) described here as a probabilistic graphical model (Höhna et al. 2014a) and run the empirical analyses in `RevBayes` are available at `http://github.com/wf8/ChromoSSE`. The `RevGadgets` `R` package (available at `https://github.com/revbayes/RevGadgets`) contains functions to summarize results and generate plots of inferred ancestral chromosome numbers over a phylogeny.

The MCMC proposals used are outlined in Section 2.5.2. Aside from the reversible jump MCMC proposals described above, all other proposals are standard except for the ElementSwapSimplex move operated on the Dirichlet distributed root frequencies parameter. This move randomly selects two elements r_1 and r_2 from the root frequencies vector and swaps their values. The reverse move, swapping the original values of r_1 and r_2 back, will have the same probability as the initial move since r_1 and r_2 were drawn from a uniform distribution.

Thus, the Hasting ratio is 1 and the ElementSwapSimplex move is a symmetric Metropolis move.

2.2.4 Simulations

We conducted a series of simulations to: 1) test the effect of unobserved speciation events due to extinction on chromosome number estimates when using a model that does not account for unobserved speciation, 2) compare the accuracy of models of chromosome evolution that account for unobserved speciation versus those that do not, 3) test the effect of jointly estimating speciation and extinction rates with chromosome number evolution, 4) test for identifiability of cladogenetic parameters, and 5) test the effect of incomplete sampling of extant lineages on ancestral chromosome number estimates. We will refer to each of the 5 simulations above as experiment 1, experiment 2, experiment 3, experiment 4, and experiment 5. Detailed descriptions of each experiment and the methods used to simulate trees and chromosome counts are in Section 2.5.3.

For all 5 experiments, MCMC analyses were run for 5000 iterations, where each iteration consisted of 28 different moves in a random move schedule with 79 moves per iteration (see Section 2.5.2). Samples were drawn with each iteration, and the first 1000 samples were discarded as burn in. Effective sample sizes (ESS) for all parameters in all simulation replicates were over 200, and the mean ESS values of the posterior for the replicates was 1470.3. See Section 2.5.4 for more on convergence of simulation replicates. To perform all 5 experiments 2100 independent MCMC analyses were run requiring a total of 89170.6 CPU hours on the Savio computational cluster at the University of California, Berkeley.

Summarizing Simulation Results

To summarize the results of our simulations, we measured the accuracy of ancestral state estimates as the percent of simulation replicates in which the true root chromosome number 8 was found to be the maximum a posteriori (MAP) estimate. To evaluate the uncertainty of the simulations, we calculated the mean posterior probability of root chromosome number for the simulation replicates that correctly found 8 to be the MAP estimate. We also calculated the proportion of simulation replicates for which the true model of chromosome number evolution used to simulate the data (as given by the table in Section 2.5.3) was estimated to be the MAP model, and calculated the mean posterior probabilities of the true model. To compare the accuracy of model averaged parameter value estimates we calculated coverage probabilities. Coverage probabilities are the percentage of simulation replicates for which the true parameter value falls within the 95% highest posterior density (HPD). High accuracy is shown when coverage probabilities approach 1.0.

2.2.5 Empirical Data

Phylogenetic data and chromosomes counts from five plant genera were analyzed (see Table 2.2). Like in Mayrose et al. (2010) we assumed each species had a single cytotype, however polymorphism could be accounted for by a vector of probabilities for each chromosome count. Sequence data for *Aristolochia* was downloaded from TreeBASE (Vos et al. 2010) study ID

1586. Sequences for *Helianthus*, *Mimulus* sensu lato, and *Primula* were downloaded directly from GenBank (Benson et al. 2005), reconstructing the sequence matrices from Timme et al. (2007), Beardsley et al. (2004), and Guggisberg et al. (2009). For each of these four datasets phylogenetic analyses were performed with all gene regions concatenated and unpartitioned, assuming the general time-reversible (GTR) nucleotide substitution model (Tavaré 1986; Rodriguez et al. 1990) with among-site rate variation modeled using a discretized gamma distribution (Yang 1994) with four rate categories. Since divergence time estimation in years is not the objective of this study, and only relative branching times are needed for our models of chromosome number evolution, a birth-death tree prior was used with a fixed root age of 10.0 time units. The MCMC analyses were performed in `RevBayes`, and were sampled every 100 iterations and run for a total of 400000 iterations, with samples from the first 100000 iterations discarded as burnin. Convergence was assessed by ensuring that the effective sample size for all parameters was over 200. The maximum a posteriori tree was calculated and used for further chromosome evolution analyses. For *Carex* section *Spirostachyae* the time calibrated tree from Escudero et al. (2010) was used.

Ancestral chromosome numbers and chromosome evolution model parameters were then estimated for each of the five clades. Since testing the effect of incomplete taxon sampling on chromosome evolution inference of the empirical datasets was not a goal of this work, we focus here on results using a taxon sampling fraction ρ_s of 1.0 (though see the Discussion section for more on this). MCMC analyses were run in `RevBayes` for 11000 iterations, where each iteration consisted of 28 different Metropolis-Hastings moves in a random move schedule with 79 moves per iteration (see Section 2.5.2). Samples were drawn each iteration, and the first 1000 samples were discarded as burn in. Effective sample sizes for all parameters were over 200. For all datasets except *Primula* we used priors as outlined in Table 2.1. To demonstrate the flexibility of our Bayesian implementation and its capacity to incorporate prior information we used an informative prior for the root chromosome number in the *Primula* section *Aleuritia* analysis. Our dataset for *Primula* section *Aleuritia* also included samples from *Primula* sections *Armerina* and *Sikkimensis*. Since we were most interested in estimating chromosome evolution within section *Aleuritia*, we used an informative Dirichlet prior $\{1, ..., 1, 100, 1....1\}$ (with 100 on the 11th element) to bias the root state towards the reported base number of *Primula* $x = 11$ (Conti et al. 2000). Note all priors can be easily modified in our implementation, thus the impact of priors can be efficiently tested.

2.3 Results

2.3.1 Simulations

General Results

In all simulations, the true model of chromosome number evolution was infrequently estimated to be the MAP model ($< 36\%$ of replicates), and when it was the posterior probability of the MAP model was very low (< 0.12; Table 2.3). We found that the accuracy of root chromosome number estimation was similar whether the process that generated the simulated data was cladogenetic-only or anagenetic-only (Tables 2.3 and 2.4). However, when the

Table 2.2: Empirical data sets analysed.

Clade	Study	Gene region	Alignment length (bp)	Number of OTUs	Haploid chromosome numbers range
Aristolochia	Ohi-Toma et al. (2006)	matK	1268	34	3 - 16
Carex section *Spirostachyae*	Escudero et al. (2010)	ITS, trnK intron	see Escudero et al. (2010)	24	30 - 42
Helianthus	Timme et al. (2007)	ETS	3085	102	17 - 51
Mimulus sensu lato	Beardsley et al. (2004)	trnL intron, ETS, ITS	2210	115	8 - 46
Primula section *Aleuritia*	Guggisberg et al. (2009)	rpl16 intron, rps16 intron, trnL intron, trnL-trnF spacer, trnT-trnL spacer, trnD-trnT region	5705	56	9 - 36

data was simulated under a process that included both cladogenetic and anagenetic evolution we found a decrease in accuracy in the root chromosome number estimates in all cases.

Experiment 1 Results

The presence of unobserved speciation in the process that generated the simulated data decreased the accuracy of ancestral state estimates (Figure 2.3, Table 2.3). Similarly, uncertainty in root chromosome number estimates increased with unobserved speciation (lower mean posterior probabilities; Table 2.3). The accuracy of parameter value estimates as measured by coverage probabilities was similar (results not shown).

Experiment 2 Results

When comparing estimates from ChromoSSE that account for unobserved speciation to estimates from the non-SSE model that does not account for unobserved speciation, we found that the accuracy in estimating model parameter values was mostly similar, though for some cladogenetic parameters there was higher accuracy with the model that did account for unobserved speciation (ChromoSSE; Figure 2.4). For both models estimates of anagenetic parameters were more accurate than estimates of cladogenetic parameters when the true generating model included cladogenetic changes.

We found that ChromoSSE had more uncertainty in root chromosome number estimates (lower mean posterior probabilities) compared to the non-SSE model that did not account for unobserved speciation. Similarly, the root chromosome number was estimated with slightly lower accuracy (Table 2.4).

Experiment 3 Results

We found that jointly estimating speciation and extinction rates with chromosome number evolution using ChromoSSE slightly decreased the accuracy of root chromosome number estimates, and further it increased the uncertainty of the inferred root chromosome number (as reflected in lower mean posterior probabilities; Table 2.4). Fixing the speciation and extinction rates to their true value removed much of the increased uncertainty associated with using a model that accounts for unobserved speciation (Table 2.4).

Experiment 4 Results

Under simulation scenarios that had cladogenetic changes but no anagenetic changes, we found that ChromoSSE overestimated anagenetic parameters and underestimated cladogenetic parameters (Figure 2.5 A), which explains the lower coverage probabilities of cladogenetic parameters reported above for experiment 2 (Figure 2.4). When anagenetic parameters were fixed to 0.0 cladogenetic parameters were no longer underestimated (Figure 2.5 A), and the coverage probabilities of cladogenetic parameters increased slightly (Figure 2.5 B).

Experiment 5 Results

We found that incomplete sampling of extant lineages had a minor effect on the accuracy of ancestral chromosome number estimates (Figure 2.6). Accuracy only slightly decreased as the percentage of extant lineages sampled declined from 100% to 50%, and decreased more rapidly when the percentage went to 10%. As measured by the proportion of simulation replicates that inferred the MAP root chromosome number to be the true root chromosome number, the accuracy of ancestral states estimated under ChromoSSE declined from 0.80 accuracy at 100% taxon sampling to 0.69 at 10% taxon sampling. Essentially no difference in accuracy was detected between the non-SSE model that does not take unobserved speciation into account and ChromoSSE. Furthermore, little difference in accuracy was detected using ChromoSSE with the taxon sampling probability ρ_s set to 1.0 compared to ChromoSSE with ρ_s set to the true value (0.1, 0.5, or 1.0; Figure 2.6).

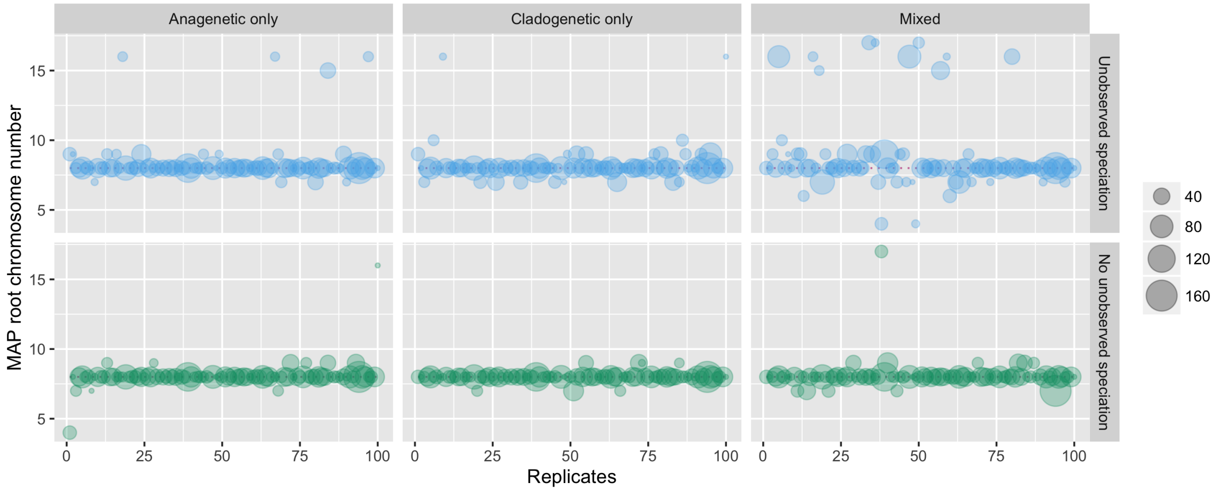

Figure 2.3: Experiment 1 results: the effect of unobserved speciation events on the maximum a posteriori (MAP) estimates of root chromosome number. Model averaged MAP estimates of the root chromosome number for 100 replicates of each simulation type on datasets that included unobserved speciation and datasets that did not include unobserved speciation. Each circle represents a simulation replicate, where the size of the circle is proportional to the number of lineages that survived to the present (the number of extant tips in the tree). The true root chromosome number used to simulate the data was 8 and is marked with a pink dotted line.

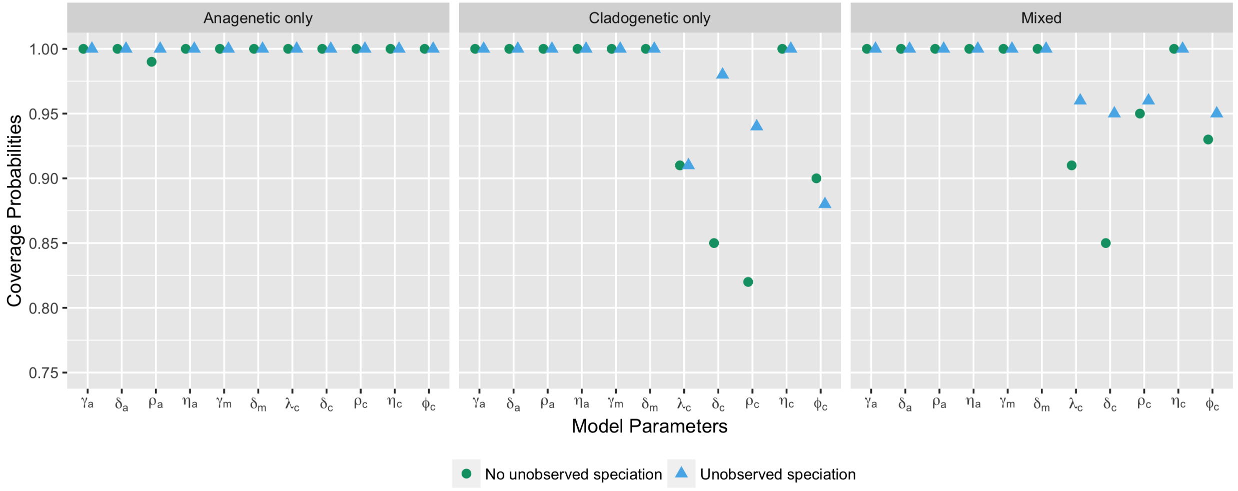

Figure 2.4: Experiment 2 results: the effect of using a model that accounts for unobserved speciation on coverage probabilities of chromosome model parameters. Each point represents the proportion of simulation replicates for which the 95% HPD interval contains the true value of the model parameter. Coverage probabilities of 1.00 mean perfect coverage. The circles represent coverage probabilities for estimates made using the non-SSE model that does not account for unobserved speciation, and the triangles represent coverage probabilities for estimates made using ChromoSSE that does account for unobserved speciation.

Table 2.3: Experiment 1 results: the effect of ignoring unobserved speciation events on chromosome evolution estimates. Regardless of the true mode of chromosome evolution, the presence of unobserved speciation events in the process that generated the simulated data decreased accuracy in estimating the true root state. The columns from left to right are: 1) an indication of whether or not the data was simulated with a process that included unobserved speciation, 2) the true mode of chromosome evolution used to simulate the data, (for description see main text and Section 2.5.3), 3) the percent of simulation replicates in which the true chromosome number at the root used to simulate the data was found to be the maximum a posteriori (MAP) estimate, 4) the mean posterior probability of the MAP estimate of the true root chromosome number, 5) the percent of simulation replicates in which the true model used to simulate the data was also found to be the MAP model, and 6) the mean posterior probability of the MAP estimate of the true model.

Unobserved Speciation Events Included When Simulating Data?	Mode of Evolution Used to Simulate Data	True Root State Estimated (%)	Mean Posterior of True Root State	True Model Estimated (%)	Mean Posterior of True Model
No	Cladogenetic	93	0.92	13	0.10
No	Anagenetic	89	0.91	31	0.12
No	Mixed	88	0.84	0	0.0
Yes	Cladogenetic	78	0.87	15	0.09
Yes	Anagenetic	83	0.91	36	0.12
Yes	Mixed	62	0.80	2	0.10

Table 2.4: Experiments 2 and 3 results: the effects of using a model that accounts for unobserved speciation and of jointly estimating diversification rates on ancestral chromosome number estimates. This table compares estimates of chromosome evolution using a non-SSE model that does not account for unobserved speciation events with ChromoSSE that does account for unobserved speciation events (Experiment 2), and compares estimates of chromosome evolution when jointly estimated with speciation and extinction rates versus when the true speciation and extinction rates are given (Experiment 3). Regardless of the true mode of chromosome evolution, the use of a model that accounts for unobserved speciation increases uncertainty in root state estimates. The columns from left to right are: 1) an indication of which experiment the results pertain to, 2) an indication of whether or not the estimates were made with ChromoSSE (that accounts for unobserved speciation), 3) whether diversification rates were jointly estimated with chromosome evolution, 4) the percent of simulation replicates in which the true chromosome number at the root used to simulate the data was found to be the MAP estimate, 5) the mean posterior probability of the MAP estimate of the true root chromosome number.

Experiment #	Estimates Made w/ Model That Accounted for Unobserved Speciation?	Speciation and Extinction Rates Jointly Estimated?	Mode of Evolution Used to Simulate Data	True Root State Estimated (%)	Mean Posterior of True Root State
2	No	No	Cladogenetic	78	0.87
2	No	No	Anagenetic	83	0.91
2	No	No	Mixed	62	0.80
2 & 3	Yes	Yes	Cladogenetic	78	0.81
2 & 3	Yes	Yes	Anagenetic	80	0.86
2 & 3	Yes	Yes	Mixed	61	0.72
3	Yes	No	Cladogenetic	78	0.84
3	Yes	No	Anagenetic	83	0.90
3	Yes	No	Mixed	62	0.76

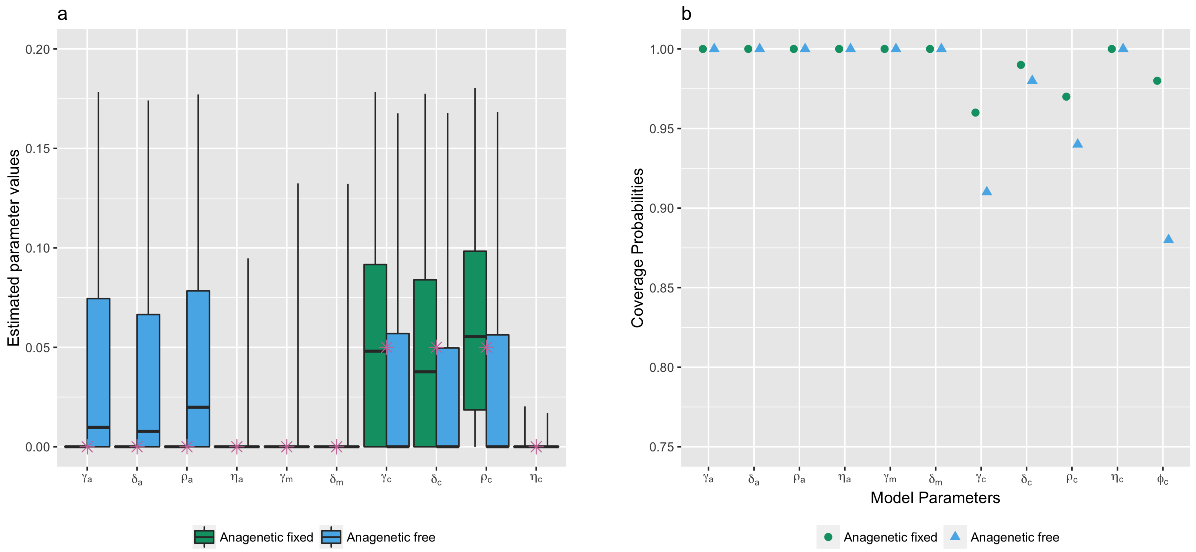

Figure 2.5: Experiment 4 results: testing identifiability of cladogenetic parameters under ChromoSSE. a) Chromosome parameter value estimates from 100 simulation replicates under a simulation scenario with no anagenetic changes (cladogenetic only). The stars represent true values. The box plots compare parameter estimates made when anagenetic parameters were fixed to 0 to estimates made when all parameters were free. When all parameters were free the anagenetic parameters were overestimated and cladogenetic parameters were underestimated. When the anagenetic parameters were fixed to 0 the estimates for the cladogenetic parameters were more accurate. b) Coverage probabilities of chromosome evolution parameters under the cladogenetic only model of chromosome evolution. The accuracy of cladogenetic parameter estimates increased when anagenetic parameters were fixed to 0.

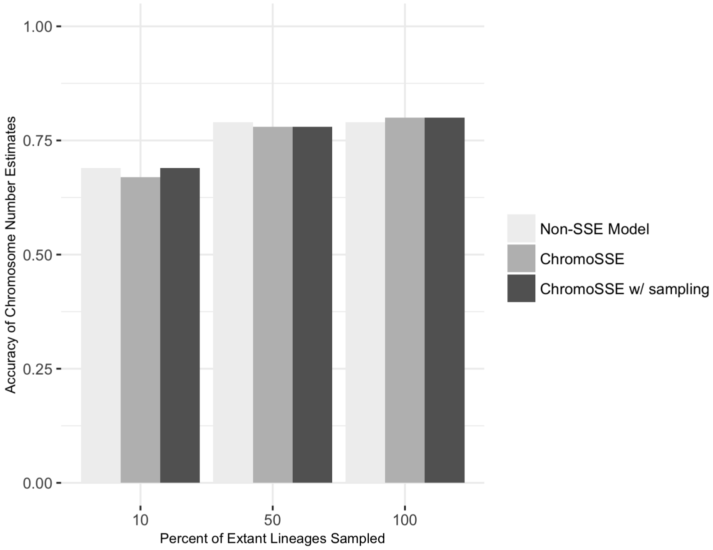

Figure 2.6: Experiment 5 results: the effect of incomplete sampling. The accuracy of ancestral chromosome number estimates slightly declined as the percentage of sampled extant lineages decreased from 100% to 50%, and decreased more quickly once the percentage of extant lineages decreased to 10%. There was little difference between the non-SSE model (light grey) that does not take into account unobserved speciation and ChromoSSE (medium and dark grey) which does take into account unobserved speciation. Furthermore, little difference in accuracy was detected using ChromoSSE with the taxon sampling probability ρ_s set to 1.0 (medium grey) and with ρ_s set to the true value (0.1, 0.5, or 1.0; dark grey). The accuracy of chromosome number estimates was measured by the proportion of simulation replicates for which the estimated MAP root chromosome number corresponded with the true chromosome number used to simulate the data.

2.3.2 Empirical Data

Model averaged MAP estimates of ancestral chromosome numbers for each of the five empirical datasets are show in Figures 2.7, 2.8, 2.9, 2.10, and 2.11. The mean model-averaged chromosome number evolution parameter value estimates for the empirical datasets are reported in Table 2.5. Posterior probabilities for the MAP model of chromosome number evolution were low for all datasets, varying between 0.04 for *Carex* section *Spirostachyae* and 0.21 for *Helianthus* (Table 2.6). Bayes factors supported unique, clade-specific combinations of anagenetic and cladogenetic parameters for all five datasets (Table 2.6). None of the clades had support for purely anagenetic or purely cladogenetic models of chromosome evolution.

The ancestral state reconstructions for *Aristolochia* were highly similar to those found by Mayrose et al. (2010). We found a moderately supported root chromosome number of 8 (posterior probability 0.45), and a polyploidization event on the branch leading to the *Isotrema* clade which has a base chromosome number of 16 with high posterior probability (0.88; Figure 2.7). On the branch leading to the main *Aristolochia* clade we found a dysploid loss of a single chromosome. Overall, we estimated moderate rates of anagenetic dysploid and polyploid changes, and the rates of cladogenetic change were 0 except for a moderate rate of cladogenetic dysploid loss (Tables 2.5). There was only one cladogenetic change inferred in the MAP ancestral state reconstruction, which was a recent possible dysploid speciation event that split the sympatric west-central Mexican species *Aristolochia tentaculata* and *A. taliscana*.

In *Helianthus*, on the other hand, we found high rates of cladogenetic polyploidization, and low rates of anagenetic change (Tables 2.5). 12 separate possible polyploid speciation events were identified over the phylogeny (Figure 2.8), and cladogenetic polyploidization made up 16% of all observed and unobserved speciation events. Bayes factors gave very strong support for models that included cladogenetic polyploidization as well as anagenetic demi-polyploidization (Table 2.6), the latter explaining the frequent anagenetic transitions from 34 to 51 chromosomes found in the MAP ancestral state reconstruction. The well supported root chromosome number of 17 (posterior probability 0.91) corresponded with the findings of Mayrose et al. (2010).

As opposed to the *Helianthus* results, the *Carex* section *Spirostachyae* estimates had very low rates of polyploidization and instead had high rates of cladogenetic dysploid change (Tables 2.5). An estimated 36.9% of all observed and unobserved speciation events included a cladogenetic gain or loss of a single chromosome. Overall, the rates of anagenetic changes were estimated to be much lower than the rates of cladogenetic changes. Bayes factors did not support either anagenetic or cladogenetic polyploidization (Table 2.6). The MAP root chromosome number of 37, despite being very weakly supported (0.08), corresponds with the findings of Escudero et al. (2014), where it was also poorly supported (Figure 2.9).

In *Primula*, we found a base chromosome number for section *Aleuritia* of 9 with high posterior probability (0.82; Figure 2.10), which agrees with estimates from Glick and Mayrose (2014). We estimated moderate rates of anagenetic and cladogenetic changes, including both cladogenetic polyploidization and demi-polyploidization (Table 2.5). The MAP ancestral state estimates include an inferred history of possible polyploid and demi-polyploid speciation events in the clade containing the tetraploid *Primula halleri* and the hexaploid *P. scotica*. *Primula* is the only dataset out of the five analysed here for which Bayes factors supported the inclusion of cladogenetic demi-polyploidization (Table 2.6).

The well supported root chromosome number of 8 (posterior probability 0.90) found for *Mimulus* s.l. corresponds with the inferences reported in Beardsley et al. (2004). We estimated moderate rates of anagenetic dysploid gains and losses, as well as a moderate rate of cladogenetic polyploidization (Table 2.5). Bayes factors also supported models that included anagenetic dysploid gain and loss, as well as cladogenetic polyploidization (Table 2.6). The MAP ancestral state reconstruction revealed that most of the possible polyploid speciation events took place in the *Diplacus* clade, particularly in the clade containing the tetraploids *Mimulus cupreus*, *M. glabratus*, *M. luteus*, and *M. yecorensis* (Figure 2.11). Additionally, an ancient cladogenetic polyploidization event is inferred for the split between the two main *Diplacus* clades at about 5 million time units ago.

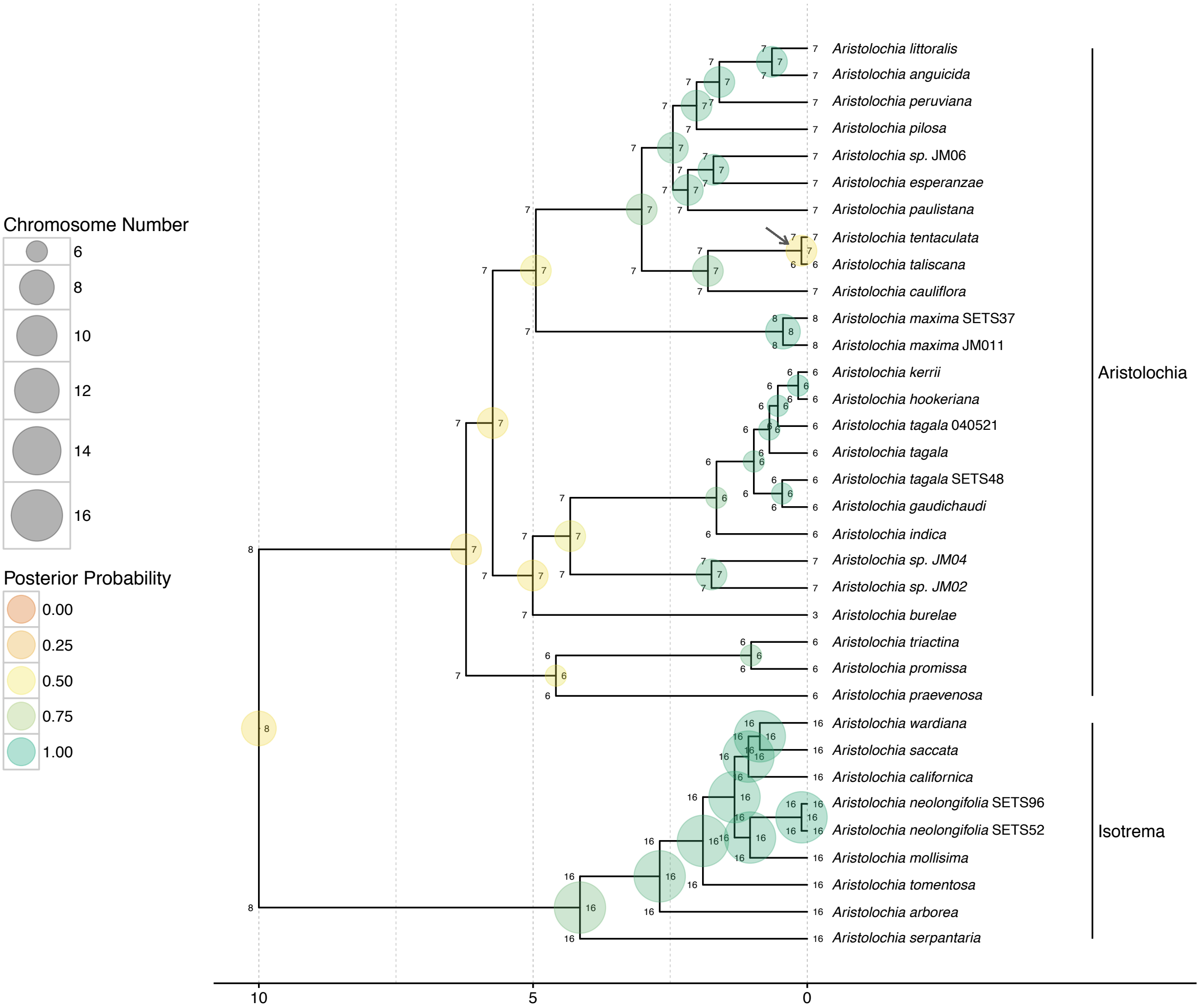
Chromosome Number
6
8
10
12
14
16
Posterior Probability
0.00
0.25
0.50
0.75
1.00
7 Aristolochia littoralis
7 Aristolochia anguicida
7 Aristolochia peruviana
7 Aristolochia pilosa
7 Aristolochia sp. JM06
7 Aristolochia esperanzae
7 Aristolochia paulistana
7 Aristolochia tentaculata
6 Aristolochia taliscana
7 Aristolochia cauliflora
8 Aristolochia maxima SETS37
8 Aristolochia maxima JM011
6 Aristolochia kerrii
6 Aristolochia hookeriana
6 Aristolochia tagala 040521
6 Aristolochia tagala
6 Aristolochia tagala SETS48
6 Aristolochia gaudichaudi
6 Aristolochia indica
7 Aristolochia sp. JM04
7 Aristolochia sp. JM02
3 Aristolochia burelae
6 Aristolochia triactina
6 Aristolochia promissa
6 Aristolochia praevenosa
16 Aristolochia wardiana
16 Aristolochia saccata
16 Aristolochia californica
16 Aristolochia neolongifolia SETS96
16 Aristolochia neolongifolia SETS52
16 Aristolochia mollisima
16 Aristolochia tomentosa
16 Aristolochia arborea
16 Aristolochia serpantaria
Aristolochia
Isotrema
10
5
0

Figure 2.7: Ancestral chromosome number estimates of *Aristolochia*. The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the "shoulders" of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 8 with a posterior probability of 0.45. The grey arrow highlights the possible dysploid speciation event leading to the west-central Mexican species *Aristolochia tentaculata* and *A. taliscana*. Clades corresponding to subgenera are indicated at right.

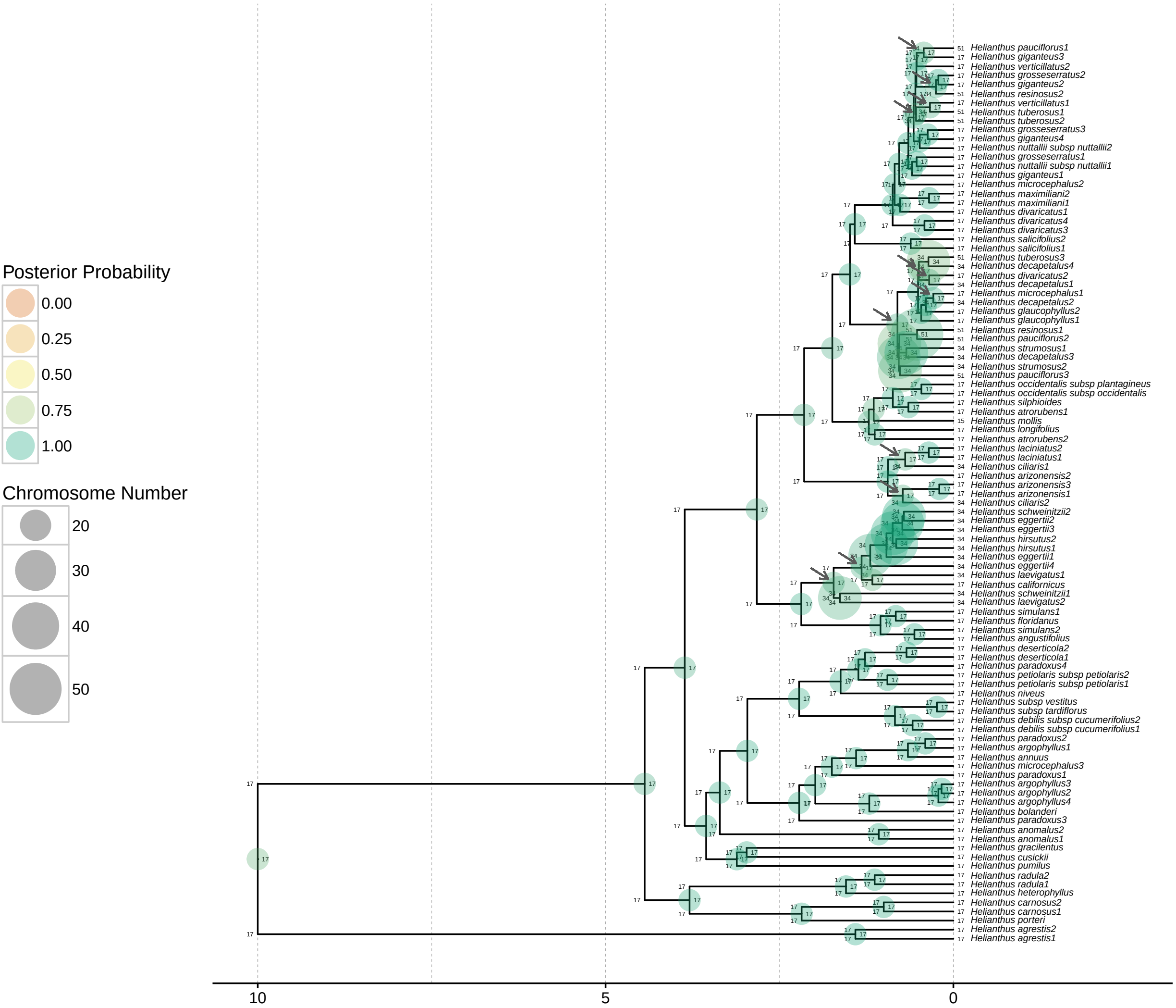
Posterior Probability
0.00
0.25
0.50
0.75
1.00
Chromosome Number
20
30
40
50
51 Helianthus pauciflorus1
17 Helianthus giganteus3
17 Helianthus verticillatus2
17 Helianthus grosseserratus2
17 Helianthus giganteus2
51 Helianthus resinosus2
17 Helianthus verticillatus1
51 Helianthus tuberosus1
51 Helianthus tuberosus2
17 Helianthus grosseserratus3
17 Helianthus giganteus4
17 Helianthus nuttallii subsp nuttallii2
17 Helianthus grosseserratus1
17 Helianthus nuttallii subsp nuttallii1
17 Helianthus giganteus1
17 Helianthus microcephalus2
17 Helianthus maximiliani2
17 Helianthus maximiliani1
17 Helianthus divaricatus1
17 Helianthus divaricatus4
17 Helianthus divaricatus3
17 Helianthus salicifolius2
17 Helianthus salicifolius1
51 Helianthus tuberosus3
34 Helianthus decapetalus4
17 Helianthus divaricatus2
34 Helianthus decapetalus1
17 Helianthus microcephalus1
34 Helianthus decapetalus2
17 Helianthus glaucophyllus2
17 Helianthus glaucophyllus1
51 Helianthus resinosus1
51 Helianthus pauciflorus2
34 Helianthus strumosus1
34 Helianthus decapetalus3
34 Helianthus strumosus2
51 Helianthus pauciflorus3
17 Helianthus occidentalis subsp plantagineus
17 Helianthus occidentalis subsp occidentalis
17 Helianthus silphioides
17 Helianthus atrorubens1
15 Helianthus mollis
17 Helianthus longifolius
17 Helianthus atrorubens2
17 Helianthus laciniatus2
17 Helianthus laciniatus1
34 Helianthus ciliaris1
17 Helianthus arizonensis2
17 Helianthus arizonensis3
17 Helianthus arizonensis1
34 Helianthus ciliaris2
34 Helianthus schweinitzii2
34 Helianthus eggertii2
34 Helianthus eggertii3
34 Helianthus hirsutus2
34 Helianthus hirsutus1
34 Helianthus eggertii1
34 Helianthus eggertii4
34 Helianthus laevigatus1
17 Helianthus californicus
34 Helianthus schweinitzii1
34 Helianthus laevigatus2
17 Helianthus simulans1
17 Helianthus floridanus
17 Helianthus simulans2
17 Helianthus angustifolius
17 Helianthus deserticola2
17 Helianthus deserticola1
17 Helianthus paradoxus4
17 Helianthus petiolaris subsp petiolaris2
17 Helianthus petiolaris subsp petiolaris1
17 Helianthus niveus
17 Helianthus subsp vestitus
17 Helianthus subsp tardiflorus
17 Helianthus debilis subsp cucumerifolius2
17 Helianthus debilis subsp cucumerifolius1
17 Helianthus paradoxus2
17 Helianthus argophyllus1
17 Helianthus annuus
17 Helianthus microcephalus3
17 Helianthus paradoxus1
17 Helianthus argophyllus3
17 Helianthus argophyllus2
17 Helianthus argophyllus4
17 Helianthus bolanderi
17 Helianthus paradoxus3
17 Helianthus anomalus2
17 Helianthus anomalus1
17 Helianthus gracilentus
17 Helianthus cusickii
17 Helianthus pumilus
17 Helianthus radula2
17 Helianthus radula1
17 Helianthus heterophyllus
17 Helianthus carnosus2
17 Helianthus carnosus1
17 Helianthus porteri
17 Helianthus agrestis2
17 Helianthus agrestis1
10
5
0

Figure 2.8: Ancestral chromosome number estimates of *Helianthus*. The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the "shoulders" of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 17 with a posterior probability of 0.91. The grey arrows show the locations of 12 inferred polyploid speciation events.

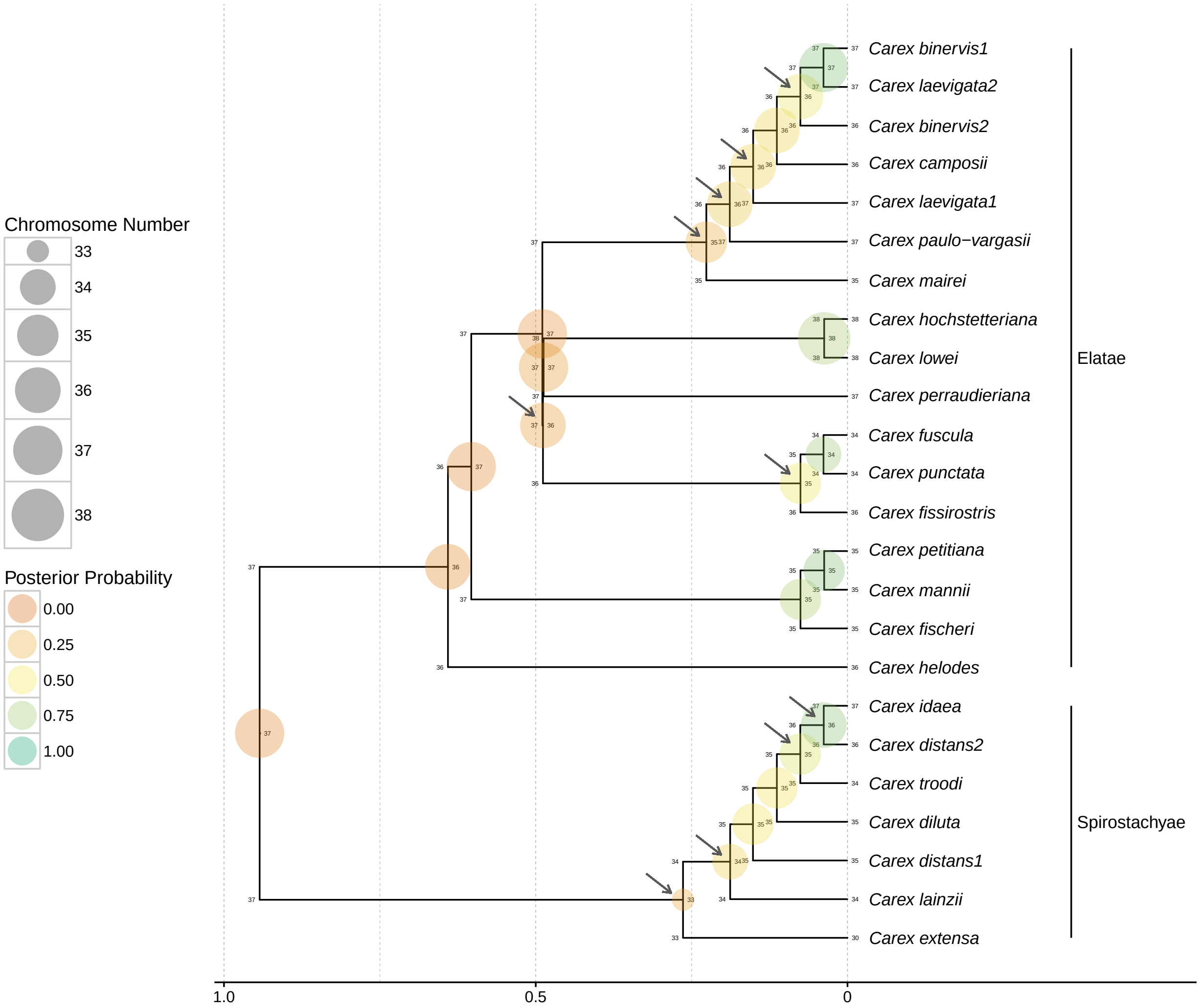

Figure 2.9: Ancestral chromosome number estimates of *Carex* section *Spirostachyae*. The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the "shoulders" of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 37 with a posterior probability of 0.08. Grey arrows indicate the location of possible dysploid speciation events. 36.9% of all speciation events include a cladogenetic gain or loss of a single chromosome. Clades corresponding to subsections are indicated at right.

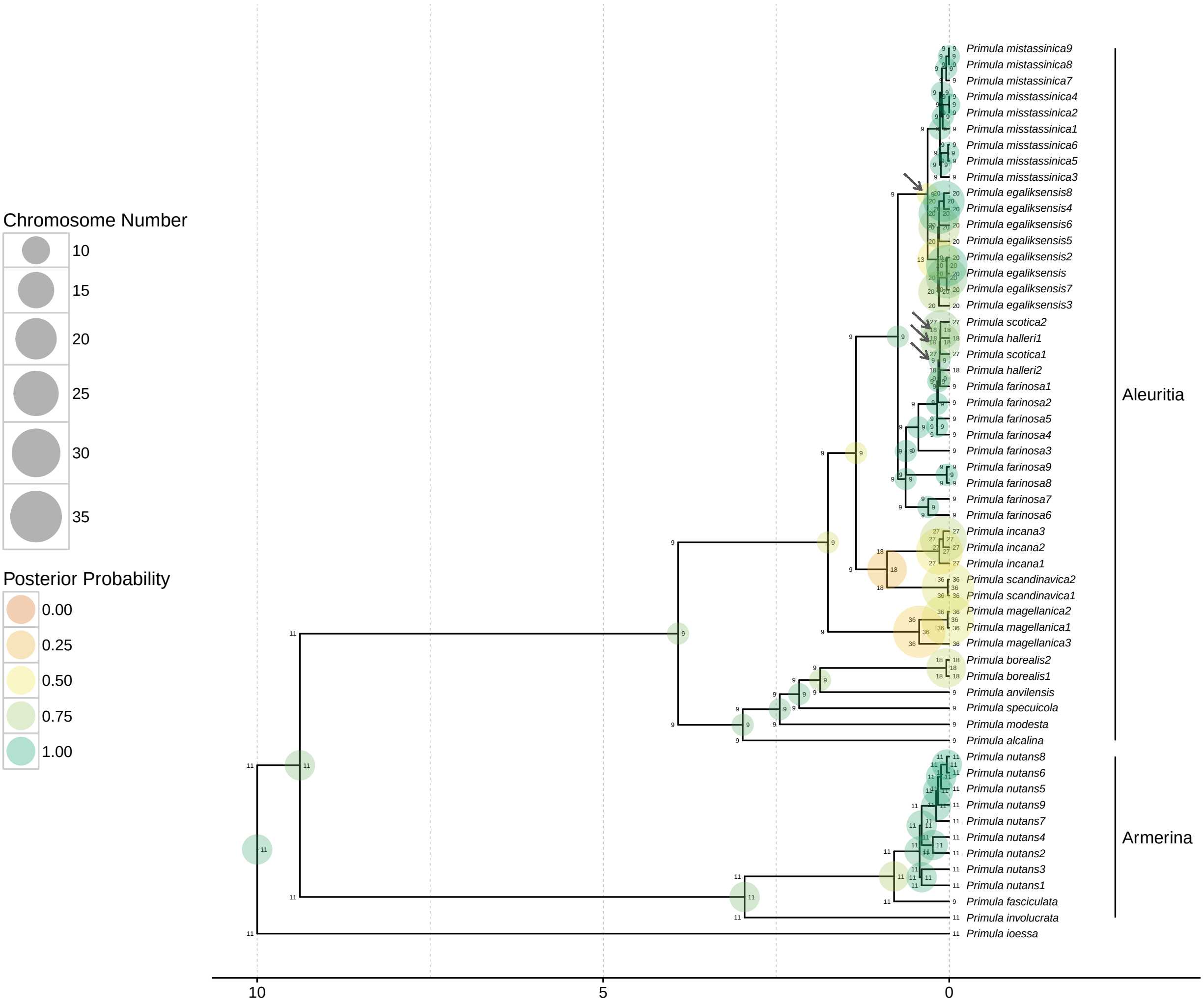

Figure 2.10: Ancestral chromosome number estimates of *Primula* section *Aleuritia*. The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the "shoulders" of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number of section *Aleuritia* is 9 with a posterior probability of 0.82. The arrows show the inferred history of possible polyploid and demi-polyploid speciation events in the clade containing the tetraploids *Primula egaliksensis* and *P. halleri* and the hexaploid *P. scotica*. Clades corresponding to sections are indicated at right.

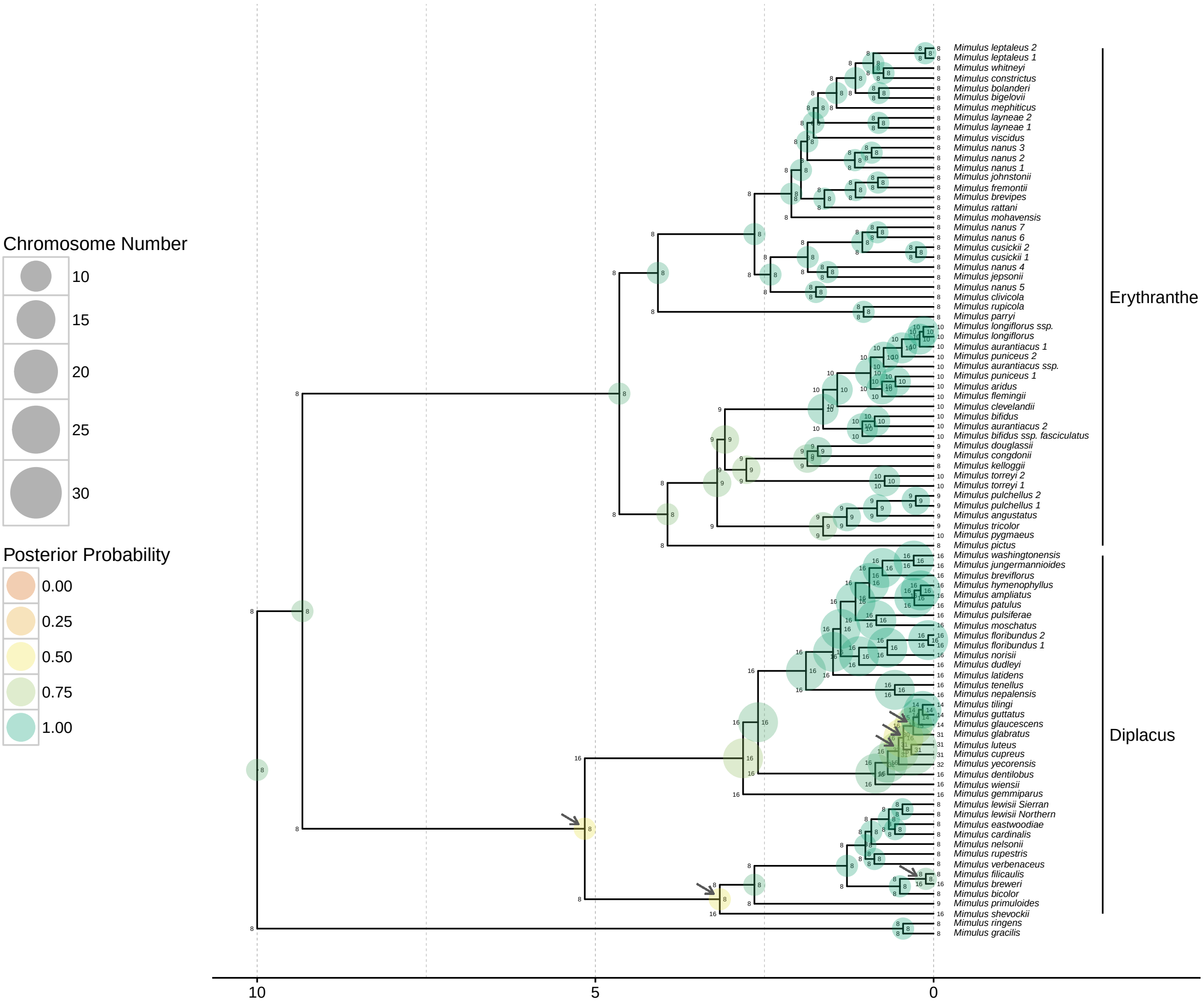

Figure 2.11: Ancestral chromosome number estimates of *Mimulus* sensu lato. The model averaged MAP estimate of ancestral chromosome numbers are shown at each branch node. The states of each daughter lineage immediately after cladogenesis are shown at the "shoulders" of each node. The size of each circle is proportional to the chromosome number and the color represents the posterior probability. The MAP root chromosome number is 8 with a posterior probability of 0.90. The arrows highlight the inferred history of repeated polyploid speciation events in the Diplacus clade, which contains the tetraploids *Mimulus cupreus*, *M. glabratus*, *M. luteus*, and *M. yecorensis*. Clades corresponding to segregate genera are indicated at right.

Table 2.5: Mean model-averaged parameter value estimates for empirical datasets. Rates for all parameters are given in units of chromosome changes per branch length unit except for μ which is given in extinction events per time units.

Clade	γ_a	δ_a	ρ_a	η_a	γ_m	δ_m	ϕ_c	γ_c	δ_c	ρ_c	η_c	μ
Aristolochia	0.02	0.05	0.01	0.0	-0.01	-0.01	0.43	0.0	0.04	0.0	0.0	0.19
Carex section *Spirostachyae*	0.19	0.79	0.16	0.13	0.0	0.04	2.49	2.15	0.15	0.95	0.5	2.26
Helianthus	0.0	0.02	0.0	0.03	-0.0	-0.0	0.68	0.0	0.0	0.13	0.0	0.09
Mimulus s.l.	0.03	0.02	0.01	0.0	0.02	0.02	0.65	0.0	0.0	0.05	0.0	0.16
Primula section *Aleuritia*	0.01	0.05	0.01	0.01	-0.0	-0.0	2.39	0.01	0.03	0.15	0.09	2.47

Table 2.6: Best supported chromosome evolution models for empirical datasets. The MAP model of chromosome evolution and its corresponding posterior probability are shown with Bayes factors (BF) for models that include each parameter. Parameters with $BF > 1$ are in bold and indicate support for models that include that parameter. Parameters with "positive" and "strong" support according to Kass and Raftery (1995) are marked with * and **, respectively.

Clade	MAP Model	Posterior Probability of MAP Model (%)	$BF\gamma_a$	$BF\delta_a$	$BF\rho_a$	$BF\eta_a$	$BF\gamma_m$	$BF\delta_m$	$BF\gamma_c$	$BF\delta_c$	$BF\rho_c$	$BF\eta_c$
Aristolochia	$\delta_a, \gamma_a, \rho_a$	0.05	**3.08***	**8.34***	**2.52**	0.42	0.55	0.61	0.15	**1.09**	0.06	0.03
Carex section *Spirostachyae*	$\delta_a, \delta_m, \gamma_c$	0.04	**1.11**	**42.67****	0.95	0.89	0.37	**6.33***	**37.02****	0.25	0.65	0.44
Helianthus	δ_a, η_a, ρ_c	0.22	0.35	**143.07****	0.51	**>1000****	0.15	0.87	0.02	0.04	**>1000****	0.16
Mimulus s.l.	$\gamma_a, \delta_a, \gamma_m, \delta_m, \rho_c$	0.13	**101.04****	**24.0****	0.86	0.31	**1.57**	**1.55**	0.07	0.1	**20.41****	0.02
Primula section *Aleuritia*	δ_a, ρ_c, η_c	0.06	0.63	**5.61***	0.95	0.58	0.23	0.64	0.17	0.54	**76.83****	**14.89***

2.4 Discussion

The results from the empirical analyses show that the ChromoSSE models detect strikingly different modes of chromosome evolution with clade-specific combinations of anagenetic and cladogenetic processes. Anagenetic dysploid gains and losses were supported in nearly all clades; however, cladogenetic dysploid changes were supported only in *Carex*. The occurrence of anagenetic dysploid changes in all clades suggest that small chromosome number changes due to gains and losses may frequently have a minimal effect on the formation of reproductive isolation, though our results suggest that *Carex* may be a notable exception. Anagenetic polyploidization was only supported in *Aristolochia*, while cladogenetic polyploidization was supported in *Helianthus*, *Mimulus* s.l., and *Primula*. These findings confirm the evidence presented by Zhan et al. (2016) that polyploidization events could play a significant role during plant speciation.

Our models shed new light on the importance of whole genome duplications as a key driver in evolutionary diversification processes. *Helianthus* has long been understood to have a complex history of polyploid speciation (Timme et al. 2007), but our results here are the first to statistically show the prevalance of cladogenetic polyploidization in *Helianthus* (occuring at 16% of all speciation events) and how few of the chromosome changes are estimated to be anagenetic. Polyploid speciation has also been suspected to be common in *Mimulus* s.l. (Vickery 1995), and indeed we estimated that 7% of speciation events were cladogenetic polyploidization events. We also estimated that the rates of cladogenetic dysploidization in *Mimulus* s.l. were 0, which is in contrast to the parsimony based inferences presented in Beardsley et al. (2004), which estimated 11.5% of all speciation events included polyploidization and 13.3% included dysploidization. Their estimates, however, did not distinguish cladogenetic from anagenetic processes, and so they likely underestimated anagenetic changes. Our ancestral state reconstructions of chromosome number evolution for *Helianthus*, *Mimulus* s.l., and *Primula* show that polyploidization events generally occurred in the relatively recent past; few ancient polyploidization events were reconstructed (one exception being the ancient cladogenetic polyploidization event in *Mimulus* clade *Diplacus*). This pattern appears to be consistent with recent studies that show polyploid lineages may undergo decreased net diversification (Mayrose et al. 2011; Scarpino et al. 2014), leading some to suggest that polyploidization may be an evolutionary dead-end (Arrigo and Barker 2012). While in the analyses presented here we fixed rates of speciation and extinction through time and across lineages, an obvious extension of our models would be to allow these rates to vary across the tree and statistically test for rate changes in polyploid lineages.

Our findings also suggest dysploid changes may play a significant role in the speciation process of some lineages. The genus *Carex* is distinguished by holocentric chromosomes that undergo common fusion and fission events but rarely polyploidization (Hipp 2007). This concurs with our findings from *Carex* section *Spirostachyae*, where we saw no support for models including either anagenetic or cladogenetic polyploidization. Instead we found high rates of cladogenetic dysploid change, which is congruent with earlier results that show that *Carex* diversification is driven by processes of fission and fusion occurring with cladogenetic shifts in chromosome number (Hipp 2007; Hipp et al. 2007). Hipp (2007) proposed a speciation scenario for *Carex* in which the gradual accumulation of chromosome fusions, fissions, and rearrangements in recently diverged populations increasingly reduce the fertility of hybrids between populations, resulting in high species richness. More recently, Escudero et al. (2016) found that chromosome number differences in *Carex scoparia* led to reduced germination rates, suggesting hybrid dysfunction could spur chromosome speciation in *Carex*. Holocentricity has arisen at least 13 times independently in plants and animals (Melters et al. 2012), thus future work could examine chromosome number evolution in other holocentric clades and test for similar patterns of cladogenetic fission and fusion events.

The models presented here could also be used to further study the role of divergence in genomic architecture during sympatric speciation. Chromosome structural differences have been proposed to perform a central role in sympatric speciation, both in plants (Gottlieb 1973) and animals (Feder et al. 2005; Michel et al. 2010). In *Aristolochia* we found most changes in chromosome number were estimated to be anagenetic, with the only cladogenetic change occuring among a pair of recently diverged sympatric species. By coupling our chromosome evolution models with models of geographic range evolution it would be possible to statistically test whether the frequency of cladogenetic chromosome changes increase in sympatric speciation events compared to allopatric speciation events, thereby testing for interaction between these two different processes of reproductive isolation and evolutionary divergence.

The simulation results from Experiment 1 demonstrate that extinction reduces the accuracy of inferences made by models of chromosome evolution that do not take into account unobserved speciation events. Furthermore, the simulations performed in Experiments 2 and 3 show that the substantial uncertainty introduced in our analyses by jointly estimating diversification rates and chromosome evolution resulted in lower posterior probabilities for ancestral state reconstructions. We feel that this is a strength of our method; the lower posterior probabilities incorporate true uncertainty due to extinction and so represent more conservative estimates. Additionally, the simulation results from Experiment 4 reveal that rates of anagenetic evolution were overestimated and rates of cladogenetic change were underestimated when the generating process consisted only of cladogenetic events. This suggests the possibility that our models of chromosome number evolution are only partially identifiable, and that the results of our empirical analyses may have a similar bias towards overestimating anagenetic evolution and underestimating cladogenetic evolution. This bias may be an issue for all ClaSSE type models, but the practical consequences here are conservative estimates of cladogenetic chromosome evolution.

An important caveat for all phylogenetic methods is that estimates of model parameters and ancestral states can be highly sensitive to taxon sampling (Heath et al. 2008). All of the empirical datasets examined here included non-monophyletic taxa that were treated as separate lineages. We made the unrealistic assumptions that 1) each of the non-monophyletic lineages sharing a taxon name have the same cytotype, and 2) the taxon sampling probability (ρ_s) for the birth-death process was 1.0. The former assumption could drastically affect ancestral state estimates, but its effect can only be confirmed by obtaining chromosome counts for each lineage regardless of taxon name. While the results from simulation Experiment 5 showed that fixing ρ_s to 1.0 did not decrease the accuracy of inferred ancestral states, we still performed extra analyses of the empirical datasets with different values of ρ_s (results not shown). The results indicated that total speciation and extinction rates are sensitive to ρ_s, but the relative speciation rates (e.g. between ϕ_c and γ_c) remained similar. The ancestral state estimates of cladogenetic and anagenetic chromosome changes were robust to different values of ρ_s. This could vary among datasets and care should be taken when considering which lineages to sample.

Bayesian model averaging is particularly appropriate for models of chromosome number evolution since conditioning on a single model ignores the considerable degree of model uncertainty found in both the simulations and the empirical analyses. In the simulations the true model of chromosome evolution was rarely inferred to be the MAP model ($< 39\%$ of replicates), and in the instances it was correctly identified the posterior probability of the MAP model was < 0.13. The posterior probabilities of the MAP models for the empirical datasets were similarly low, varying between 0.04 and 0.22. Conditioning on a single poorly fitting model of chromosome evolution, even when it is the best model available, results in an underestimate of the uncertainty of ancestral chromosome numbers. Furthermore, Bayesian model averaging enabled us to detect different modes of chromosome number evolution without the limitation of traditional model testing procedures in which multiple analyses are performed that each condition on a different single model. This is a particularly useful approach when the space of all possible models is large.

Our `RevBayes` implementation facilitates model modularity and easy experimentation. Experimenting with different priors or MCMC moves is achieved by simply editing the `Rev` scripts that describe the model. Though in our analyses here we ignored phylogenetic uncertainty by assuming a fixed known tree, we could easily incorporate this uncertainty by modifying a couple lines of the `Rev` script to integrate over a previously estimated posterior distribution of trees. We could also use molecular sequence data simultaneously with the chromosome models to jointly infer phylogeny and chromosome evolution, allowing the chromosome data to help inform tree topology and divergence times. In this paper we chose not to perform joint inference so that we could isolate the behavior of the chromosome evolution models; however, this is a promising direction for future research.

There are a number of challenging directions for future work on phylogenetic chromosome evolution models. Models that incorporate multiple aspects of chromosome morphology such as translocations, inversions, and other gene synteny data as well as the presence of ring and/or B chromosomes have yet to be developed. None of our models currently account for allopolyploidization; indeed few phylogenetic comparative methods can handle reticulate evolutionary scenarios that result from allopolyploidization and other forms of hybridization (Marcussen et al. 2015). A more tractable problem is mapping chromosome number changes along the branches of the phylogeny, as opposed to simply making estimates at the nodes as we have done here. Since the approach described here models both anagenetic and cladogenetic chromosome evolution processes while accounting for unobserved speciation events, the rejection sampling procedure used in standard stochastic character mapping (Nielsen 2002; Huelsenbeck et al. 2003) is not sufficient. While data augmentation approaches such as those described by Bokma (2008) could be utilized, they require complex MCMC algorithms that may have difficulty mixing. Another option is to extend the method described in this paper to draw joint ancestral states by numerically integrating root-to-tip over the tree into a new procedure called joint conditional character mapping. This sort of approach would infer the joint MAP history of chromosome changes both at the nodes and along the branches of the tree, and provide an alternative to stochastic character mapping that will work for all ClaSSE type models.

2.4.1 Conclusions

The analyses presented here show that the ChromoSSE models of chromosome number evolution successfully infer different clade-specific modes of chromosome evolution as well as the history of anagenetic and cladogenetic chromosome number changes for a clade, including reconstructing the timing and location of possible chromosome speciation events over the phylogeny. These models will help investigators study the mode and history of chromosome evolution within individual clades of interest as well as advance understanding of how fundamental changes in the architecture of the genome such as whole genome duplications affect macroevolutionary patterns and processes across the tree of life.

2.5 Supporting Information

2.5.1 Validating Ancestral State Estimates

Ancestral State Estimates of SSE Models

The code repository `http://github.com/wf8/anc_state_validation` contains scripts to validate the Monte Carlo method of ancestral state estimation for state-dependent speciation and extinction (SSE) models we implemented in `RevBayes` (Höhna et al. 2016) against the analytical marginal ancestral state estimation implemented in the `R` package `diversitree` (FitzJohn 2012).

Although the closest model to ChromoSSE implemented in `diversitree` is ClaSSE (Goldberg and Igić 2012), ancestral state estimation for ClaSSE is not implemented in `diversitree`. Therefore here we compare the ancestral state estimates for BiSSE (Maddison et al. 2007) as implemented in `diversitree` to the estimates made by `RevBayes`. Note that as implemented in `RevBayes` the BiSSE, ChromoSSE, ClaSSE, MuSSE (FitzJohn 2012), and HiSSE (Beaulieu and OMeara 2016) models use the same `C++` classes and algorithms for parameter and ancestral state estimation, so validating ancestral state estimates for BiSSE should provide confidence in estimates made by `RevBayes` for all these SSE models.

In `RevBayes` we sample ancestral states for SSE models from their joint distribution conditional on the tip states and the model parameters during the MCMC. However, in this work we summarize the MCMC samples by calculating the marginal posterior probability of each node being in each state. So the `RevBayes` marginal ancestral state reconstructions which are estimated via MCMC are directly comparable to the analytical marginal ancestral states computed by `diversitree`. It would be possible to summarize the samples from the MCMC to reconstruct the maximum a posteriori joint ancestral state reconstruction, but we have not done so in this work.

Comparison of `RevBayes` Estimates to `diversitree`

Here we show ancestral state estimates under BiSSE for an example where the tree and tip data were simulated in `diversitree` with the following parameters: $\lambda_0 = 0.2, \lambda_1 = 0.4, \mu_0 = 0.01, \mu_1 = 0.1$, and $q_{01} = q_{10} = 0.1$. The ancestral state reconstructions from `RevBayes` and `diversitree` are shown in Figures 2.13 and 2.14, respectively.

The log-likelihood as computed by `diversitree` was -109.46, whereas with `RevBayes` it was -109.71. Small differences in the log-likelihoods are expected due to differences in the way `diversitree` and `RevBayes` calculate probabilities at the root, and also due to numerical approximations. However both reconstructions should return the same probabilities for ancestral states at the root, and indeed `diversitree` calculated the root probability of being in state 0 as 0.555 and `RevBayes` calculated it as 0.554. The estimated posterior probabilities are very close for all nodes. This is shown in a plot comparing the marginal posterior probabilities for all nodes being in state 1 as estimated by `RevBayes` against the `diversitree` estimates (Figure 2.12).

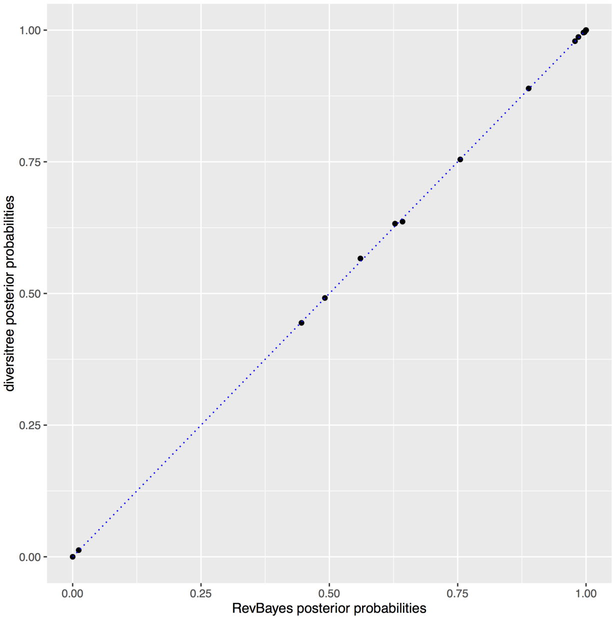

Figure 2.12: Posterior probabilities of marginal ancestral state estimates. Each point represents the marginal posterior probability of a node being in state 1 as estimated by `RevBayes` plotted against the estimates made by `diversitree`. The marginal ancestral states were estimated under BiSSE from a tree and tip data simulated with the following parameters: $\lambda_0 = 0.2, \lambda_1 = 0.4, \mu_0 = 0.01, \mu_1 = 0.1$, and $q_{01} = q_{10} = 0.1$. The full ancestral state reconstructions from `RevBayes` and `diversitree` are shown in Figures 2.13 and 2.14, respectively.

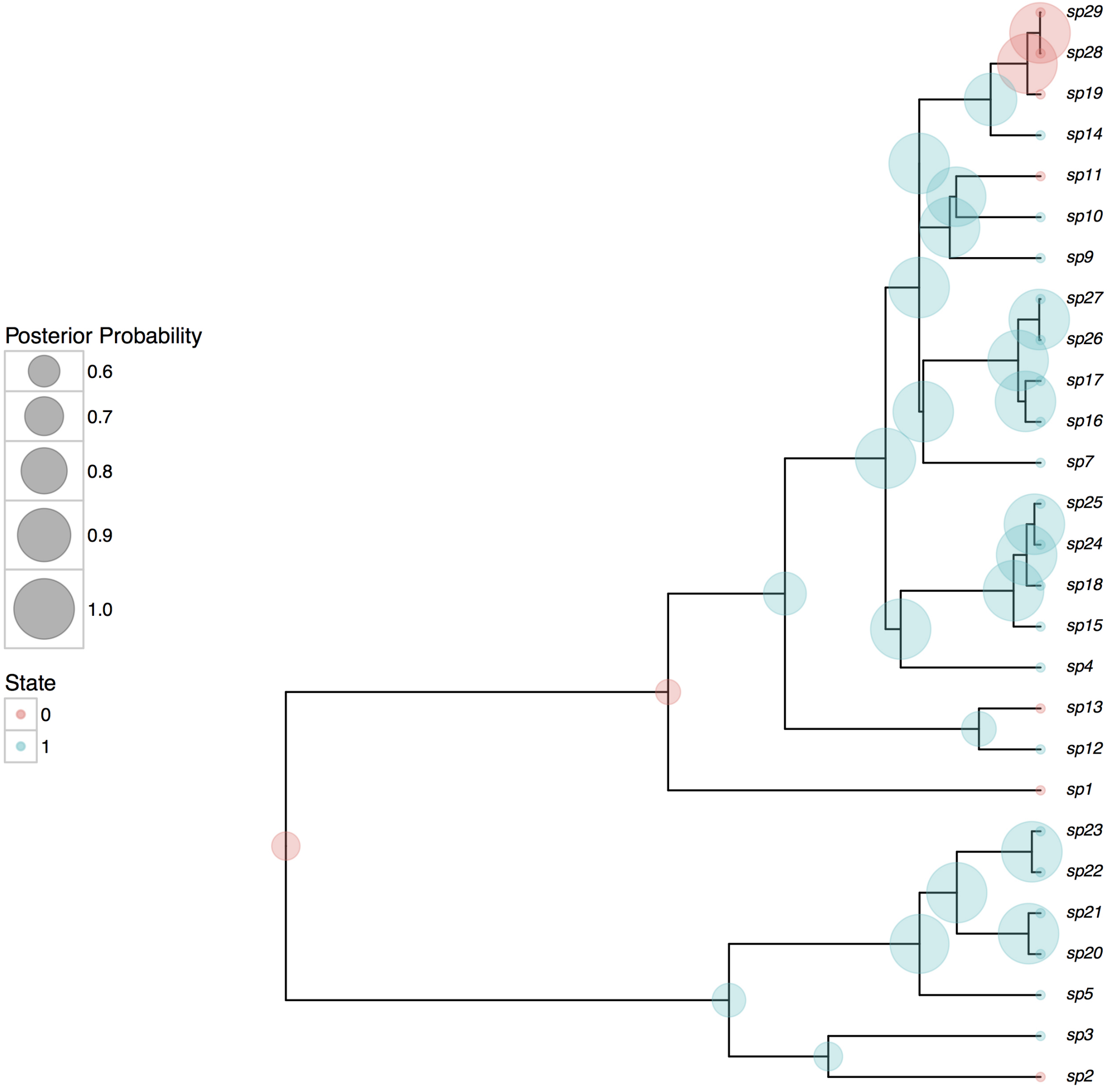

Figure 2.13: Ancestral state estimates from `RevBayes`. Marginal ancestral states estimated under BiSSE from a tree and tip data simulated with the following parameters: $\lambda_0 = 0.2, \lambda_1 = 0.4, \mu_0 = 0.01, \mu_1 = 0.1$, and $q_{01} = q_{10} = 0.1$.

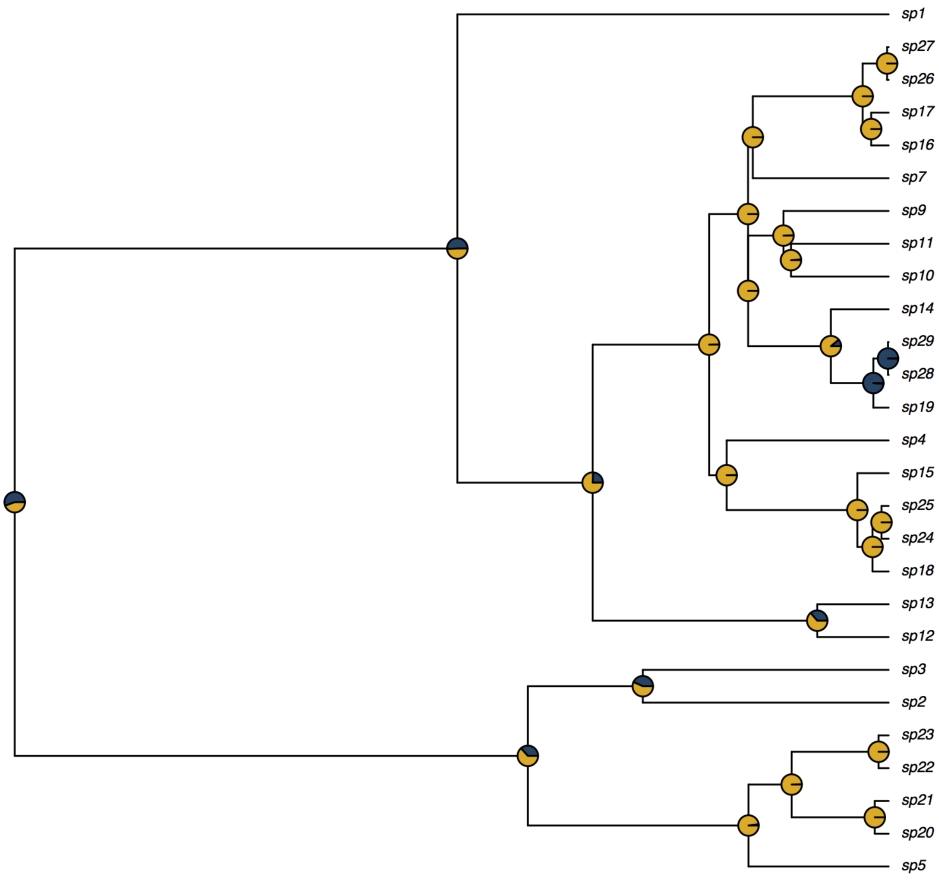

Figure 2.14: Ancestral state estimates from `diversitree`. Marginal ancestral states estimated under BiSSE from a tree and tip data simulated with the following parameters: $\lambda_0 = 0.2, \lambda_1 = 0.4, \mu_0 = 0.01, \mu_1 = 0.1$, and $q_{01} = q_{10} = 0.1$. Dark blue represents state 0 and yellow represents state 1.

2.5.2 Metropolis-Hastings Moves

The Metropolis-Hastings moves used in all ChromoSSE analyses are outlined in Table 2.7. All MCMC proposals are standard except the ElementSwapSimplex move and the reversible jump MCMC proposals. These are described in detail in the main text. MCMC analyses were run in `RevBayes` for 11000 iterations, where each iteration consisted of 79 MCMC moves per iteration. The 79 moves were randomly drawn from the 28 different Metropolis-Hastings moves listed in Table 2.7 using the weights listed. Samples of parameter values and joint ancestral states were drawn each iteration, and the first 1000 samples were discarded as burn in.

Table 2.7: MCMC moves used for chromosome number evolution analyses. See the main text for further explanations of the moves used. Samples were drawn from the MCMC each iteration, where each iteration consisted of 28 different moves in a random move schedule with 79 moves per iteration.

	Parameter	X	Move	Weight
Anagenetic	Chromosome gain rate	γ_a	Scale($\lambda = 1$)	2
	Chromosome gain rate	γ_a	Reduce/Augment	2
	Chromosome loss rate	δ_a	Scale($\lambda = 1$)	2
	Chromosome loss rate	δ_a	Reduce/Augment	2
	Polyploidization rate	ρ_a	Scale($\lambda = 1$)	2
	Polyploidization rate	ρ_a	Reduce/Augment	2
	Demi-polyploidization rate	η_a	Scale($\lambda = 1$)	2
	Demi-polyploidization rate	η_a	Reduce/Augment	2
	Linear component of gain rate	γ_m	Slide($\delta = 0.1$)	1
	Linear component of gain rate	γ_m	Slide($\delta = 0.001$)	1
	Linear component of gain rate	γ_m	Reduce/Augment	2
	Linear component of loss rate	δ_m	Slide($\delta = 0.1$)	1
	Linear component of loss rate	δ_m	Slide($\delta = 0.001$)	1
	Linear component of loss rate	δ_m	Reduce/Augment	2
Cladogenetic	No change	ϕ_c	Scale($\lambda = 5$)	2
	Chromosome gain	γ_c	Scale($\lambda = 5$)	2
	Chromosome gain	γ_c	Reduce/Augment	2
	Chromosome loss	δ_c	Scale($\lambda = 5$)	2
	Chromosome loss	δ_c	Reduce/Augment	2
	Polyploidization	ρ_c	Scale($\lambda = 5$)	2
	Polyploidization	ρ_c	Reduce/Augment	2
	Demi-polyploidization	η_c	Scale($\lambda = 5$)	2
	Demi-polyploidization	η_c	Reduce/Augment	2
	All cladogenetic rates	$\phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$	Joint Up-Down Scale($\lambda = 0.5$)	2
Other	Root frequencies	π	BetaSimplex($\alpha = 0.5$)	10
	Root frequencies	π	ElementSwapSimplex	20
	Relative-extinction	r	Scale($\lambda = 5$)	3
	Relative-extinction and all clado rates	$r, \phi_c, \gamma_c, \delta_c, \rho_c, \eta_c$	Joint Up-Down Scale($\lambda = 0.5$)	2
Total			**28**	**79**

2.5.3 Simulation Details

Description of Simulation Experiments

Experiment 1

In experiment 1 we tested the effect of unobserved speciation events due to extinction on chromosome number estimates when using a model that does not account for unobserved speciation. Is the additional model complexity required to account for unobserved speciation necessary, or are the effects of unobserved speciation negligible and safe to ignore? Using the non-SSE model described above that does not account for unobserved speciation, ancestral chromosome numbers and chromosome evolution model parameters were estimated for each of the 600 datasets.

Experiment 2

Here we compared the accuracy of models of chromosome evolution that account for unobserved speciation versus those that do not. Since extinction can safely be assumed to be present to some extent in all clades, it is likely that all empirical datasets contain some unobserved speciation. Do we see an increase in accuracy when we account for unobserved speciation events, or conversely do we see an increase in the variance of our estimates that perhaps describes true uncertainty due to extinction? To test this, we estimated ancestral chromosome numbers and chromosome evolution model parameters over the simulated datasets that included unobserved speciation using both ChromoSSE that accounts for unobserved speciation as well as the non-SSE model that does not.

Experiment 3

In experiment 3 we tested the effect of jointly estimating speciation and extinction rates with chromosome number evolution. Estimating speciation and extinction rates accurately is notoriously challenging (Nee et al. 1994a; Rabosky 2010; Beaulieu and O'Meara 2015; May et al. 2016), so how much of the variance in chromosome evolution estimates made with models that jointly estimate speciation and extinction are due to uncertainty in diversification rates? Here we compared our estimates of ancestral chromosome numbers and chromosome evolution model parameters using ChromoSSE that accounts for unobserved speciation (and in which speciation and extinction rates are jointly estimated) with estimates made from ChromoSSE but where the true rates of speciation and extinction used to simulate the data were fixed. The latter analyses were given the true rates of total speciation and extinction, but still had to estimate the proportion of speciation events for each type of cladogenetic event.

Experiment 4

Since we model the same chromosome number transitions as both cladogenetic and anagenetic processes, it is possible that the two processes could be confounded and our models may not be fully identifiable. Furthermore, preliminary results suggested our models overestimate anagenetic changes and underestimate cladogenetic changes when the true generating process

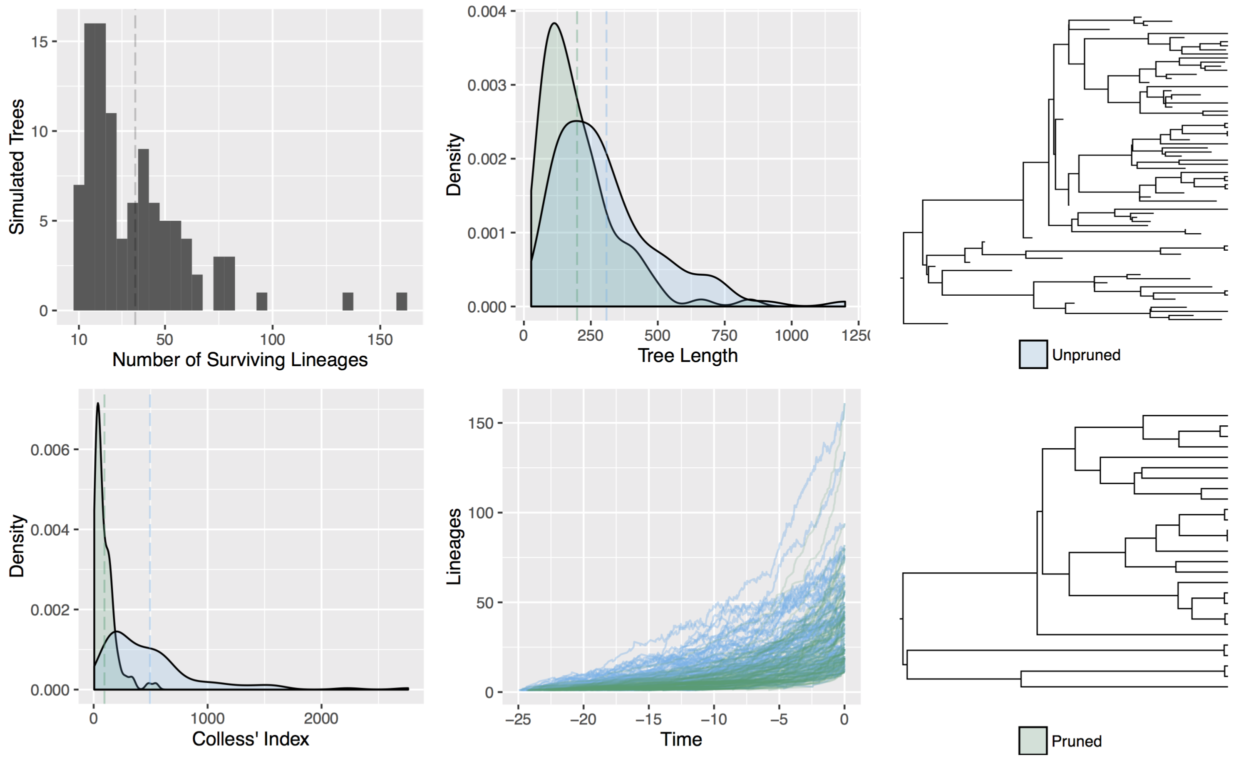

Figure 2.15: Tree simulations. 100 trees were simulated under the birth-death process as described in the main text for Experiments 1, 2, 3, and 4. Chromosome number evolution was simulated over the unpruned trees that included all extinct lineages, as well as over the same trees but with extinct lineages pruned. This resulted in two simulated datasets: one simulated under a process that did have unobserved speciation events, and one simulated with no unobserved speciation events. Shown above is a histogram of the number of lineages that survived to the present, the tree lengths, Colless' Index (a measure of tree imbalance; Colless 1982), and lineage through time plots of the 100 pruned and unpruned trees.

includes cladogenetic evolution. Here we compared cladogenetic and anagenetic estimates made by ChromoSSE under simulation scenarios that only included cladogenetic changes. Do we see an increase in accuracy of cladogenetic parameter estimates when anagenetic changes are disallowed (fixed to 0)?

Experiment 5

Experiments 1-3 deal with the increase in uncertainty caused by unobserved speciation events due to extinction. Here we focused on the effect of unobserved speciation due to incomplete taxon sampling by comparing chromosome number estimates at 3 levels of taxon sampling: 100%, 50%, and 10%. We compared estimates made by both the ChromoSSE model and the non-SSE model, as well as compared estimates made by ChromoSSE using the true taxon sampling probability ρ_s versus estimates made by ChromoSSE using ρ_s fixed to 1.0.

Table 2.8: Simulation parameter values. Parameter values used to simulate datasets. The top 3 rows show the 3 modes of chromosome number evolution simulated for Experiments 1, 2, 3, and 4: anagenetic only, cladogenetic only, and mixed. Row 4 shows the parameter values used to simulate data for Experiment 5. The total speciation rate $\lambda_t = 0.25$ and the extinction rate $\mu = 0.15$. The root state was fixed to 8.

Simulation mode	γ_a	δ_a	ρ_a	η_a	γ_m	δ_m	ϕ_c	γ_c	δ_c	ρ_c	η_c
Anagenetic	0.0085	0.0085	0.0085	-	-	-	λ_t	-	-	-	-
Cladogenetic	-	-	-	-	-	-	$0.85\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	-
Mixed	0.0085	0.0085	0.0085	-	-	-	$0.85\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	$0.05\lambda_t$	-
Experiment 5	0.0025	0.0025	0.0025	-	-	-	$0.93\lambda_t$	$0.02\lambda_t$	$0.02\lambda_t$	$0.02\lambda_t$	-

Methods Used to Simulate Data

For experiments 1, 2, 3, and 4 the same set of simulated trees and chromosome counts were used. Since ChromoSSE assumes the total rates of speciation and extinction are fixed over the tree (see Equation 2.5), trees were first simulated with constant diversification rates, and then cladogenetic and anagenetic chromosome evolution was simulated over the trees. 100 trees were simulated under the birth-death process with $\lambda = 0.25$ and $\mu = 0.15$ (see Figure 2.15) using the R package `diversitree` (FitzJohn 2012). The trees were conditioned on an age of 25.0 time units and a minimum of 10 extant lineages. To test the effect of unobserved speciation events due to lineages going extinct on cladogenetic estimates, chromosome number evolution was simulated along the trees including their extinct lineages (unpruned) and the same 100 trees but with the extinct lineages pruned. All chromosome number simulations were performed using `RevBayes` (Höhna et al. 2016).

Three models were used to generate simulated chromosome counts: a model where all chromosome evolution was anagenetic, a model where all chromosome evolution was cladogenetic, and a model that mixed both anagenetic and cladogenetic changes (Table 2.8). Parameter values were roughly informed by the mean values estimated from the empirical datasets. The mean length of the simulated trees was 253.5 (Figure 2.15). Hence, the anagenetic rates were set to $2/235.5 \approx 0.0085$ which corresponds to an expected value of 2 events over the tree for each of the four transition types. The root chromosome number was fixed to be 8. Simulating data for all 3 models over both the pruned and unpruned tree resulted in 600 simulated datasets. To reproduce the effect of using reconstructed phylogenies all inferences were performed using the trees with extinct lineages pruned and with chromosome counts from extinct lineages removed.

Since Experiment 5 focused on the effect of incomplete taxon sampling on chromosome number estimates, the trees used needed to be conditioned on a known number of extant tips. The trees used for the previous simulations were conditioned only on age and a minimum of 10 extant lineages and so were not appropriate. To simulate 100 trees conditioned on 200 extant lineages we used the R package `TreeSim` (Stadler 2011) with $\lambda = 0.25$ and $\mu = 0.15$ (like above). Complete trees with both extant and extinct lineages were simulated, and then chromosome evolution was simulated over the complete tree. Since these trees had a significantly longer mean length (2020.1 compared to 253.5) we used different rates of chromosome evolution to simulate data compared to Experiments 1, 2, 3, and 4 (Table 2.8). Chromosome numbers were only simulated using a mixed anagenetic and cladogenetic model.

The anagenetic rates were set to $5/2020.1 \approx 0.0025$ which corresponds to an expected value of 5 events over the tree for each of the four transition types. Like Experiments 1, 2, 3, and 4, the root chromosome number was fixed to be 8. Once chromosome data was simulated over the complete trees, the extinct taxa were pruned off leaving trees with 100% taxon sampling. 50% of the tips were randomly pruned off to create trees with 50% taxon sampling, and 90% of the tips were randomly pruned off to create trees with 10% taxon sampling.

2.5.4 MCMC Convergence of Simulation Replicates

Effective sample sizes (ESS) for all parameters in all simulation replicates were over 200, and the mean ESS values of the posterior for the replicates was 1470.3. Since the space of possible models is so large (1024 possible models, see main text), we replicated all analyses that included unobserved speciation in Experiment 1 three independent times to ensure that MCMC convergence was not an issue in detecting the true model of chromosome number evolution used to simulate the data. The results displayed in Table 2.9 show that the percentage of simulation replicates in which the true model was inferred to be the MAP model, and the mean posterior of the true model, converged and were stable across all three independent runs.

Table 2.9: Simulation Experiment 1 replicated 3 times. Estimates of the true model that generated the simulated data and estimates of the posterior probability of the true model were stable and converged across multiple independent replicates of the experiment.

Replicate	Mode of Evolution Used to Simulate Data	True Model Estimated (%)	Mean Posterior of True Model
1	Cladogenetic	15	0.09
1	Anagenetic	36	0.12
1	Mixed	2	0.10
2	Cladogenetic	15	0.09
2	Anagenetic	36	0.12
2	Mixed	2	0.09
3	Cladogenetic	15	0.09
3	Anagenetic	36	0.12
3	Mixed	2	0.10

Chapter 3

Stochastic character mapping of state-dependent diversification

Abstract

A major goal of evolutionary biology is to identify key evolutionary transitions that correspond with shifts in speciation and extinction rates. Stochastic character mapping has become the primary method used to infer the timing, nature, and number of character state transitions along the branches of a phylogeny. The method is widely employed for standard substitution models of character evolution. However, current approaches cannot be used for models that specifically test the association of character state transitions with shifts in diversification rates such as state-dependent speciation and extinction (SSE) models. Here we introduce a new stochastic character mapping algorithm that overcomes these limitations, and apply it to study mating system evolution over a densely sampled fossil-calibrated phylogeny of the plant family Onagraceae. Utilizing a hidden state SSE model we tested the association of the loss of self-incompatibility with shifts in diversification rates. Confirming long standing theory, we found that self-compatible lineages have higher extinction rates and lower net diversification rates compared to self-incompatible lineages. Further, our mapped character histories show that the loss of self-incompatibility is followed by a short-term spike in speciation rates, which declines after a time lag of several million years resulting in negative net diversification. Lineages that have long been self-compatible such as *Fuchsia* and *Clarkia* are in a previously unrecognized and ongoing evolutionary decline. Our results demonstrate that stochastic character mapping of SSE models is a powerful tool for examining the timing and nature of both character state transitions and shifts in diversification rates over the phylogeny.

3.1 Introduction

Evolutionary biologists have long sought to identify key evolutionary transitions that drive the diversification of life (Szathmary and Smith 1995; Sanderson and Donoghue 1996). One frequently used method to test hypotheses about evolutionary transitions is stochastic character mapping on a phylogeny (Nielsen 2002; Huelsenbeck et al. 2003). While most ancestral state reconstruction methods estimate states only at the nodes of a phylogeny, stochastic character mapping explicitly infers the timing and nature of each evolutionary transition along the branches of a phylogeny. However, current approaches to stochastic character mapping have two major limitations: the commonly used rejection sampling approach proposed by Nielsen (2002) is inefficient for characters with large state spaces (Huelsenbeck et al. 2003; Hobolth and Stone 2009), and more importantly current methods only apply to models of character evolution that are finite state substitution processes. While the first limitation has been partially overcome through uniformization techniques (Rodrigue et al. 2008; Irvahn and Minin 2014), a novel approach is needed for models with infinite state spaces, such as models to specifically test the association of character state transitions with shifts in diversification rates. These models describe the joint evolution of both a character and the phylogeny itself, and define a class of widely used models called state-dependent speciation and extinction models (SSE models; Maddison et al. 2007; FitzJohn 2012; Goldberg and Igić 2012; Freyman and Höhna 2017a).

In this work we introduce a method to sample character histories directly from their joint distribution, conditional on the observed tip data and the parameters of the model of character evolution. The method is applicable to standard finite state Markov processes of character evolution and also more complex SSE models that are infinite state Markov processes. The method does not rely on rejection sampling and does not require complex data augmentation (Van Dyk and Meng 2001) schemes to handle unobserved speciation/extinction events. Our implementation directly simulates the number, type, and timing of diversification rate shifts and character state transitions on each branch of the phylogeny. Thus, when applying our method together with a Markov chain Monte Carlo (MCMC; Metropolis et al. 1953) algorithm we can sample efficiently from the posterior distribution of both character state transitions and shifts in diversification rates over the phylogeny.

To illustrate the usefulness of our method to sample stochastic character maps from SSE models, we applied the method to study the association of diversification rate shifts with mating system transitions in the plant family Onagraceae. The majority of flowering plants are hermaphrodites, and the loss of self-incompatibility (SI), the genetic system that encourages outcrossing and prevents self-fertilization, is a common evolutionary transition (Stebbins 1974; Grant 1981; Barrett 2002). Independent transitions to self-compatibility (SC) have occurred repeatedly across the angiosperm phylogeny (Igic et al. 2008) and within Onagraceae (Raven 1979). Despite the repeated loss of SI, outcrossing is widespread and prevalent in plants, an observation that led Stebbins to hypothesize that SC was an evolutionary dead-end (Stebbins 1957). Stebbins proposed that over evolutionary time SC lineages will have higher extinction rates due to reduced genetic variation and an inability to adapt to changing conditions. However, Stebbins also speculated that SC is maintained by providing a short-term advantage in the form of reproductive assurance. The ability of SC lineages to self reproduce has long been understood to be potentially beneficial in droughts and other conditions where pollinators are rare (Darwin 1876) or after long distance dispersal when a single individual can establish a new population (Baker 1955).

Recent studies have reported higher net diversification rates for SI lineages, supporting Stebbins' dead-end hypothesis (Goldberg et al. 2010; Ferrer and Good 2012). However, explicit phylogenetic tests for increased extinction rates in SC lineages are limited to the plant family Solanaceae, where increased rates of speciation in SC lineages were offset by higher extinction rates, leading to lower overall rates of net diversification in SC lineages compared to SI lineages (Goldberg et al. 2010). In the study by Goldberg et al., the association of mating system transitions with shifts in extinction and speciation rates was tested using the Binary State Speciation and Extinction model (BiSSE; Maddison et al. 2007). More recently, BiSSE has been shown to be prone to falsely identifying a positive association when diversification rate shifts are associated with another character *not* included in the model (Maddison and FitzJohn 2015; Rabosky and Goldberg 2015). One approach to reduce the possibility of falsely associating a character with diversification rate heterogeneity is to incorporate a second, unobserved character into the model (i.e., a Hidden State Speciation and Extinction (HiSSE) model; Beaulieu and OMeara 2016). The changes in the unobserved character's state represent background diversification rate changes that are not correlated with the observed character. Our work here is the first to apply a HiSSE-type model to test Stebbins' dead-end hypothesis. We also use simulations and Bayes factors (Kass and Raftery 1995) to evaluate the false positive error rate of our model. Additionally, we employ our novel stochastic character mapping method to reconstruct the timing of both diversification rate shifts and transitions in mating system over a densely sampled fossil calibrated phylogeny of Onagraceae. Finally, we test the hypothesis that SC lineages have higher extinction and speciation rates yet lower net diversification rates compared to SI lineages.

3.2 Methods

3.2.1 Stochastic Character Mapping Method

The primary steps of the novel stochastic character mapping algorithm introduced here are illustrated in Fig. 3.1. A pseudocode formulation of the algorithm is provided in the Supporting Information (Alg. 1). Additionally, Supporting Information Fig. 3.6 gives a side by side comparison of the standard stochastic character mapping algorithm as originally described by Nielsen (2002) and the approach introduced in this work. In standard stochastic character mapping the first step is to traverse the tree post-order (tip to root) calculating the conditional likelihood of the character being in each state at each node using Felsenstein's pruning algorithm (Felsenstein 1981). Transition probabilities are computed along each branch using matrix exponentiation. Ancestral states are then sampled at each node during a pre-order (root to tip) traversal. Finally, character histories are repeatedly simulated using rejection sampling for each branch of the tree.

In our new stochastic character mapping algorithm we begin similarly by traversing the tree post-order and calculating conditional likelihoods. However, instead of using matrix exponentiation we calculate the likelihood using a set of ordinary differential equations. We numerically integrate these equations for every arbitrarily small time interval along each branch and store a vector of conditional likelihoods for the character being in each state for every small time interval. The two functions we must numerically integrate are $D_{N,i}(t)$ which is defined as the probability that a lineage in state i at time t evolves into the observed clade N, and $E_i(t)$ which is the probability that a lineage in state i at time t goes extinct before the present, or is not sampled at the present. The equations for these two probabilities are given as Supporting Information Eq.3.6 and Eq. 3.5. Note these equations are identical to the ones

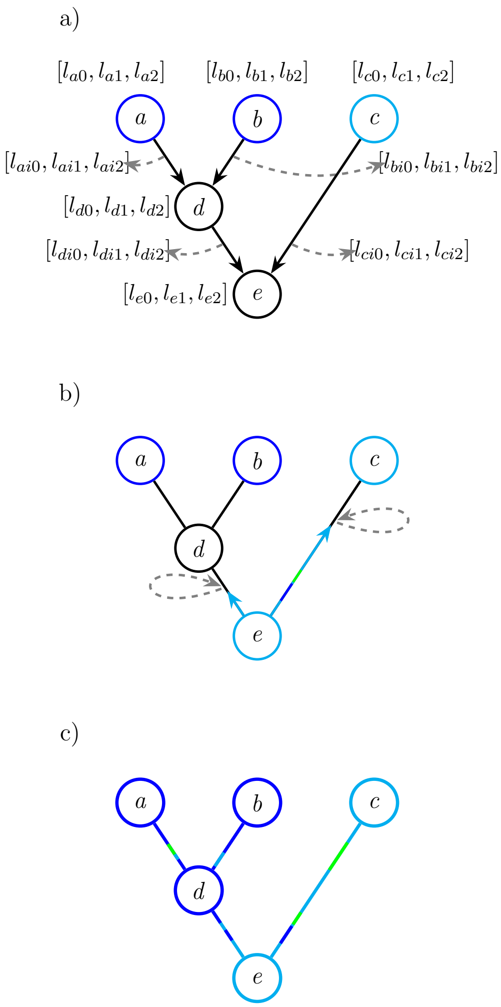
a)
$[l_{a0}, l_{a1}, l_{a2}]$
$[l_{b0}, l_{b1}, l_{b2}]$
$[l_{c0}, l_{c1}, l_{c2}]$
a
b
c
$[l_{ai0}, l_{ai1}, l_{ai2}]$
$[l_{bi0}, l_{bi1}, l_{bi2}]$
$[l_{d0}, l_{d1}, l_{d2}]$
d
$[l_{di0}, l_{di1}, l_{di2}]$
$[l_{ci0}, l_{ci1}, l_{ci2}]$
$[l_{e0}, l_{e1}, l_{e2}]$
e
b)
a
b
c
d
e
c)
a
b
c
d
e

Figure 3.1: Schematic of the new stochastic character mapping method introduced in this work. The first step in the stochastic character mapping method introduced in this work is *(a)* traversing the tree post-order (tip to root) calculating conditional likelihoods for every arbitrarily small time interval along each branch and at nodes. Next, during a pre-order traversal (root to tip) ancestral states are sampled for each time interval *(b)*, resulting in a full character history *(c)* without the need for a rejection sampling step. See Supporting Information Fig. 3.6 for a side by side comparison of the standard stochastic character mapping algorithm as originally described by Nielsen (2002) and the approach introduced in this work.

describing the Cladogenetic State Speciation and Extinction model (ClaSSE; Goldberg and Igić 2012), which all other discrete SSE models are nested within.

At the tips of the phylogeny (time $t = 0$) the extinction probabilities are $E_i(0) = 1 - \rho$ for all i where ρ is the sampling probability of including that lineage. For lineages with the observed state i, the initial condition is $D_{N,i}(0) = \rho$. The initial condition for all other states j is $D_{N,j}(0) = 0$. When a node L is reached, the probability of it being in state i is calculated by combining the probabilities of its descendant nodes M and N as such:

$$D_{L,i}(t) = \sum_j \sum_k \lambda_{ijk} D_{M,j}(t) D_{N,k}(t), \tag{3.1}$$

where the rate of a lineage in state i splitting into two lineages in states j and k is λ_{ijk}. Letting $\mathcal{X}$ represent the observed tip data, Ψ an observed phylogeny, and θ_q a particular set of character evolution model parameters, then the likelihood is given by:

$$P(\mathcal{X}, \Psi | \theta_q) = \sum_i \pi_i D_{R,i}(t), \tag{3.2}$$

where π_i is the root frequency of state i and $D_{R,i}(t)$ is the likelihood of the root node being in state i conditional on having given rise to the observed tree Ψ and the observed tip data $\mathcal{X}$ (Maddison et al. 2007; FitzJohn 2012).

We then sample a complete character history during a pre-order tree traversal in which the root state is first drawn from the marginal likelihoods at the root, and then states are drawn for each small time interval moving towards the tip of the tree conditioned on the state of the previous small time interval. We must again numerically integrate over a set differential equations during this root-to-tip tree traversal. This integration, however, is performed in forward-time, thus a different and new set of differential equations must be used. Letting the rate of anagenetic change from state i to j to be Q_{ij} and the rate of extinction in state i to be μ_i:

$$\begin{aligned} E_i(t - \Delta t) \approx E_i(t) - \Bigg[\mu_i - \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \Bigg) E_i(t) \\ + \sum_{j \neq i} Q_{ij} E_j(t) + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t) \Bigg] \Delta t, \end{aligned} \tag{3.3}$$

$$\begin{aligned} D_{N,i}(t - \Delta t) \approx D_{N,i}(t) + \Bigg[- \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \Bigg) D_{N,i}(t) \\ + \sum_{j \neq i} Q_{ji} D_{N,j}(t) + D_{N,j}(t) E_k(t) \Bigg(\sum_j \sum_k \lambda_{jik} + \sum_j \sum_k \lambda_{jki} \Bigg) \Bigg] \Delta t. \end{aligned} \tag{3.4}$$

In the Supporting Information we derive these forward-time differential equations. We demonstrate how the forward-time equations correctly handle non-reversible models of character evolution and validate the forward-time computation of $D_{N,i}(t)$ and $E_i(t)$. With this approach we can directly sample character histories from an SSE process in forward-time, resulting in a complete stochastic character map sample without the need for rejection sampling or uniformization, see Figure 3.1.

Implementation

The stochastic character mapping method described here is implemented in `C++` in the software `RevBayes` (Höhna et al. 2014b, 2016). The `RevGadgets R` package (available at `https://github.com/revbayes/RevGadgets`) can be used to generate plots from `RevBayes` output. Scripts to run all `RevBayes` analyses presented here can be found in the repository at `https://github.com/wf8/onagraceae`.

3.2.2 Onagraceae Phylogenetic Analyses

DNA sequences for Onagraceae and Lythraceae were mined from GenBank using SUMAC (Freyman 2015). Lythraceae was selected as an outgroup since previous molecular phylogenetic analyses place it sister to Onagraceae (Sytsma et al. 2004). Information about the alignments and GenBank accessions used can be found in the Supporting Information. Phylogeny and divergence times were inferred using RevBayes (Höhna et al. 2016). Details regarding the fossil and secondary calibrations, the model of molecular evolution, and MCMC analyses are given in the Supporting Information.

3.2.3 Analyses of Mating System Evolution

The mating system of Onagraceae species were scored as either self-compatible or self-incompatible following Wagner et al. (Wagner et al. 2007).

HiSSE Model

To test whether diversification rate heterogeneity is associated with shifts in mating system or changes in other unmeasured traits, we used a model with 4 states that describes the joint evolution of mating system as well as an unobserved character with hidden states a and b (Fig. 3.2). Since the RNase-based gametophytic system of self-incompatibility found in Onagraceae is ancestral for all eudicots (Steinbachs and Holsinger 2002), we used an irreversible model that only allowed transitions from self-incompatible to self-compatible. For each of the 4 states we estimated speciation (λ) and extinction (μ) rates. While estimating diversification rates, we accounted for uncertainty in phylogeny and divergence times by sampling 200 trees from the posterior distribution of trees. For details on priors used and the MCMC analyses see the Supporting Information.

Model Comparisons and Error Rates

To test whether diversification rate heterogeneity was *not* associated with shifts in mating system, we calculated a Bayes factor (Kass and Raftery 1995) to compare the mating system dependent diversification model described above with a mating system independent diversification model. The independent model had 4 states and the same parameters as the dependent model, except that the speciation and extinction rates were fixed so they only varied between the hidden states a and b. Hence, λ_{ca} was fixed to equal λ_{ia}, λ_{cb} was fixed to λ_{ib}, μ_{ca} was fixed to μ_{ia}, and μ_{cb} was fixed to μ_{ib}.

To evaluate the false positive error rate we performed a series of simulations that tested the power of our models to reject false associations between shifts in mating system and diversification rate shifts. Trees were simulated under a BiSSE model, and then diversification *independent* binary characters representing mating system were simulated over the trees. For each simulation replicate Bayes factors were calculated to compare the fit of the mating system dependent diversification model and mating system independent diversification model. Details on the simulations are provided in the Supporting Information.

All Bayes factors were calculated using the stepping stone method (Xie et al. 2011; Höhna et al. 2017) as implemented in RevBayes. Marginal likelihood estimates were run for 50 path

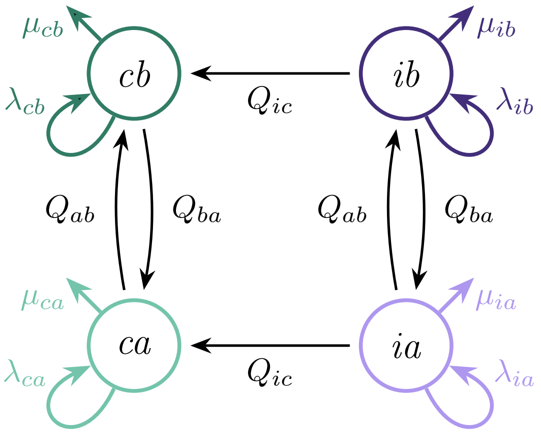

Figure 3.2: SSE model depicting states and rate parameters used to infer mating system evolution. The states are labeled ca, cb, ia, and ib, representing self-compatible hidden state a, self-compatible hidden state b, self-incompatible hidden state a, and self-incompatible hidden state b, respectively. Independent extinction μ and speciation λ rates were estimated for each of the 4 states, as well as the rate of transitioning from self-incompatible to self-compatible Q_{ic} and the rates of transitioning between the hidden states Q_{ab} and Q_{ba}.

steps and 19000 generations within each step. The Bayes factor was then calculated as twice the difference in the natural log marginal likelihoods (Kass and Raftery 1995).

3.3 Results

3.3.1 Onagraceae Phylogeny

In our estimated phylogeny, all currently recognized Onagraceae genera (Wagner et al. 2007) were strongly supported to be monophyletic with posterior probabilities > 0.98. The crown age of Onagraceae was estimated to be 98.8 Ma (94.0 Ma – 107.3 Ma 95% HPD; Fig. 3.3), and a summary of the divergence times of major clades within Onagraceae can be found in Supporting Information Table 3.3.

3.3.2 Model Comparisons and Error Rates

The state-dependent diversification model of mating system evolution (Fig. 3.2) was "decisively" supported over the state-independent diversification model with a Bayes factor ($2ln$BF) of 19.9 (Jeffreys 1961). Bayes factors calculated using simulated datasets showed that the false positive error rate was low. The false positive rate for "strong" support ($2ln$BF > 6; Kass and Raftery 1995) was 0.05, and the false positive rate for "very strong" support ($2ln$BF > 10; Kass and Raftery 1995) was 0.0.

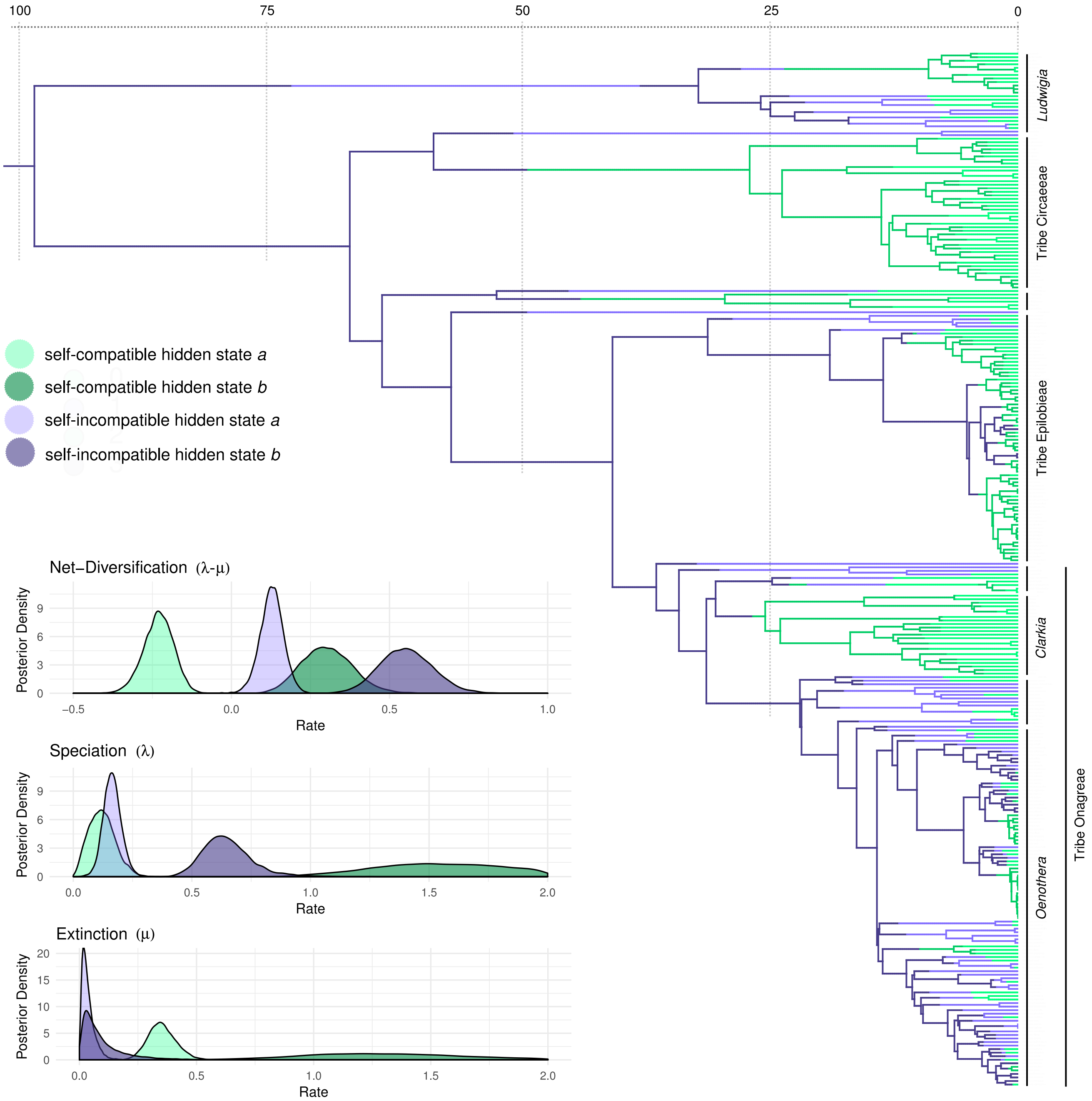

Figure 3.3: Maximum a posteriori reconstruction of mating system evolution and shifts in diversification rates in Onagraceae. Divergence times in millions of years are indicated by the axis at the top. The inset panels show posterior densities of net diversification ($\lambda - \mu$), speciation (λ), and extinction (μ) rates in millions of years. Changes in mating system and an unobserved character (hidden states a and b) are both associated with diversification rate heterogeneity. Within either hidden state (a or b) self-compatible lineages have higher extinction and speciation rates yet lower net diversification rates compared to self-incompatible lineages.

3.3.3 Stochastic Character Maps

Under the state-dependent diversification model, repeated independent losses of SI across the Onagraceae phylogeny were found to be associated with shifts in diversification rates (Fig. 3.3). Additionally, transitions between the unobserved character states a and b were also associated with diversification rate heterogeneity. Uncertainty in the timing of diversification rate shifts and character state transitions was generally low, but increased along long branches where there was relatively little information regarding the exact timing of transitions (Fig. 3.4). Following the loss of self-incompatibility, there was an evolutionary time lag (mean 18.3 My) until net diversification (speciation minus extinction) turned negative (Fig. 3.5).

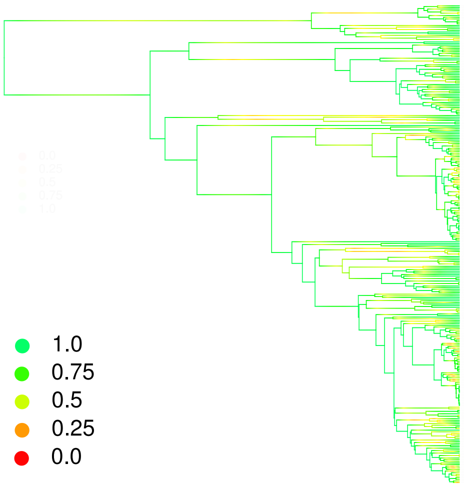

Figure 3.4: Posterior probabilities of the maximum a posteriori reconstruction of mating system evolution and shifts in diversification rates in Onagraceae. Marginal posterior probabilities of the character states shown in Fig. 3.3. Uncertainty was highest along long branches where there was relatively little information regarding the timing of transitions.

3.3.4 Diversification Rate Estimates

Within either hidden state (a or b) SC lineages had generally higher speciation and extinction rates compared to SI lineages (Fig. 3.3). SC lineages in state a had a speciation rate of 0.12 (0.02 – 0.23 95% HPD) compared to 0.16 (0.09 – 0.24 95% HPD) in SI lineages in state a. For SC lineages in state b the speciation rate was 1.66 (0.98 – 2.41 95% HPD) compared to 0.65 (0.45 – 0.85 95% HPD) in SI lineages in state b. Similarly, SC lineages in state a had an extinction rate of 0.35 (0.25 – 0.48 95% HPD) compared to 0.04 (0.00 – 0.09 95% HPD) in SI lineages in state a. For SC lineages in state b the extinction rate was 1.36 (0.65 – 2.19 95% HPD) compared to 0.10 (0.00 – 0.29 95% HPD) in SI lineages in state b.

Despite higher speciation and extinction rates, SC lineages had lower net diversification compared to SI lineages. Net diversification was found to be negative for most but not all extant SC lineages. The net diversification rate for SC lineages in state a was -0.23 (-0.32 – -0.14 95% HPD), compared to 0.13 (0.05 – 0.19 95% HPD) in SI lineages in state a. For SC

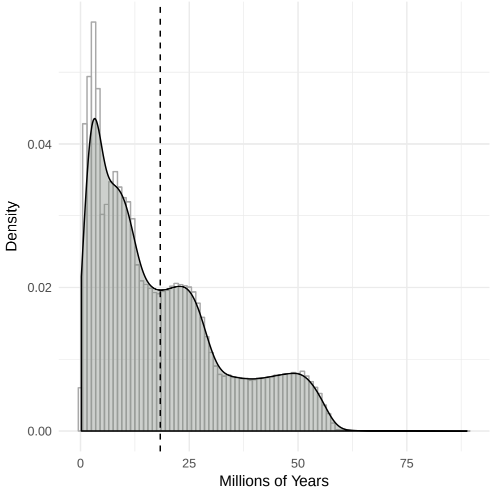

Figure 3.5: The time lag from the loss of self-incompatibility until the onset of evolutionary decline. The time in millions of years after the loss of self-incompatibility until the net diversification rate became negative measured over 10000 stochastic character map samples. The mean time lag until evolutionary decline was 18.3 million years.

lineages in state *b* the net diversification rate was 0.30 (0.15 – 0.46 95% HPD), compared to 0.55 (0.39 – 0.71 95% HPD) in SI lineages in state *b*.

3.4 Discussion

The stochastic character map results reveal that the loss of SI has different short term and long term macroevolutionary consequences. Lineages with relatively recent losses of SI like *Epilobium* are undergoing a burst in both speciation and extinction rates with a positive net diversification rate. However, lineages that have long been SC such as *Fuchsia* (Tribe Circaeeae) and *Clarkia* are in a previously unrecognized evolutionary decline. These lineages went through an increase in both speciation and extinction rates a long time ago —after the loss of SI— but now only the extinction rates remain elevated and the speciation rates have declined, resulting in negative net diversification. The stochastic character maps quantify the speed of this evolutionary decline in SC lineages; while the mean time until evolutionary decline was 18.3 My, there was a large amount of variation in time estimates (Fig. 3.5). This variation could be due to differences in realized selfing/outcrossing rates of different lineages. Lineages with higher selfing rates likely build up inbreeding depression more quickly, which could lead to a more rapid evolutionary decline. Furthermore, even if inbreeding depression is low, the loss of genetic variation in highly selfing lineages will reduce the probability that such lineages can respond adequately to natural selection, such as imposed by a changing or new environment, thus increasing potential for extinction.

These results confirm long-standing theory about the macroevolutionary consequences of SC (Darwin 1876; Stebbins 1957). These consequences include the increased probability of going extinct due to inbreeding (Charlesworth and Charlesworth 1987) and an increased rate of speciation which may be driven by higher among-population differentiation and reproductive assurance that facilitates colonization of new habitats (Baker 1955; Hartfield 2016). The advantages of reproductive assurance may explain why transitions to SC occur repeatedly (Igic et al. 2008; Lande and Schemske 1985). However, our results reveal that this advantage is short-lived; the burst of increased speciation following the loss of SI eventually declines, possibly due to decreased variation resulting from inbreeding. The overall macroevolutionary pattern is one in which SC lineages undergo rapid bursts of increased speciation that eventually decline, doomed by intensified extinction and thus supporting Stebbins' hypothesis of SC as an evolutionary dead-end (Stebbins 1957).

Our findings corroborate previous analyses performed in the plant family Solanceae (Goldberg et al. 2010), where SC lineages were also found to have higher speciation and extinction rates yet lower net diversification. Our results, however, are the first to show that this pattern is supported even when other unmeasured factors affect diversification rate heterogeneity. Intuitively it is clear that no single factor drives all diversfication rate heterogeneity in diverse and complex clades such as Onagraceae. Indeed, in some lineages of *Oenothera* the loss of sexual recombination and segregation due to extensive chromosome translocations (a condition called Permanent Translocation Heterozygosity) is associated with increased diversification rates (Johnson et al. 2011). Furthermore, other factors such as polyploidy and shifts in habitat, growth form, or life cycle may impact diversification rates (Mayrose et al. 2011; Donoghue 2005; Eriksson and Bremer 1992).

Stochastic character mapping of state-dependent diversification can be a powerful tool for examining the timing and nature of both shifts in diversification rates and character state transitions on a phylogeny. Character mapping reveals which stages of the unobserved character a lineage goes through; e.g. after the loss of self-incompatibility transitions are predominantly from hidden state b to a, representing shifts from positive net diversification to negative net diversification. Furthermore, character mapping infers the state of the lineages in the present and so reveals which tips of the phylogeny are currently undergoing positive or negative net diversification. Distributions of character map samples could be used for posterior predictive assessments of model fit (Nielsen 2002; Bollback 2006; Höhna et al. 2017) and for testing whether multiple characters coevolve (Huelsenbeck et al. 2003; Bollback 2006). Our hope is that these approaches enable researchers to examine the macroevolutionary impacts of the diverse processes shaping the tree of life with increasing quantitative rigor.

3.5 Supporting Information

3.5.1 Stochastic Character Mapping Method

Comparison of Algorithms

Our novel stochastic character mapping method is formulated as Algorithm (1). A side by side illustration comparing the primary steps of the standard stochastic character mapping algorithm originally described by Nielsen (2002) and Algorithm (1) is provided in Figure 3.6. There are three primary differences in the two algorithms. First, in the original algorithm likelihoods are calculated using matrix exponentiation whereas in the new algorithm likelihoods are calculated using numerical integration of ordinary differential equations. Second, during the post-order tree traversal the original algorithm stores conditional likelihoods of the process being in each state only at the nodes, whereas in the new algorithm they are stored for every very small time interval between which we apply a numerical integration algorithm. Third, during a pre-order tree traversal the original algorithm estimates ancestral states at all nodes and then uses rejection sampling to simulate branch histories, whereas the new method simulates branch histories by sampling directly from the process in forward time using numerical integration. The central challenge in developing our new method was deriving and validating the forward time differential equations necessary for this last pre-order tree traversal step.

Derivation of our differential equations

In the following section we will derive the differential equations for our algorithm to compute the probability of the observed lineages and the extinction probabilities both backwards and forwards in time. We additionally show how the forward-time equations must be modified to handle non-reversible models of character evolution when sampling ancestral states or stochastic character maps.

Differential equations backwards in time

The original derivation of the differential equations for the state-dependent speciation and extinction (SSE) process look backward in time (Maddison et al. 2007). Here we present a generalization of the SSE process to allow for cladogenetic events where daughter lineages may inherit different states (Goldberg and Igić 2012; Ng and Smith 2014). We repeat this known derivation of the backwards process to show the similarities to our forward in time derivation. We present an overview of the possible scenarios of what can happen in a small time interval Δt in Figure 3.7. We need to consider all these scenarios in our differential equations.

First, let us start with the computation of the extinction probability. That is, we want to compute the probability of a lineage going extinct at time $t + \Delta t$, denoted by $E(t + \Delta t)$, before the present time $t = 0$. We assume that we know the extinction probability of a lineage at time t, denoted by $E(t)$, which is provided by our initial condition that $E(t = 0) = 0$ because the probability of a lineage alive at the present cannot go extinct before the present, or $E(t = 0) = 1 - \rho$ in the case of incomplete taxon sampling. We have five different cases (top row in Figure 3.7): (1) the lineage goes extinct within the interval Δt; (2) nothing happens in the interval Δt but the lineage eventually goes extinct before the present; (3) a state-change to state j occurs and the lineage now in state j goes extinct before the present; (4) the lineage speciates, giving birth to a left daughter lineage in state j and a right daughter lineage in state k and both lineages eventually go extinct before the present, or; (5) the lineage speciates, giving birth to a left daughter lineage in state k and a right daughter lineage in state j and both lineages eventually go extinct before the present. With this description of

Algorithm 1 Stochastic character mapping algorithm. $D_{Ni}(t)$ is the probability that a lineage in state i at time t evolves into the observed clade N. $E_i(t)$ is the probability that a lineage in state i at time t goes extinct or is not sampled before the present.

```
Inputs:
    $\mathcal{X}$: the vector of observed tip states.
    $t_r$: the starting time of the process.
    $\pi$: the vector of root state frequencies.
    $\lambda$: the vector of speciation rates.
    $\mu$: the vector of extinction rates.
    $\rho$: the probability of sampling a lineage in the present.
    $Q$: The matrix of transition rates between states.
Initialize:
    $t \leftarrow 0$                                        // start at the present
    $E_i(t=0) \leftarrow 1-\rho$                            // extinction probability at present time
    if $i = \mathcal{X}_{\text{observed}}$ then
        $D_{N,i}(t=0) \leftarrow \rho$                      // probability of observed character
    else
        $D_{N,i}(t=0) \leftarrow 0$
while $t \le t_r$ do                                        // post-order tree traversal
    if node $L$ is reached then
        $D_{L,i}(t) \leftarrow \sum_j \sum_k \lambda_{ijk} D_{M,j}(t) D_{N,k}(t)$      // combine descendant probabilities
    else
        $L_{N,i}(t) \leftarrow D_{N,i}(t)$                  // store the conditional likelihoods for this time interval
        $E_i(t+\Delta t) \leftarrow E_i(t) +$               // compute conditional likelihoods for next time interval
            $\Bigg[\mu_i - \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \ne i} Q_{ij} + \mu_i\Bigg) E_i(t)$
            $+ \sum_{j \ne i} Q_{ij} E_j(t) + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t)\Bigg] \Delta t$      // backward-time Equation (3.5)
        $D_{N,i}(t+\Delta t) \leftarrow D_{N,i}(t) +$
            $\Bigg[- \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \ne i} Q_{ij} + \mu_i\Bigg) D_{N,i}(t) + \sum_{j \ne i} Q_{ij} D_{N,j}(t)$
            $+ \sum_j \sum_k \lambda_{ijk} \Bigg(D_{N,k}(t) E_j(t) + D_{N,j}(t) E_k(t)\Bigg)\Bigg] \Delta t$      // backward-time Equation (3.6)
        $t \leftarrow t + \Delta t$                         // increment the current $t$
    end if
end while
while $t \ge 0$ do                                          // pre-order tree traversal
    if $t = t_r$ then
        $s_t \sim \text{Multinomial}\big(n = 1, D_N(t_r) \times \pi\big)$       // draw character state at the root
    else
        $s_t \sim \text{Multinomial}\big(n = 1, D_N(t) \times L_N(t)\big)$      // draw character state for time $t$
    end if
    $D_{N,s_t} \leftarrow 1$                                // condition on the sampled character state
    $D_{N,i \ne s_t} \leftarrow 0$
    $E_i(t-\Delta t) \leftarrow E_i(t) -$
        $\Bigg[\mu_i - \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \ne i} Q_{ij} + \mu_i\Bigg) E_i(t)$
        $+ \sum_{j \ne i} Q_{ij} E_j(t) + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t)\Bigg] \Delta t$      // forward-time Equation (3.8)
    $D_{N,i}(t-\Delta t) \leftarrow D_{N,i}(t) +$
        $\Bigg[- \Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j \ne i} Q_{ij} + \mu_i\Bigg) D_{N,i}(t)$
        $+ \sum_{j \ne i} Q_{ji} D_{N,j}(t) + D_{N,j}(t) E_k(t) \Bigg(\sum_j \sum_k \lambda_{jik} + \sum_j \sum_k \lambda_{jki}\Bigg)$      // forward-time Equation (3.10)
    $t \leftarrow t - \Delta t$                             // decrement the current $t$
end while
return vector of all sampled character states $s$
```

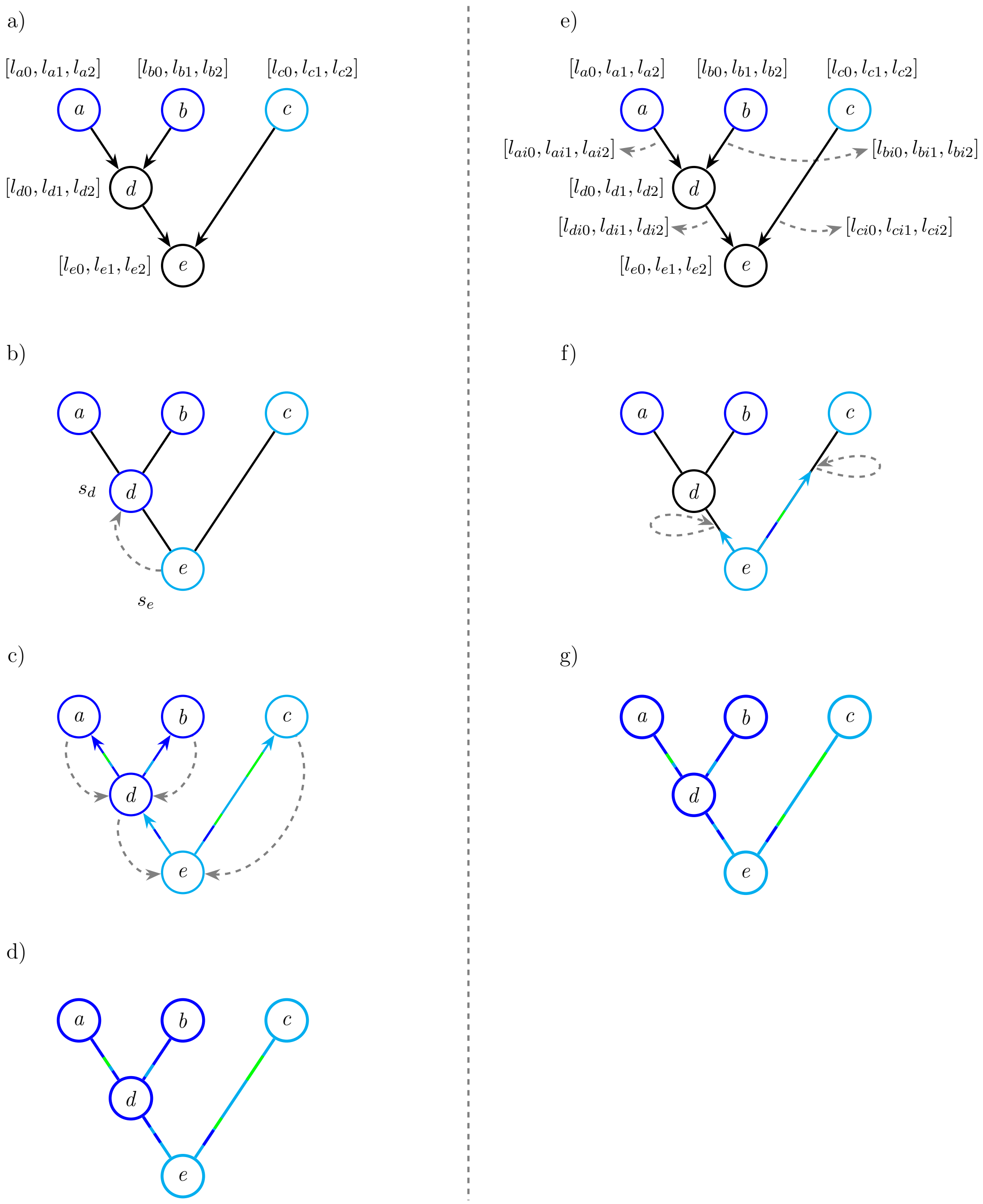
a)
$[l_{a0}, l_{a1}, l_{a2}]$
$[l_{b0}, l_{b1}, l_{b2}]$
$[l_{c0}, l_{c1}, l_{c2}]$
a
b
c
$[l_{d0}, l_{d1}, l_{d2}]$
d
$[l_{e0}, l_{e1}, l_{e2}]$
e
e)
$[l_{a0}, l_{a1}, l_{a2}]$
$[l_{b0}, l_{b1}, l_{b2}]$
$[l_{c0}, l_{c1}, l_{c2}]$
$[l_{ai0}, l_{ai1}, l_{ai2}]$
$[l_{bi0}, l_{bi1}, l_{bi2}]$
$[l_{d0}, l_{d1}, l_{d2}]$
$[l_{di0}, l_{di1}, l_{di2}]$
$[l_{ci0}, l_{ci1}, l_{ci2}]$
$[l_{e0}, l_{e1}, l_{e2}]$
b)
s_d
s_e
f)
c)
g)
d)

Figure 3.6: Comparison of stochastic character mapping methods. On the left *(a, b, c, d)* is an illustration of the standard stochastic character mapping algorithm as originally described by Nielsen (2002). On the right *(e, f, g)* is the approach introduced in this work. The first step in standard stochastic character mapping is *(a)* traversing the tree post-order (tip to root) calculating conditional likelihoods for each node. Next, ancestral states are sampled at each node during a pre-order (root to tip) traversal *(b)*. Branch by branch, character histories are then repeatedly simulated using rejection sampling *(c)*, resulting in a full character history *(d)*. The first step in the stochastic character mapping method introduced in this work is *(e)* traversing the tree post-order calculating conditional likelihoods for every arbitrarily small time interval along each branch and at nodes. Next, during a pre-order traversal ancestral states are sampled for each time interval *(f)*, resulting in a full character history *(g)* without the need for a rejection sampling step. See the main text for more details.

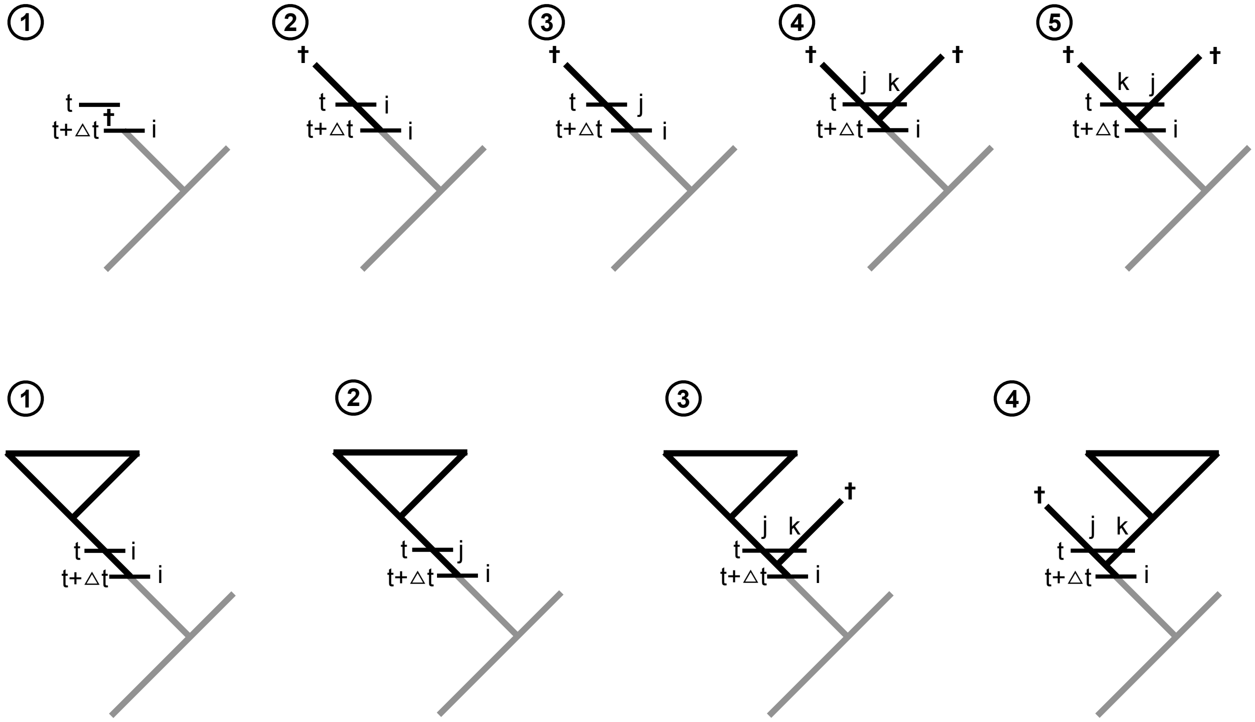

Figure 3.7: Alternative scenarios of events in a small time interval Δt looking backwards in time. The top row shows the different scenarios for a lineage that goes extinct before the present. Case 1: The lineage goes extinct in the time interval Δt. Case 2: There is no event in the time interval Δt and the lineage goes extinct before the present. Case 3: The lineage undergoes a state-shift event to state j in the time interval Δt and the lineage goes extinct before the present. Case 4: The lineage speciates and leaves a left daughter lineage in state j and a right daughter lineage in state k and both daughter lineages go extinct before the present. Case 5: The lineage speciates and leaves a left daughter lineage in state k and a right daughter lineage in state j and both daughter lineages go extinct before the present. The bottom row shows the different scenarios for an observed lineage. Case 1: There is no event in the time interval Δt. Case 2: The lineage undergoes a state-shift event to state j in the time interval Δt. Case 3: The lineage speciates and leaves a left daughter lineage in state j and a right daughter lineage in state k and only the left daughter lineage survives. Case 4: The lineage speciates and leaves a left daughter lineage in state j and a right daughter lineage in state k and only the right daughter lineage survives.

all possible scenarios we can derive the differential equation.

$$\begin{aligned} E_i(t+\Delta t) = E_i(t) + & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.5) \\ & \Bigg[\mu_i & \text{Case (1)} \\ & - \left(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \right) E_i(t) & \text{Case (2)} \\ & + \sum_{j \neq i} Q_{ij} E_j(t) & \text{Case (3)} \\ & + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t) \Bigg] \Delta t & \text{Case (4) and (5)} \end{aligned}$$

Similarly, we can consider all possible scenarios for an observed lineage. We have four different cases (bottom row in Figure 3.7): (1) nothing happens in the interval Δt; (2) a state-change to state j occurs; (3) the lineage speciates, giving birth to a left daughter lineage in state j and a right daughter lineage in state k and only the left daughter lineage survives until the present, or; (4) the lineage speciates, giving birth to a left daughter lineage in state j and a right daughter lineage in state k and only the right daughter lineage survives until the present. Again, these scenarios are sufficient to derive the differential equation for the probability of an observed lineage, denoted $D(t)$.

$$\begin{aligned} D_{N,i}(t+\Delta t) = D_{N,i}(t) + & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.6) \\ & \Bigg[- \left(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \right) D_{N,i}(t) && \text{Case (1)} \\ & + \sum_{j \neq i} Q_{ij} D_{N,j}(t) && \text{Case (2)} \\ & + \sum_j \sum_k \lambda_{ijk} \Big(D_{N,k}(t) E_j(t) + D_{N,j}(t) E_k(t) \Big) \Bigg] \Delta t && \text{Case (3) and (4)} \end{aligned}$$

Differential equations forward in time

Next, we want to compute the probability of extinction and the probability of an observed lineage forward in time. For the probability of extinction this is, in principle, almost identical to the backward in time equations. However, now we assume that we know $E(t)$ and want to compute $E(t - \Delta t)$. We already computed $E(t_{root})$ and $D(t_{root})$ in our post-order tree traversal (from the tips to root). We use $E(t_{root})$ as the initial conditions to approximate $E(t - \Delta t)$. Again, we have the same five different cases (top row in Figure 3.7): (1) the lineage goes extinct within the interval Δt; (2) nothing happens in the interval Δt but the lineage eventually goes extinct before the present; (3) a state-change to state j occurs and the lineage now in state j goes extinct before the present; (4) the lineage speciates, giving birth to a left daughter lineage in state j and a right daughter lineage in state k and both lineages eventually go extinct before the present, or; (5) the lineage speciates, giving birth to a left daughter lineage in state k and a right daughter lineage in state j and both lineages eventually go extinct before the present. However, these are the events that can happen in the future and we included the probabilities of these events already in $E(t)$. Thus, we need to subtract instead of adding all possible scenarios that lead to the extinction of the lineage in the time interval Δt from $E(t)$ to obtain $E(t - \Delta t)$. This gives us the differential equation for the extinction probability as

$$\begin{aligned} E_i(t-\Delta t) = E_i(t) - & \\ & \Bigg[\mu_i && \text{Case (1)} \\ & -\Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j\neq i} Q_{ij} + \mu_i\Bigg) E_i(t) && \text{Case (2)} \\ & + \sum_{j\neq i} Q_{ij} E_j(t-\Delta t) && \text{Case (3)} \\ & + \sum_j \sum_k \lambda_{ijk} E_j(t-\Delta t) E_k(t-\Delta t)\Bigg] \Delta t && \text{Case (4) and (5)} \end{aligned} \tag{3.7}$$

Unfortunately, we cannot solve Equation (3.7) directly because we do not know $E_j(t-\Delta t)$ and $E_k(t-\Delta t)$. Instead, we will approximate Equation (3.7) by using $E_j(t)$ instead of $E_j(t-\Delta t)$, and $E_k(t)$ instead of $E_k(t-\Delta t)$, respectively. Our approximation yields the new differential equation of the extinction probability by

$$\begin{aligned} E_i(t-\Delta t) \approx E_i(t) - & \\ & \Bigg[\mu_i && \text{Case (1)} \\ & -\Bigg(\sum_j \sum_k \lambda_{ijk} + \sum_{j\neq i} Q_{ij} + \mu_i\Bigg) E_i(t) && \text{Case (2)} \\ & + \sum_{j\neq i} Q_{ij} E_j(t) && \text{Case (3)} \\ & + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t)\Bigg] \Delta t && \text{Case (4) and (5)} \end{aligned} \tag{3.8}$$

The derivation of the probability of an observed lineage in forward time is slightly different. When sampling a character history from the process we must compute $D(t-\Delta t)$ conditioned upon the character state sampled at time t. This does not effect the probability of a lineage going extinct before the present, so we can use $E(t_{root})$ as the initial conditions to approximate $E(t-\Delta t)$. The initial conditions for the probability of an observed lineage, on the other hand, must account for the sampled character state. For example, if we sample the state a at time t our initial conditions to compute $D(t-\Delta t)$ must be $D_a(t) = 1.0$ and $D_b(t) = 0.0$ for all other character states b. Additionally, we must consider the process in forward time with all possible scenarios instead of backwards in time and subtracting the possible scenarios. We have four different cases that are similar to the cases for the backward in time computation (bottom row in Figure 3.7), however here the character state transitions are reversed since we are looking forward in time: (1) nothing happens in the interval Δt; (2) with probability $D_{N,j}(t)$ the lineage was in state j and then a state-change to state i

occurs; (3) with probability $D_{N,j}(t)$ the lineage was in state j and then speciates, giving birth to a left daughter lineage in state i and a right daughter lineage in state k and only the left daughter lineage survives until the present (the probability of extinction of the right daughter lineage is given by $E_k(t-\Delta t)$), or; (4) with probability $D_{N,j}(t)$ the lineage was in state j and then speciates, giving birth to a left daughter lineage in state k and a right daughter lineage in state i and only the right daughter lineage survives until the present (the probability of extinction of the left daughter lineage is given by $E_k(t-\Delta t)$). From these four scenarios we derive the differential equation.

$$\begin{aligned} D_{N,i}(t-\Delta t) = D_{N,i}(t) + & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.9) \\ & \Bigg[- \left(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \right) D_{N,i}(t) \qquad \text{Case (1)} \\ & + \sum_{j \neq i} Q_{ji} D_{N,j}(t) \qquad \text{Case (2)} \\ & + D_{N,j}(t) E_k(t-\Delta t) \left(\sum_j \sum_k \lambda_{jik} + \sum_j \sum_k \lambda_{jki} \right) \Bigg] \Delta t \quad \text{Case (3) and (4)} \end{aligned}$$

As before, we cannot solve Equation (3.9) directly because we do not know $E_k(t-\Delta t)$. Thus, we use the same approximation as before and substitute $E_k(t)$ for $E_k(t-\Delta t)$. This substitution gives our approximated differential equation.

$$\begin{aligned} D_{N,i}(t-\Delta t) \approx D_{N,i}(t) + & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.10) \\ & \Bigg[- \left(\sum_j \sum_k \lambda_{ijk} + \sum_{j \neq i} Q_{ij} + \mu_i \right) D_{N,i}(t) \qquad \text{Case (1)} \\ & + \sum_{j \neq i} Q_{ji} D_{N,j}(t) \qquad \text{Case (2)} \\ & + D_{N,j}(t) E_k(t) \left(\sum_j \sum_k \lambda_{jik} + \sum_j \sum_k \lambda_{jki} \right) \Bigg] \Delta t \quad \text{Case (3) and (4)} \end{aligned}$$

To sample character histories from an SSE process in forward-time during Algorithm (1) we calculate $E(t-\Delta t)$ using the approximation given by Equation (3.8) and $D(t-\Delta t)$ using Equation (3.10).

Correctness of the forward time equations

For the purpose of demonstrating our forward time equations, we will use a non-symmetrical BiSSE model with states 0 and 1 which have the speciation rates $\lambda_0 = 1$ and $\lambda_1 = 2$, the extinction rates $\mu_0 = 0.5$ and $\mu_1 = 1.5$, and the transition rates $Q_{01} = 0.2$ and $Q_{10} = 2.0$. For simplicity we assume that there are no state changes at speciation events. We will first show that the approximations given by Equation (3.8) actually converge to the true probability of extinction if the time interval Δt is very small (goes to zero). Note that we cannot show the same behavior for the forward in time probabilities of the observed lineage, $D(t)$, because when conditioning on a sampled character state the forward in time probabilities will be different than the backward in time probabilities. For these probabilities we provide a different type of validation in Section 3.5.1.

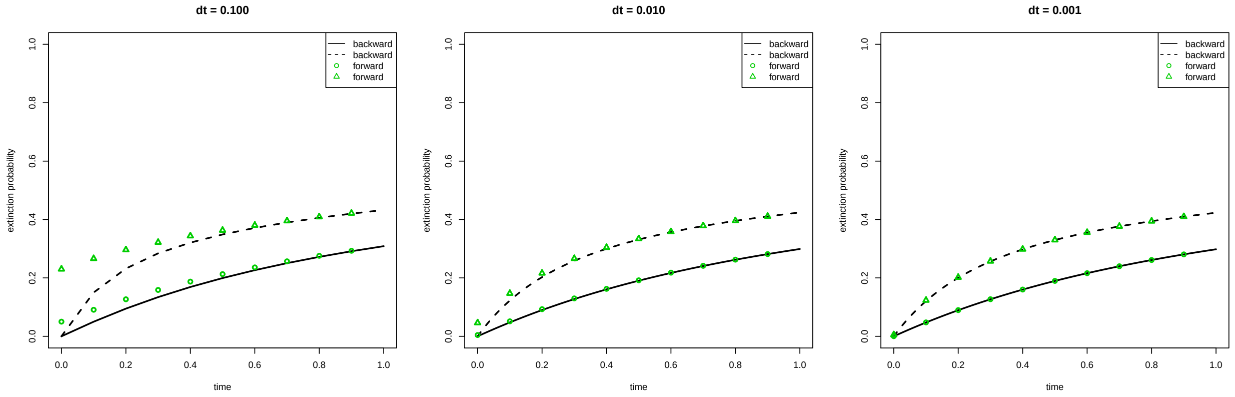

Figure 3.8: The probability of extinction computed backward and forward in time. Here we compute the extinction probabilities $E_0(t)$ and $E_1(t)$ for a BiSSE model backward and forward in time. Details about the parameters of the BiSSE model are given in the text. We varied the step-size Δt for the numerical integration between 0.1, 0.01, and 0.001 to show that both computations give the same probabilities once Δt is small enough.

We start by computing the probability of extinction and the probability of an observed lineage backward in time for a total time interval of 1.0. We initialize the computation with $E_i(t = 0) = 0$ and then compute $E_0(t)$ and $E_1(t)$ backward in time. Then, we use the computed values of $E_i(t = 1)$ as the initial values for our forward in time computation. If our approximation is correct, then we should get identical values for the extinction probabilities $E_i(t)$ for any value of t.

Figure 3.8 shows our computation using three different values for Δt: 0.1, 0.01 and 0.001. We observe that our approximation of the forward in time computation of the probabilities converges to the backward in time computation when $\Delta t \leq 0.001$, which confirms our expectation. An explanation for the convergence is that $E_0(t)$ will be approximately equal to $E_0(t - \Delta t)$, (and $E_1(t)$ to $E_1(t - \Delta t)$) the smaller Δt becomes. In our actual implementation in `RevBayes` we use an initial step-size of $\Delta t = 10^{-7}$ but apply an adaptive numerical integration routine to minimize the error in the integrated function.

Validation of the forward time equations against `diversitree`

Finally, we validate our method of sampling character histories from an SSE process in forward-time by testing it against the analytical marginal ancestral state estimation implemented in the `R` package `diversitree` (FitzJohn 2012). Our method as implemented in `RevBayes` works for sampling both ancestral states and stochastic character maps, however `diversitree` can not sample stochastic character maps. Thus we limit our comparison to ancestral states estimated at the nodes of a phylogeny. Though our method works for all SSE models nested within ClaSSE, ancestral state estimation for ClaSSE is not implemented in `diversitree`, so we further limit our comparison to ancestral state estimates for a BiSSE model. Note that as implemented in `RevBayes` the BiSSE, ClaSSE, MuSSE (FitzJohn 2012),

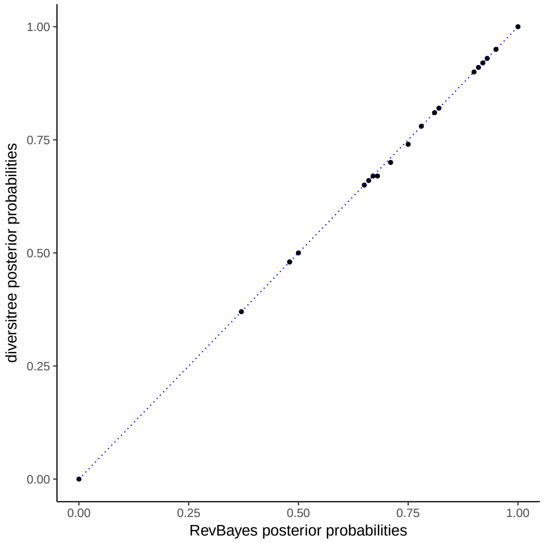

Figure 3.9: Comparing marginal posterior ancestral state estimates from `diversitree` to those calculated in `RevBayes`. Each point represents the posterior probability of a given node having the ancestral state 0. On the y-axis are the posterior probabilities as analytically calculated by `diversitree`. On the x-axis are the posterior probabilities as calculated by `RevBayes` using Algorithm (1). Our approximation given in Equation (3.10) yields the same posterior probabilities of the ancestral states as `diversitree`. Scripts to repeat this test with various parameter settings are provided in `https://github.com/wf8/anc_state_validation`.

HiSSE (Beaulieu and OMeara 2016), ChromoSSE (Freyman and Höhna 2017a), and GeoSSE (Goldberg et al. 2011) models use the same C++ classes and algorithms for parameter and ancestral state estimation, so validating under BiSSE should provide confidence in estimates made by `RevBayes` for all these SSE models.

Our method samples character histories from SSE models from their joint distribution conditioned on the tip states and the model parameters during MCMC. In contrast, `diversitree` computes marginal ancestral states analytically. Thus to directly compare results from these two approaches we calculated the marginal posterior probability of each node being in each state from a set of 10000 samples drawn by our Monte Carlo method. Figure 3.9 compares these estimates under a non-reversible BiSSE model where the tree and tip data were simulated in `diversitree` with the following parameters: $\lambda_0 = 0.2, \lambda_1 = 0.4, \mu_0 = 0.01, \mu_1 = 0.1$, and $q_{01} = 0.1, q_{10} = 0.4$. Figure 3.9 shows that using the approximation of $E(t - \Delta t)$ given by Equation (3.8) and the approximation to compute $D(t - \Delta t)$ in Equation (3.10) during Algorithm (1) results in marginal posterior estimates for the ancestral states that are nearly identical (up to some expected numerical and sampling errors) to those calculated analytically by `diversitree`. Scripts to perform this test with various parameter settings are provided in `https://github.com/wf8/anc_state_validation`.

MCMC Sampling and Computational Efficiency

Our method approximates the posterior distribution of the timing and nature of all character transitions and diversification rate shifts by sampling a large number of stochastically mapped character histories using MCMC. Uncertainty in the phylogeny and other parameters is incorporated by integrating over all possible phylogenetic trees and other parameters jointly. From these sampled character histories the maximum a posteriori character history can be summarized in a number of ways. The approach presented here is to calculate the marginal probabilities of character states for every small time interval along each branch, however one could also calculate the joint posterior probability of an entire character history.

During Algorithm (1) the rate-limiting step is writing conditional likelihood vectors for every small time interval along every branch on the tree, particularly when the state space of the model is large. The time required is of order $O(n \times m \times r)$, where n is the number of taxa in the tree, m is the number of character states, and r is the number of time intervals. This is reduced by only storing conditional likelihood vectors for all time intervals during the MCMC iterations that are sampled. During unsampled (*i.e.*, thinned) MCMC iterations the likelihood is calculated in the standard way storing conditional likelihood vectors only at the nodes, thus the use of the stochastic mapping algorithm has little impact on the overall computation time.

3.5.2 Onagraceae Phylogenetic Analyses

Methods

Supermatrix Assembly

DNA sequences for Onagraceae and Lythraceae were mined from GenBank using `SUMAC` (Freyman 2015). Lythraceae was selected as an outgroup since previous molecular phylogenetic analyses place it sister to Onagraceae (Sytsma et al. 2004). `SUMAC` assembled an 8 gene supermatrix (7 chloroplast loci plus the nuclear ribosomal internal transcribed spacer region) representing a total of 340 taxa. Table 3.1 summarizes the genes used, their length, and the percent of missing data. Sequences were aligned using `MAFFT` v7.123b (Katoh and Standley 2013). The default settings in `MAFFT` were used except that proper sequence polarity was ensured by using the direction adjustment option. Alignments were then concatenated resulting in chimeric operational taxonomic units (OTUs) that do not necessarily represent a single individual.

Table 3.1: DNA regions mined from GenBank. A total of 340 taxa were included.

DNA Region	# Taxa	Aligned Length	# Variable Sites	Missing data (%)	Taxon Coverage Density
ITS	250	904	481	26.5	0.735
matK	42	895	276	87.6	0.124
ndhF	39	1085	429	88.5	0.115
pgiC	66	7664	3828	80.6	0.194
rbcL	108	1427	388	68.2	0.318
rpl16	54	1139	343	84.1	0.159
rps16	78	1056	335	77.1	0.229
trnL-trnF	261	1434	644	23.2	0.768

Table 3.2: Fossil and secondary calibrations used as priors in the Bayesian divergence time analysis. Units are in millions of years.

Group	Calibration Type	Placement	Prior Distribution	Mean	SD	Offset	Reference
Circaea	fossil	stem	lognormal	10	2	12	(Grímsson et al. 2012)
Tribe Epilobieae	fossil	stem	lognormal	10	2	12	(Grímsson et al. 2012)
Fuchsia section *Skinnera*	fossil	stem	lognormal	10	2	23	(Lee et al. 2013)
Lythraceae	fossil	crown	lognormal	20	2	81.5	(Graham 2013)
Ludwigia	fossil	stem	lognormal	10	2	57.6	(Zhi-Chen et al. 2004)
Onagraceae + Lythraceae	secondary	crown	normal	93	5	0	(Sytsma et al. 2004)

Phylogenetic Analyses

Divergence times and phylogeny were jointly estimated using `RevBayes` (Höhna et al. 2014a, 2016). Estimates were time calibrated using six node calibrations: four stem fossil calibrations, one crown fossil calibration, and a secondary calibration for the root split between Onagraceae and Lythraceae (Table 3.2). An uncorrelated lognormal relaxed clock model was used, and each of the eight gene partitions were assigned independent GTR substitution models (Tavaré 1986; Rodriguez et al. 1990). Rate variation across sites was modeled under a gamma distribution approximated by four discrete rate categories (Yang 1994). The constant rate birth-death-sampling tree prior (Nee et al. 1994b; Yang and Rannala 1997) was used with the probability of sampling species at the present (ρ) set to 0.27. ρ was calculated by dividing the number of extant species sampled in the supermatrix (340) by the sum of the number of species recognized in Onagraceae (~650) and in Lythraceae (~620).

Four independent MCMC analyses were performed. Each MCMC ran for 15000 generations, where each generation consisted of 837 randomly scheduled Metropolis-Hastings moves. This resulted in four chains that each performed a total of 12,555,000 MCMC steps. Samples of the posterior distribution were drawn every 10 generations, and the first 50% of samples from each chain were discarded as burnin resulting in 750 trees sampled from each of the 4 independent chains. Convergence was assessed by ensuring the effective sample size of each parameter was over 200 for each independent chain. The maximum a posteriori (MAP) tree was then calculated from the combined 3000 tree samples of all 4 chains.

Results

All Onagraceae genera described in Wagner et al. (2007) were recovered as monophyletic clades in the MAP summary tree with posterior probabilities > 0.95 (Figure 3.10). Onagraceae was found to diverge from Lythraceae at 111.3 My (95% HPD interval 106.0 - 116.6 My). Divergence time estimates of other major clades and 95% HPD intervals can be seen in Table 3.3.

3.5.3 Mating System Evolution Analyses

Model Priors

Model parameter priors are listed in Table 3.4. The rate of loss of self-incompatibility (q_{ic}), and the rates of switching between hidden states a and b (q_{ab} and q_{ba}) were each given an exponential distribution with a mean of n/Ψ_l, where Ψ_l is the length of the tree Ψ and n is

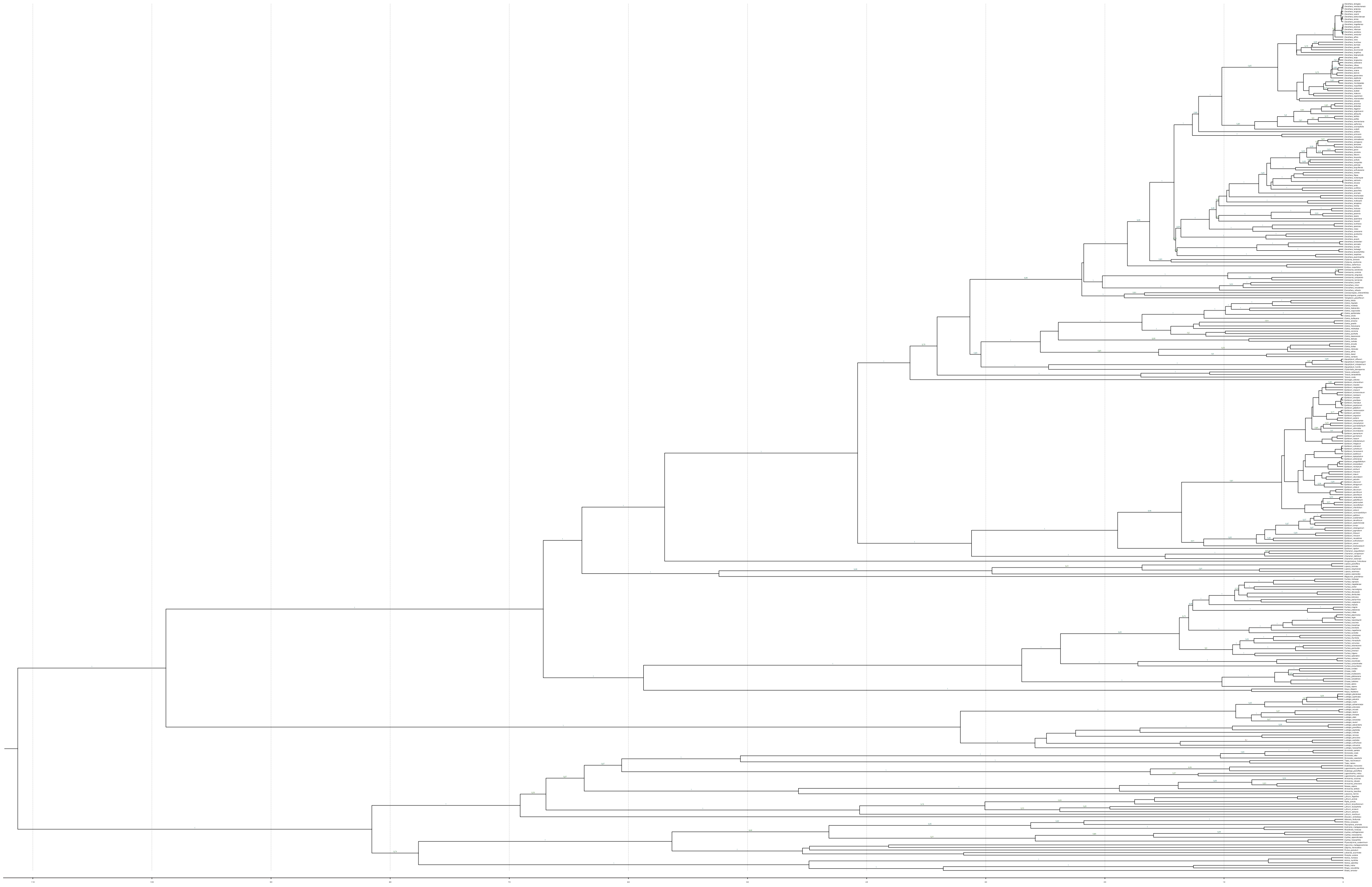

Figure 3.10: Maximum a posteriori estimate of phylogeny and divergence times of Onagraceae. Divergence times in millions of years are indicated by the axis at the bottom. Posterior probabilities > 0.50 are labeled on the branches.

Table 3.3: Divergence time estimates of major clades.

Clade	Age Type	Mean Age (Ma)	95% HPD Min	95% HPD Max
Onagraceae + Lythraceae	crown	111.3	106.0	116.6
Onagraceae	crown	98.8	94.0	107.3
Ludwigia	crown	32.1	31.3	50.6
Tribe Circaeeae	stem	58.7	42.3	65.0
Tribe Circaeeae	crown	27.0	24.6	29.8
Fuchshia	crown	23.7	23.0	24.1
Lopezia	crown	29.5	25.3	36.7
Tribe Epilobieae	stem	40.8	39.3	47.0
Tribe Epilobieae	crown	31.2	31.2	40.6
Chamerion	crown	14.9	12.6	25.2
Epilobium	crown	18.9	15.7	22.9
Tribe Onagreae	stem	40.8	39.3	47.0
Tribe Onagreae	crown	36.4	31.6	40.5
Taraxia	crown	17.0	9.9	19.5
Gayophytum	crown	3.1	1.8	6.1
Clarkia	crown	25.4	24.6	31.2
Eremothera	crown	10.4	8.1	15.1
Camissonia	crown	9.7	6.4	14.6
Eulobus	crown	4.7	2.6	8.1
Chylismia	crown	14.4	10.8	19.0
Oenothera	crown	14.2	13.0	17.6

the expected number of transitions. n was given an exponential hyperprior with a mean of 20.

The speciation and extinction rates were drawn from exponential priors with a mean equal to an estimate of the net diversification rate $\hat{d}$. Under a constant rate birth-death process not conditioning on survival of the process, the expected number of lineages at time t is given by:

$$E(N_t) = N_0 e^{td}, \tag{3.11}$$

where N_0 is the number of lineages at time 0 and d is the net diversification rate $\lambda - \mu$ (Nee et al. 1994b; Höhna 2015). Therefore, we estimate $\hat{d}$ as:

$$\hat{d} = (\ln N_t - \ln N_0)/t, \tag{3.12}$$

where N_t is the number of lineages in the clade that survived to the present, t is the age of the root, and $N_0 = 2$. The root state probabilities π were set to start the process equally in either self-incompatible hidden state a or self-incompatible hidden state b.

MCMC Analyses

To account for uncertainty in phylogeny and divergence times 200 independent MCMC analyses were performed, each sampling a tree from the posterior distribution of trees generated

Table 3.4: Model parameter names and prior distributions. See the main text for complete description of model parameters and prior distributions. Ψ_l represents the length of tree Ψ and $\hat{d}$ is the expected diversification rate under a constant rate birth-death process.

Parameter	X	$f(X)$
Speciation self-incompatible a	λ_{ia}	Exponential($\lambda = 1/\hat{d}$)
Speciation self-incompatible b	λ_{ib}	Exponential($\lambda = 1/\hat{d}$)
Speciation self-compatible a	λ_{ca}	Exponential($\lambda = 1/\hat{d}$)
Speciation self-compatible b	λ_{cb}	Exponential($\lambda = 1/\hat{d}$)
Extinction self-incompatible a	μ_{ia}	Exponential($\lambda = 1/\hat{d}$)
Extinction self-incompatible b	μ_{ib}	Exponential($\lambda = 1/\hat{d}$)
Extinction self-compatible a	μ_{ca}	Exponential($\lambda = 1/\hat{d}$)
Extinction self-compatible b	μ_{cb}	Exponential($\lambda = 1/\hat{d}$)
Rate of loss of self-incompatibility	q_{ic}	Exponential($\lambda = \Psi_l/n$)
Rate of $a \rightarrow b$	q_{ab}	Exponential($\lambda = \Psi_l/n$)
Rate of $b \rightarrow a$	q_{ba}	Exponential($\lambda = \Psi_l/n$)
Expected number of transitions	n	Exponential($\lambda = 1/20$)

during the phylogenetic analyses. All outgroup (Lythraceae) lineages were pruned off. Each MCMC run drew 10000 samples from the posterior distribution, with 190 randomly scheduled Metropolis-Hastings moves per sample. The first 10% of samples from each run were discarded as burnin. For each run, all parameters had effective sample sizes greater than 200, and the mean effective sample size of the posterior across all 200 tree samples was 1161.6. Estimates of the diversification rates were made by combining samples from all 200 independent runs.

3.5.4 Simulations

Simulated Datasets

100 datasets were simulated under a model where the observed binary character was diversification rate independent yet an unobserved binary character drove background diversification rate heterogeneity. First trees were simulated under BiSSE (Maddison et al. 2007) as implemented in the R package `diversitree` (FitzJohn 2012). The binary character represented hidden states a and b with diversification rates $\lambda_a = 1.0, \lambda_b = 2.0, \mu_a = 0.4$, and $\mu_b = 0.1$. The rate of change between hidden states a and b was set to $q_{ab} = q_{ba} = 0.1$. This resulted in trees that were qualitatively similar in shape to the empirically estimated Onagraceae tree, with a mix of early diverging depauperate clades and more rapidly radiating recent clades (Figure 3.12). To simulate incomplete sampling, 55% of the extant tips were randomly pruned off the tree. After pruning, tree samples were discarded unless they had between 100 and 200 sampled lineages that survived to the present. This restriction ensured that the simulated datasets were not too small for reliable inference and yet not so large to be computationally infeasible. Furthermore, we discarded datasets that had fewer than 20% of the tips in either hidden state to ensure that the trees were generated under a sufficiently heterogenous process.

Once the trees were simulated, diversification independent binary characters were simulated over the trees. These characters represented the observed character (mating system) and so were simulated under an irreversible model where the allowed transition occurred with the rate $10/\Psi_s$, where Ψ_s is the length of the simulated tree. This represents an expected 10 irreversible transitions over the length of the tree, and resulted in simulated datasets with a proportion of either state similar to the proportion of self-compatible/self-incompatible in the empirical Onagraceae dataset. These diversification independent characters were then used to calculate Bayes factors that compared the fit of the diversification dependent model to the diversification independent model of mating system. For details on how Bayes factors were calculated see Section 3.2.3. The false positive error rate was calculated as the percent of simulation replicates in which the Bayes factor supported the false dependent model over the true independent model.

Simulation Results

Bayes factors calculated using simulated datasets showed that the false positive error rate was low (Figure 3.11). The false positive rate for "strong" support ($2ln\text{BF} > 6$; Kass and Raftery 1995) was 0.05, and the false positive rate for "very strong" support ($2ln\text{BF} > 10$; Kass and Raftery 1995) was 0.0.

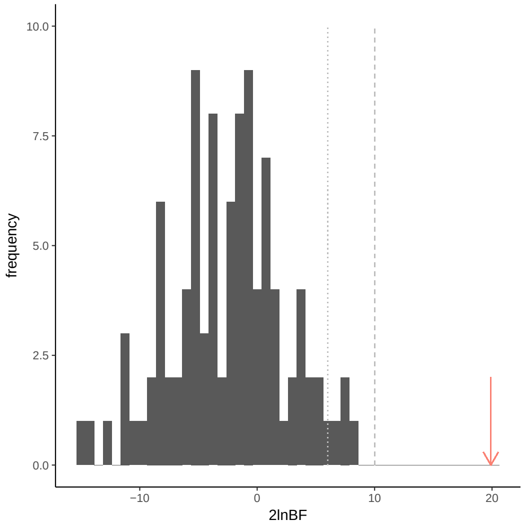

Figure 3.11: Bayes factors ($2ln$BF) comparing the fit of the state-dependent diversification model of mating system evolution with the state-independent diversification model. The red arrow indicates the "decisive" support found for the empirical Onagraceae data ($2ln\text{BF} = 19.9$; Jeffreys 1961). The dark grey bars represent Bayes factors calculated for 100 datasets simulated under a state-independent diversification model. The dotted light grey line indicates "strong" support ($2ln\text{BF} > 6$; Kass and Raftery 1995), and the dashed light grey line indicates "very strong" support ($2ln\text{BF} > 10$; Kass and Raftery 1995).

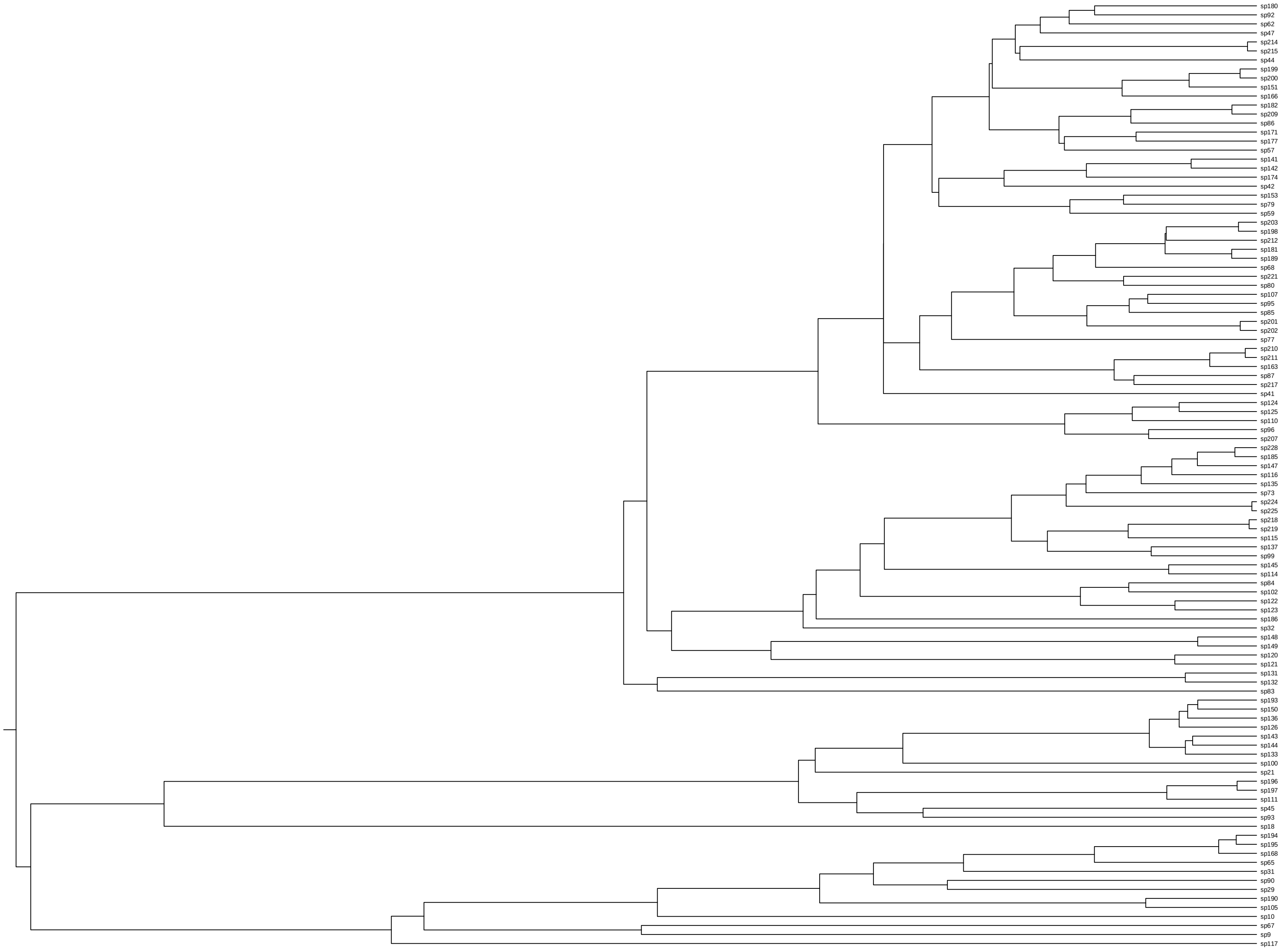

Figure 3.12: One of the trees simulated under BiSSE used to calculate the false positive error rate. Trees were simulated under a heterogenous diversification process to result in a mix of early diverging depauperate clades and more rapidly radiating recent clades. The tree is shown after 55% of the extant lineages were randomly pruned to replicate incomplete sampling in a reconstructed phylogeny.

Bibliography

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Arrigo, N. and M. S. Barker. 2012. Rarely successful polyploids and their legacy in plant genomes. Current Opinion in Plant Biology 15:140–146.

Ayala, F. J. and M. Coluzzi. 2005. Chromosome speciation: humans, *Drosophila*, and mosquitoes. Proceedings of the National Academy of Sciences USA 102:6535–6542.

Baker, H. G. 1955. Self-compatibility and establishment after "long-distance" dispersal. Evolution 9:347–349.

Barrett, S. C. 2002. The evolution of plant sexual diversity. Nature Reviews Genetics 3:274–284.

Beardsley, P. M., S. E. Schoenig, J. B. Whittall, and R. G. Olmstead. 2004. Patterns of evolution in western North American *Mimulus* (Phrymaceae). American Journal of Botany 91:474–489.

Beaulieu, J. M. and B. C. O'Meara. 2015. Extinction can be estimated from moderately sized molecular phylogenies. Evolution 69:1036–1043.

Beaulieu, J. M. and B. C. OMeara. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology 65:583–601.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2005. Genbank. Nucleic Acids Research 33:D34–D38.

Bokma, F. 2002. Detection of punctuated equilibrium from molecular phylogenies. Journal of Evolutionary Biology 15:1048–1056.

Bokma, F. 2008. Detection of "punctuated equilibrium" by Bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. Evolution 62:2718–2726.

Bollback, J. P. 2006. Simmap: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Charlesworth, D. and B. Charlesworth. 1987. Inbreeding depression and its evolutionary consequences. Annual Review of Ecology and Systematics 18:237–268.

Cock, P. J., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al. 2009. BioPython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Colless, D. H. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. Systematic Zoology 31:100–104.

Conti, E., E. Suring, D. Boyd, J. Jorgensen, J. Grant, and S. Kelso. 2000. Phylogenetic relationships and character evolution in *Primula* L.: the usefulness of ITS sequence data. Plant Biosystems 134:385–392.

Coyne, J. A., H. A. Orr, et al. 2004. Speciation. Sinauer Associates Sunderland, MA.

Darwin, C. 1876. The effects of cross and self fertilization in the vegetable kingdom.

de Queiroz, A. and J. Gatesy. 2007. The supermatrix approach to systematics. Trends in Ecology & Evolution 22:34–41.

Dobzhansky, T. G. 1937. Genetics and the Origin of Species. Columbia University Press.

Donoghue, M. J. 2005. Key innovations, convergence, and success: macroevolutionary lessons from plant phylogeny. Paleobiology 31:77–93.

Eriksson, O. and B. Bremer. 1992. Pollination systems, dispersal modes, life forms, and diversification rates in angiosperm families. Evolution 46:258–266.

Escudero, M., M. Hahn, B. H. Brown, K. Lueders, and A. L. Hipp. 2016. Chromosomal rearrangements in holocentric organisms lead to reproductive isolation by hybrid dysfunction: The correlation between karyotype rearrangements and germination rates in sedges. American Journal of Botany 103:1529–1536.

Escudero, M., A. L. Hipp, and M. Luceño. 2010. Karyotype stability and predictors of chromosome number variation in sedges: a study in *Carex* section *Spirostachyae* (Cyperaceae). Molecular Phylogenetics and Evolution 57:353–363.

Escudero, M., S. Martín-Bravo, I. Mayrose, M. Fernández-Mazuecos, O. Fiz-Palacios, A. L. Hipp, M. Pimentel, P. Jiménez-Mejías, V. Valcárcel, P. Vargas, et al. 2014. Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. PLOS ONE 9:e85266.

Feder, J. L., X. Xie, J. Rull, S. Velez, A. Forbes, B. Leung, H. Dambroski, K. E. Filchak, and M. Aluja. 2005. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. Proceedings of the National Academy of Sciences USA 102:6573–6580.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Ferrer, M. M. and S. V. Good. 2012. Self-sterility in flowering plants: preventing self-fertilization increases family diversification rates. Annals of Botany Page mcs124.

Fischer, M. 2012. Perfect taxon sampling and phylogenetically decisive taxon coverage. arXiv preprint arXiv:1206.3472 .

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution 3:1084–1092.

Freyman, W. A. 2015. SUMAC: Constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. Evolutionary Bioinformatics 11:263.

Freyman, W. A. and S. Höhna. 2017a. Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. Systematic Biology syx065.

Freyman, W. A. and S. Höhna. 2017b. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible lineages. bioRxiv 210484.

Glick, L. and I. Mayrose. 2014. ChromEvol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. Molecular Biology and Evolution 31:1914–1922.

Goldberg, E. E. and B. Igić. 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709.

Goldberg, E. E., J. R. Kohn, R. Lande, K. A. Robertson, S. A. Smith, and B. Igić. 2010. Species selection maintains self-incompatibility. Science 330:493–495.

Goldberg, E. E., L. T. Lancaster, and R. H. Ree. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. Systematic Biology 60:451–465.

Gottlieb, L. D. 1973. Genetic differentiation, sympatric speciation, and the origin of a diploid species of *Stephanomeria*. American Journal of Botany Pages 545–553.

Graham, S. A. 2013. Fossil records in the Lythraceae. The Botanical Review 79:48–145.

Grant, V. 1981. Plant speciation. New York: Columbia University Press xii, 563p.-illus., maps, chrom. nos.. En 2nd edition. Maps, Chromosome numbers. General (KR, 198300748).

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Grímsson, F., R. Zetter, and Q. Leng. 2012. Diverse fossil Onagraceae pollen from a Miocene palynoflora of north-east China: early steps in resolving the phytogeographic history of the family. Plant Systematics and Evolution 298:671–687.

Guggisberg, A., G. Mansion, and E. Conti. 2009. Disentangling reticulate evolution in an arctic–alpine polyploid complex. Systematic Biology 58:55–73.

Hartfield, M. 2016. Evolutionary genetic consequences of facultative sex and outcrossing. Journal of Evolutionary Biology 29:5–22.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of phylogenetic analyses. Journal of Systematics and Evolution 46:239–257.

Hipp, A. L. 2007. Nonuniform processes of chromosome evolution in sedges (*Carex*: Cyperaceae). Evolution 61:2175–2194.

Hipp, A. L., P. E. Rothrock, A. A. Reznicek, and P. E. Berry. 2007. Chromosome number changes associated with speciation in sedges: a phylogenetic study in *Carex* section *Ovales* (Cyperaceae) using AFLP data. Aliso: A Journal of Systematic and Evolutionary Botany 23:193–203.

Hipp, D. R. and D. Kennedy. 2007. SQLite. http://www.sqlite.org.

Hobolth, A. and E. A. Stone. 2009. Efficient simulation from finite-state, continuous-time Markov chains with incomplete observations. Annals of Applied Statistics 3:1204–1231.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science 14:382–401.

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

Höhna, S., L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown. 2017. P^3: Phylogenetic Posterior Prediction in RevBayes. Molecular Biology and Evolution msx286.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014a. Probabilistic graphical model representation in phylogenetics. Systematic Biology 63:753–771.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014b. Probabilistic Graphical Model Representation in Phylogenetics. Systematic Biology 63:753–771.

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

Höhna, S., M. L. Landis, and J. P. Huelsenbeck. 2017. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. bioRxiv 104422.

Huelsenbeck, J. P. and J. P. Bollback. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Systematic Biology 50:351–366.

Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock 154:1879–1892.

Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. Stochastic mapping of morphological characters. Systematic Biology 52:131–158.

Igic, B., R. Lande, and J. R. Kohn. 2008. Loss of self-incompatibility and its evolutionary consequences. International Journal of Plant Sciences 169:93–104.

Irvahn, J. and V. N. Minin. 2014. Phylogenetic stochastic mapping without matrix exponentiation. Journal of Computational Biology 21:676–690.

Jeffreys, H. 1961. Theory of probability. 3 ed. Oxford University Press.

Johnson, M. T., R. G. FitzJohn, S. D. Smith, M. D. Rausher, and S. P. Otto. 2011. Loss of sexual recombination and segregation is associated with increased diversification in evening primroses. Evolution 65:3230–3240.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773–795.

Katoh, K., K. Misawa, K.-i. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research 30:3059–3066.

Katoh, K. and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772–780.

Lande, R. and D. W. Schemske. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. Evolution 39:24–40.

Landis, M. J. 2017. Biogeographic dating of speciation times using paleogeographically informed processes. Systematic Biology 66:128–144.

Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. Systematic Biology 62:789–804.

Lee, D. E., J. G. Conran, J. M. Bannister, U. Kaulfuss, and D. C. Mildenhall. 2013. A fossil *Fuchsia* (Onagraceae) flower and an anther mass with in situ pollen from the early Miocene of New Zealand. American Journal of Botany 100:2052–2065.

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

Maddison, W. P. and R. G. FitzJohn. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Systematic Biology 64:127–136.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56:701–710.

Madigan, D. and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 89:1535–1546.

Marcussen, T., L. Heier, A. K. Brysting, B. Oxelman, and K. S. Jakobsen. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). Systematic Biology 64:84–101.

May, M. R., S. Höhna, and B. R. Moore. 2016. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. Methods in Ecology and Evolution 7:947–959.

Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. Systematic Biology 59:132–144.

Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257–1257.

Melters, D. P., L. V. Paliulis, I. F. Korf, and S. W. Chan. 2012. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Research 20:579–593.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21:1087–1092.

Michel, A. P., S. Sim, T. H. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010. Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences USA 107:9724–9729.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can be estimated from molecular phylogenies. Philosophical Transactions of the Royal Society B: Biological Sciences 344:77–82.

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society B: Biological Sciences 344:305–311.

Ng, J. and S. D. Smith. 2014. How traits shape trees: new approaches for detecting character state-dependent lineage diversification. Journal of Evolutionary Biology 27:2035–2045.

Nielsen, R. 2002. Mapping mutations on phylogenies. Systematic Biology 51:729–739.

Ohi-Toma, T., T. Sugawara, H. Murata, S. Wanke, C. Neinhuis, and J. Murata. 2006. Molecular phylogeny of *Aristolochia* sensu lato (Aristolochiaceae) based on sequences of rbcL, matK, and phyA genes, with special reference to differentiation of chromosome numbers. Systematic Botany 31:481–492.

Pagel, M. and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. The American Naturalist 167:808–25.

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Systematic Biology 53:673–684.

Pires, J. C. and K. L. Hertweck. 2008. A renaissance of cytogenetics: Studies in polyploidy and chromosomal evolution. Annals of the Missouri Botanical Garden 95:275–281.

Posada, D. and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53:793–808.

Pupko, T., I. Pe, R. Shamir, and D. Graur. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution 17:890–896.

Python Software Foundation. 2008. Python multiprocessing module. http://docs.python.org/library/multiprocessing.html.

Rabosky, D. L. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64:1816–1824.

Rabosky, D. L. and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Systematic Biology 64:340–355.

Raven, P. H. 1979. A survey of reproductive biology in Onagraceae. New Zealand Journal of Botany 17:575–593.

Ree, R. H. and S. A. Smith. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology 57:4–14.

Rieseberg, L. H. and J. H. Willis. 2007. Plant speciation. Science 317:910–914.

Rodrigue, N., H. Philippe, and N. Lartillot. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics 24:56–62.

Rodriguez, F., J. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology 142:485–501.

Sanderson, M. J., D. Boss, D. Chen, K. A. Cranston, and A. Wehe. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Systematic Biology 57:335–46.

Sanderson, M. J. and M. J. Donoghue. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. Trends in Ecology & Evolution 11:15–20.

Sanderson, M. J., M. M. McMahon, and M. Steel. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC evolutionary biology 10:155.

Scarpino, S. V., D. A. Levin, and L. A. Meyers. 2014. Polyploid formation shapes flowering plant diversity. The American Naturalist 184:456–465.

Schwarz, G. 1978. Estimating the dimension of a model. The Annals of Statistics 6:461–464.

Sibson, R. 1973. SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal 16:30–34.

Smith, S. A., J. M. Beaulieu, and M. J. Donoghue. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evolutionary Biology 9:37.

Sneath, P. H. A. 1957. The application of computers to taxonomy. Journal of General Microbiology 17:201–226.

Stadler, T. 2011. Simulating trees with a fixed number of extant species. Systematic Biology 60:676–684.

Stebbins, G. L. 1957. Self fertilization and population variability in the higher plants. The American Naturalist 91:337–354.

Stebbins, G. L. 1971. Chromosomal evolution in higher plants. Edward Arnold Ltd., London.

Stebbins, G. L. 1974. Flowering plants: evolution above the species level. Edward Arnold Ltd., London.

Steel, M. and M. J. Sanderson. 2010. Characterizing phylogenetically decisive taxon coverage. Applied Mathematics Letters 23:82–86.

Steinbachs, J. and K. Holsinger. 2002. S-RNase–mediated gametophytic self-incompatibility is ancestral in eudicots. Molecular Biology and Evolution 19:825–829.

Sytsma, K., A. Litt, and M. Zjhra. 2004. Clades, clocks, and continents: Historical and biogeographical analysis of Myrtaceae, Vochysiaceae, and relatives in the southern Hemisphere source. International Journal of Plant Sciences 165:S85–S105.

Szathmary, E. and J. M. Smith. 1995. The major evolutionary transitions. Nature 374:227.

Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C. E. Hinchliff, J. W. Brown, E. B. Sessa, and L. J. Harmon. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytologist 207:454–467.

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

Timme, R. E., B. B. Simpson, and C. R. Linder. 2007. High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. American Journal of Botany 94:1837–1852.

Van Dyk, D. A. and X.-L. Meng. 2001. The art of data augmentation. Journal of Computational and Graphical Statistics 10:1–50.

Vickery, R. K. 1995. Speciation by aneuploidy and polyploidy in *Mimulus* (Scrophulariaceae). The Great Basin Naturalist 55:174–176.

Von Haeseler, A. 2012. Do we still need supertrees? BMC biology 10:13.

Vos, R. A., H. Lapp, W. H. Piel, and V. Tannen. 2010. TreeBase2: rise of the machines. Nature Precedings 4600.1.

Wagner, W. L., P. C. Hoch, and P. H. Raven. 2007. Revised classification of the Onagraceae. Systematic Botany Monographs 83.

White, M. J. D. 1978. Modes of speciation. San Francisco: WH Freeman 455p.-Illus., maps, chrom. nos.. General (KR, 197800185).

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Systematic Biology 60:150–60.

Xu, X., D. Dimitrov, C. Rahbek, and Z. Wang. 2015. NCBIminer: sequences harvest from Genbank. Ecography 38:426–430.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306–314.

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution 14:717–724.

Zhan, S. H., M. Drori, E. E. Goldberg, S. P. Otto, and I. Mayrose. 2016. Phylogenetic evidence for cladogenetic polyploidization in land plants. American Journal of Botany 103:1252–1258.

Zhi-Chen, S., W. Wei-Ming, and H. Fei. 2004. Fossil pollen records of extant angiosperms in China. The Botanical Review 70:425–458.